

# AI Mastery Course

## Module 4

Supervised machine learning and  
predictive modelling


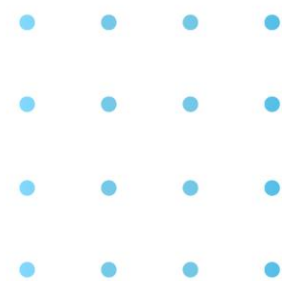
### Section

Linear regression



# Learning Objectives

Di akhir modul ini, Anda akan dapat:

- Memahami lebih detil tentang teknik regression dan berbagai use case nya
  - Melihat bagaimana konsep gradient descent di linear regression
  - Menjelaskan fungsi dari cost function
  - Melakukan training model untuk memahami konsep gradient descent
  - Mendeskripsikan regularisasi pada regression seperti: Lasso dan Ridge Regression
  - Mendefinisikan Generalized Linear Regression (GLM)
- 
- 



# Agenda

01

## LINEAR REGRESSION

- Regression Introduction
- Objectives of linear regression
- Variables affecting LR

02

## LINEAR REGRESSION

- Understanding linear regression
- Error function
- Optimization – Gradient descent

03

## K FOLD

- Cross validation
- K fold cross validation

04

## REGULARIZATION

- Lasso
- Ridge

05

## CONCLUSION

- Summary



# 01

## LINEAR REGRESSION

- Regression Introduction
- Objectives of linear regression
- Variables affecting LR

# Regression Technique

Analisa regression digunakan:

- Untuk memprediksi nilai-nilai dari dependent variable Y, berdasarkan relasinya dengan nilai-nilai dari independent variable X (minimal 1)
- Untuk menjelaskan apa pengaruh dari perubahan pada sebuah independent variable terhadap dependent variable melalui estimasi nilai numerik dari hubungan yang ada

Dependent variable  
(Y)

Variable yang ingin diperjelas

Independent  
variable (X)

Variable yang digunakan untuk menjelaskan dependent variable

Coefficients

Nilai yang memuat penjelasan tentang hubungan terhadap dependent variable yang dihitung dengan regression tool

Residuals

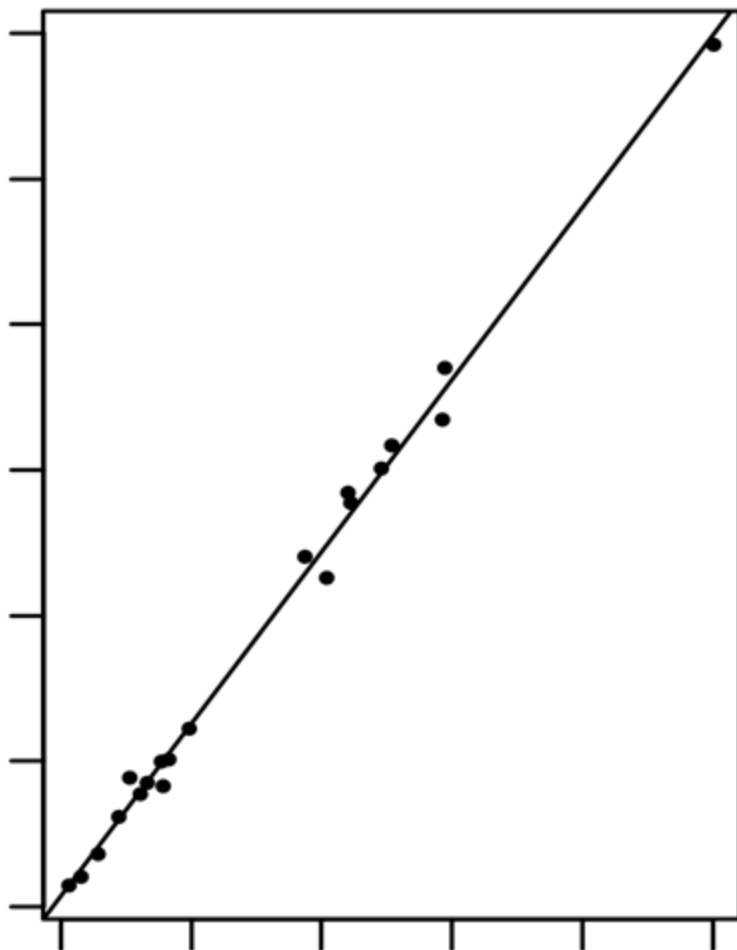
Porsi dari dependent variable yang tidak dijelaskan oleh model (model melakukan under dan over predictions)

# Regression in Real World

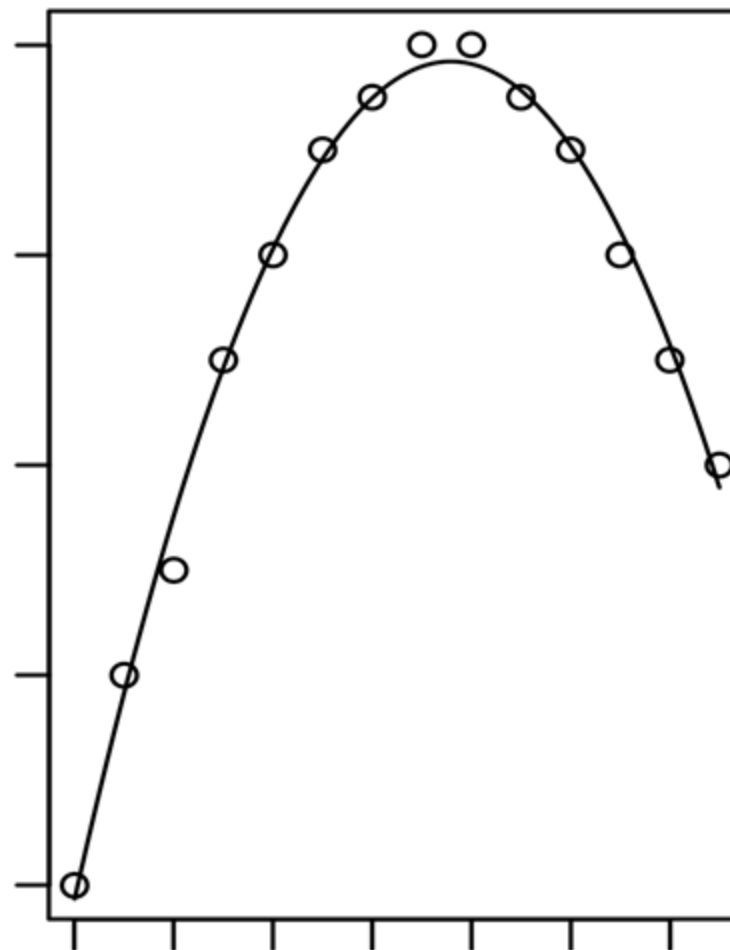
- Dalam penerapannya kita sering mendapati data-data numerik yang disimpan dalam bentuk tabel. Regression analysis sebagai tool yang sangat baik bisa digunakan untuk menganalisa data numerik tersebut. Secara umum, analisa regresi adalah sebuah proses untuk menemukan best fits dari sekumpulan data points.
- Contohnya: Kita memiliki beberapa deskriptor dari sebuah lagu seperti genre, author, duration, lyrics dan lain sebagainya. Tujuan dari masalah ini adalah untuk memprediksi tahun ketika lagu tersebut diproduksi. Pada dasarnya masalah ini bisa dikategorikan sebagai masalah regresi, karena variabel target yang ingin diprediksi adalah angka dalam kisaran antara tahun 1922 dan 2011.

# Regression Types

Linear



Non - Linear



# Linear Regression Types

**Number of the explanatory variables**



```
graph TD; A[Number of the explanatory variables] --> B[Simple regression]; A --> C[Multiple regression]
```

Simple regression

Multiple regression



# Objectives of Linear Regression

- Menganalisa relasi antara dua variable

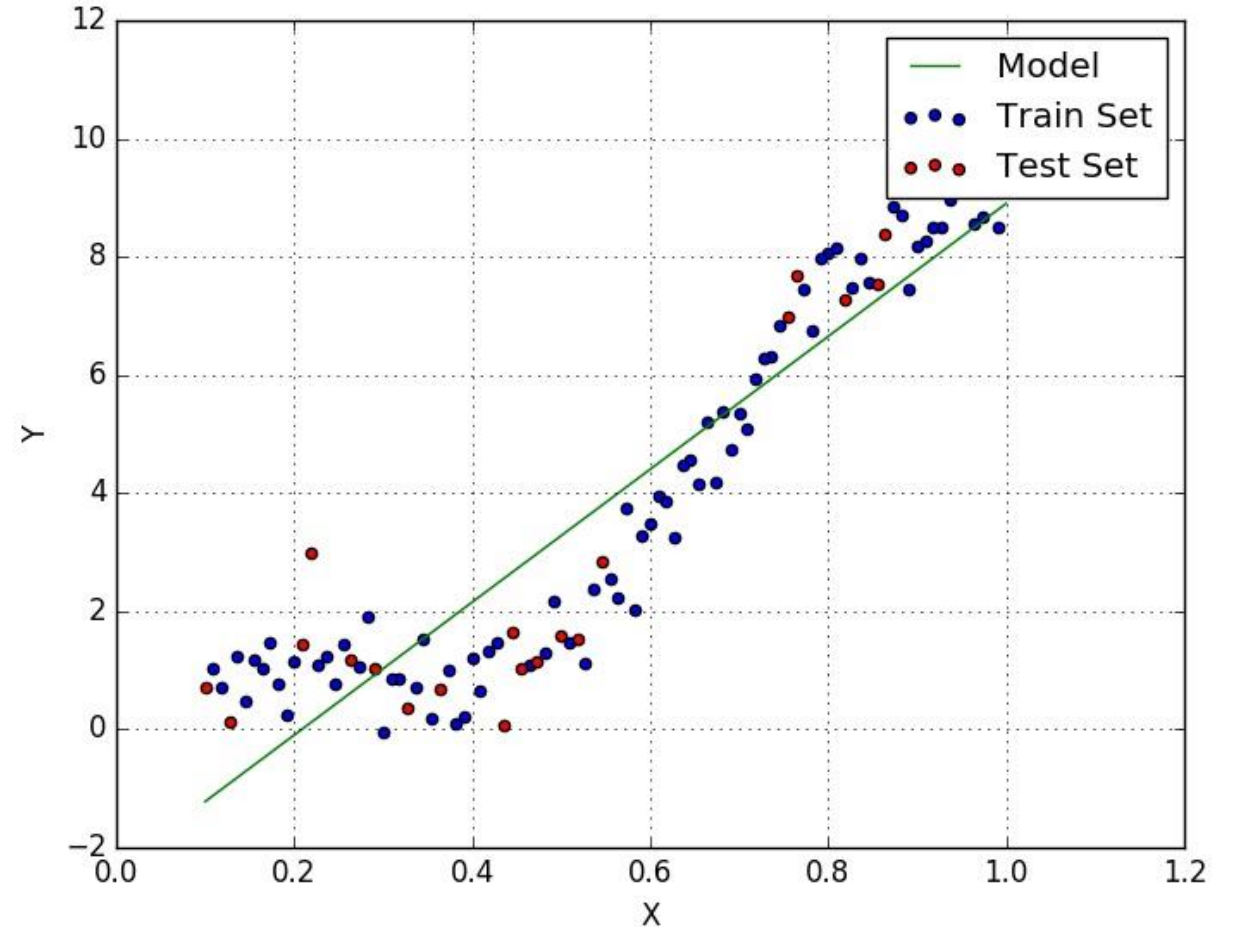
Contoh – relasi antara harga dan area/luas rumah, jumlah jam belajar dengan nilai yang didapat, pendapatan dan pengeluaran dan lain sebagainya

- Memprediksi nilai baru yang mungkin

Berdasarkan area/luas rumah memprediksi harga rumah tersebut di bulan tertentu, berdasarkan jumlah jam belajar memprediksi kemungkinan nilai yang akan didapat, memprediksi total penjualan di 3 bulan berikutnya dan lain sebagainya

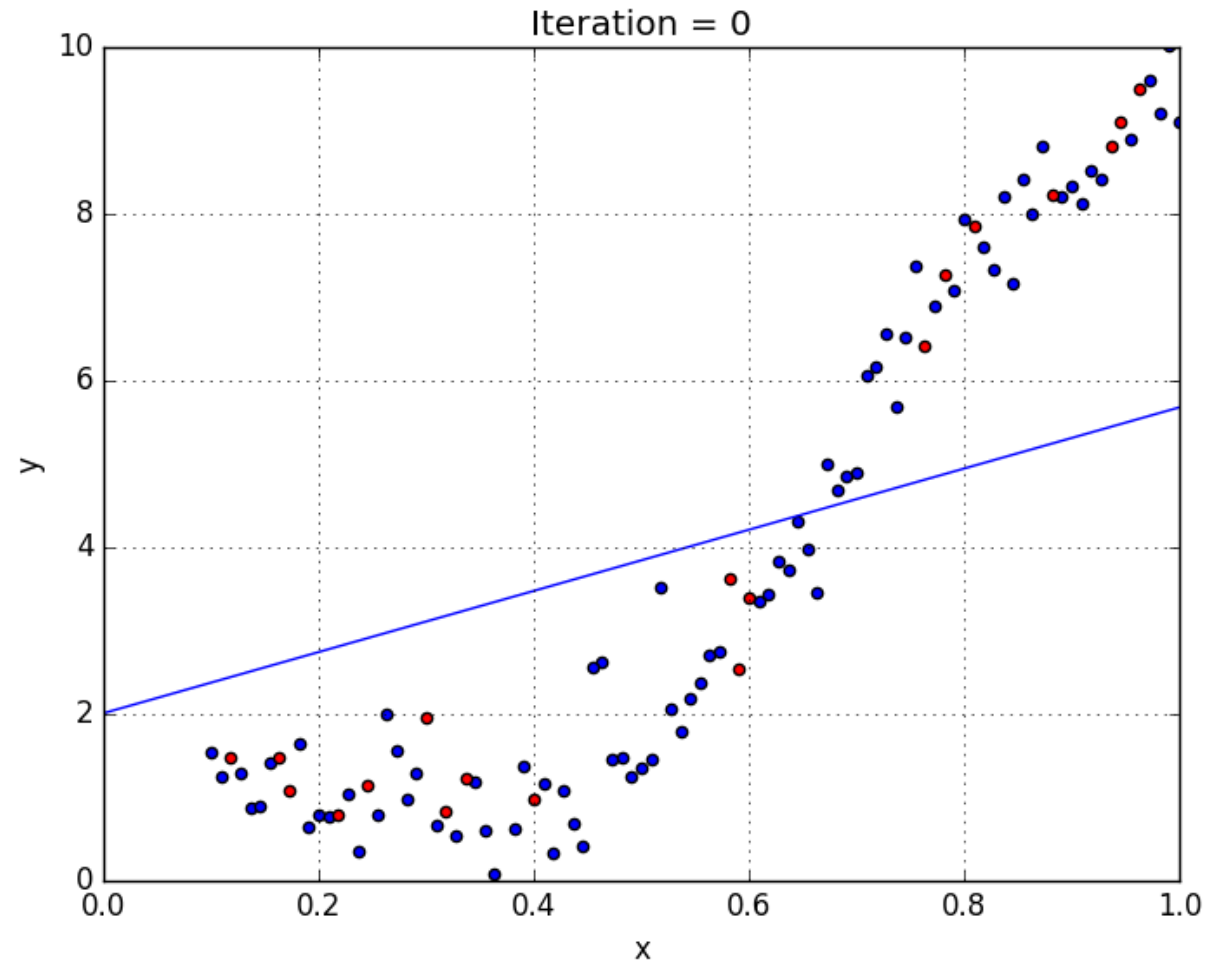
# What is Linear Regression?

- Algoritma supervised learning yang belajar dari sekumpulan sampel data latih
- Linear Regression mengestimasi hubungan antara dependent variable (target/label) dan 1 atau lebih independent variable (predictors).



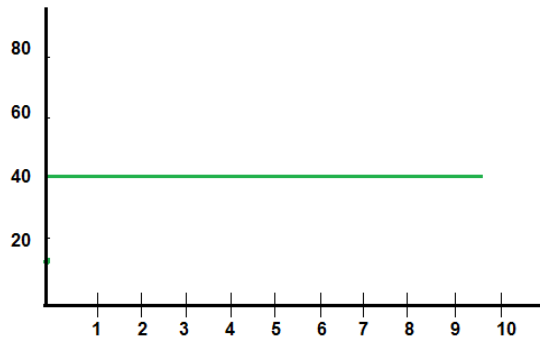
# Univariate Linear Regression

Selama periode latih (training), garis regresi akan sampai pada kondisi "fit".

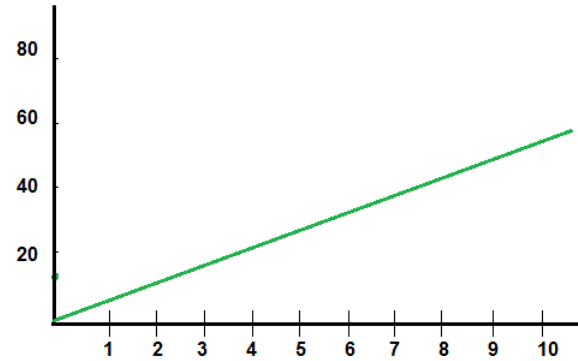


# Linear Regression – Variables affecting Regression Equation

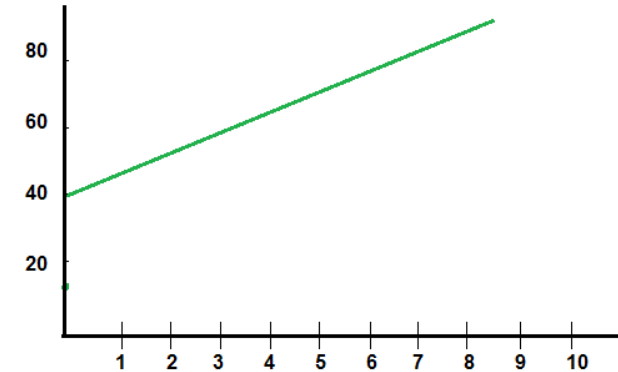
$$m = 0$$
$$c = 40$$



$$m = 0.8$$
$$c = 0$$



$$m = 0.8$$
$$c = 40$$



$$\hat{y} = mx + c$$



## 02

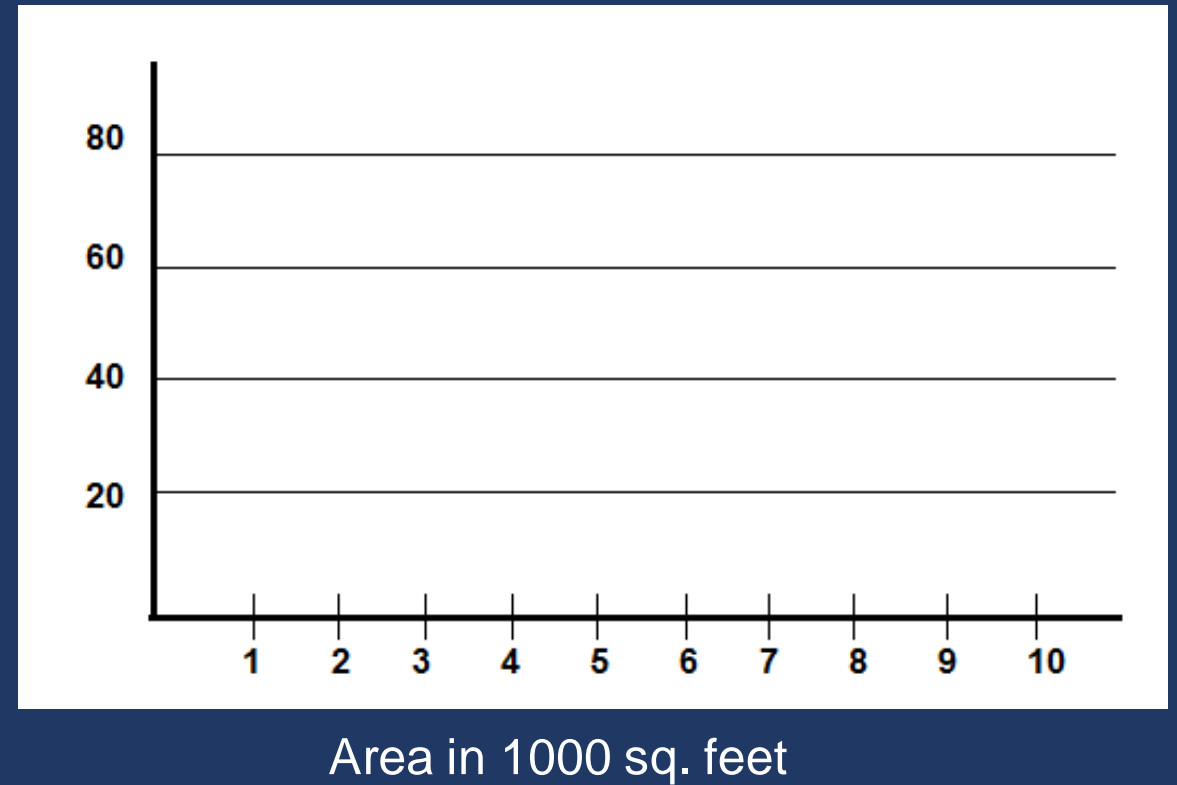
## LINEAR REGRESSION

- Understanding linear regression
- Error function
- Optimization – Gradient descent

# Linear Regression – Housing prices prediction

Area ( sq ft)	Price In IDR
1200	2,600,000K
1800	4,200,000K
3200	7,400,000K
3800	6,200,000K
4200	7,050,000K

Price in 100 ML (IDR)

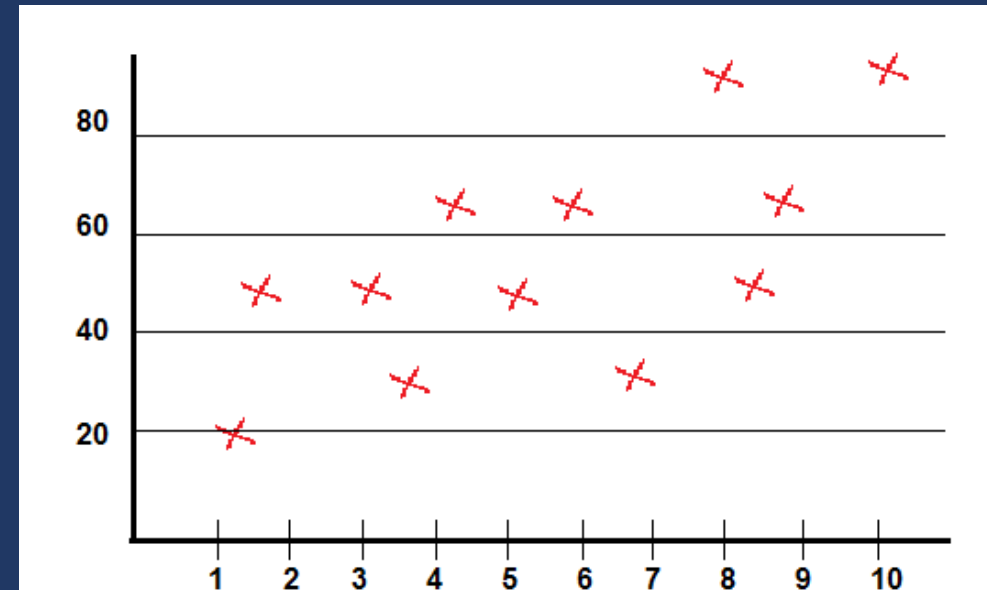


# Linear Regression – Housing prices prediction

Area ( sq ft)	Price In IDR
1200	2,600,000K
1800	4,200,000K
3200	7,400,000K
3800	6,200,000K
4200	7,050,000K

y: Dependent Variable, criterion variable.  
x: Independent variable, predictor variables.

Price in 100 ML (IDR)



Area in 1000 sq. feet

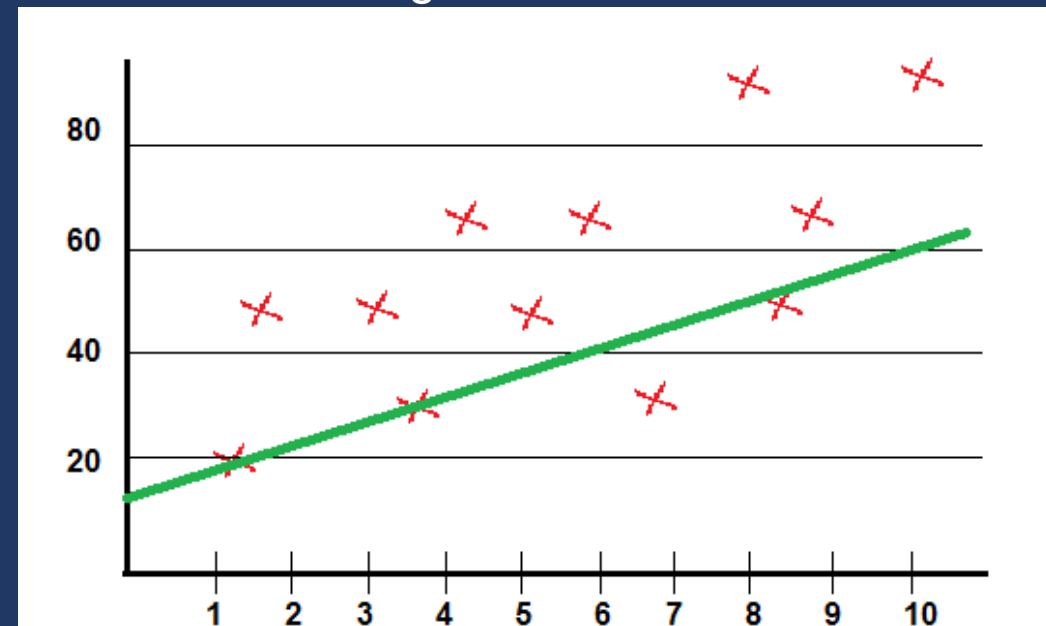
# Linear Regression – Housing prices prediction

Area ( sq ft)	Price In IDR
1200	2,600,000K
1800	4,200,000K
3200	7,400,000K
3800	6,200,000K
4200	7,050,000K

$$\hat{y} = mx + c$$

$\hat{y}$  = Value predicted by current Algorithm  
Linear Regression in one Variable

Price in 100 ML (IDR)



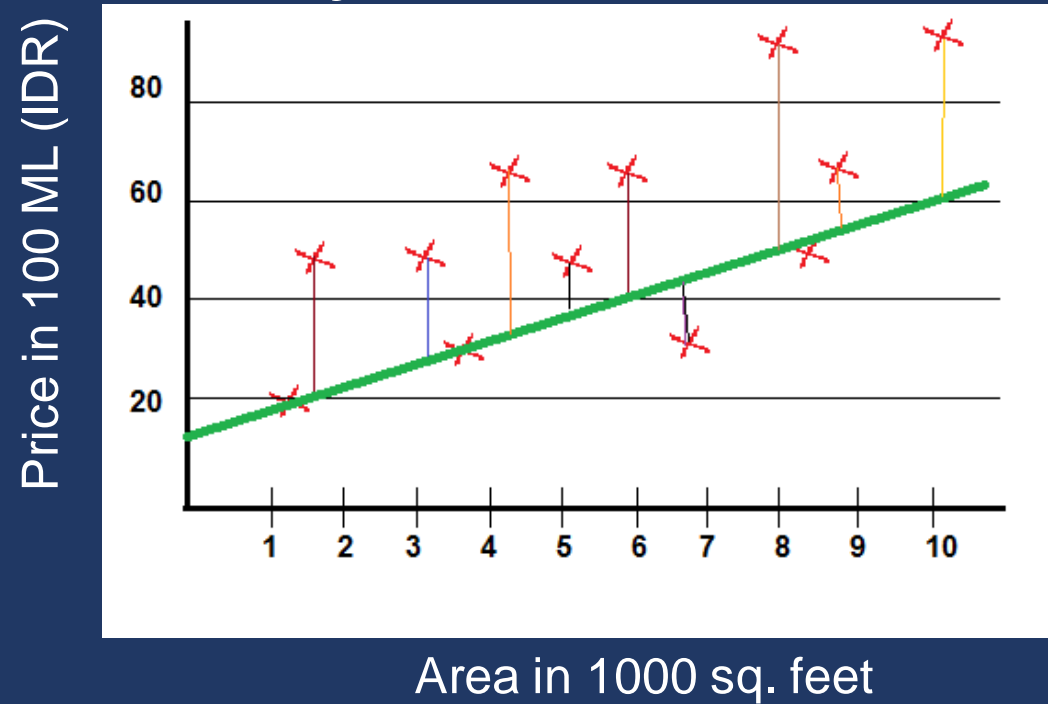
Area in 1000 sq. feet



# Linear Regression – Housing prices prediction

*minimize*  
 $(y - \hat{y})$

*Predictor*  
 $\hat{y} = mx + c$



# Linear Regression – Housing prices prediction

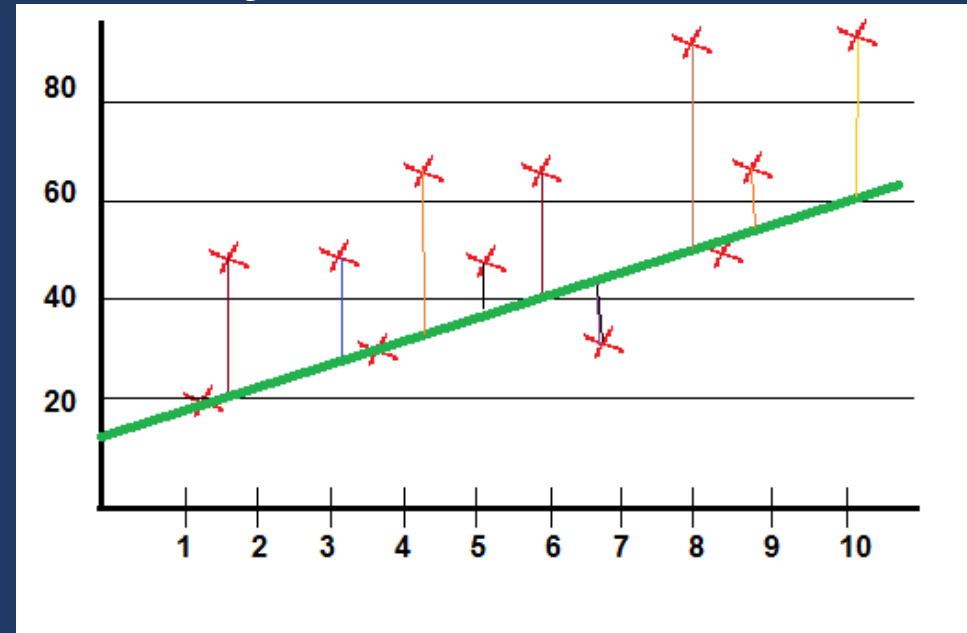
*Cost Function*

$$J = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$j(m_i, c) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

*Predictor*  
 $\hat{y} = mx + c$

Price in 100 ML (IDR)



Area in 1000 sq. feet

# Linear Regression

Regression Equation:

$$\hat{y} = mx + c$$

Parameters

$$m_i, c$$

Cost Function:

$$j(m_i, c) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Goal

$$\underset{m_i, c}{\text{minimize}} J(m_i, c)$$

# Linear Regression — Gradient Descent Algorithm

Repeat Until converge

$$w_j := w_j - lr \frac{\partial}{\partial w} J(w_j)$$

simultaneously update,  $j = 0, j = 1$   
where,  $w$ =parameter (coefficient & constant)

# Learning Rate - $\eta$

- Learning rate  $\eta$  akan mengontrol seberapa besar update yang akan terjadi pada parameter weight
- Jika  $\eta$  terlalu kecil, maka gradient descent menjadi sangat lamban
- Jika  $\eta$  terlalu besar, maka gradient descent akan mengalami overshoot dan pada akhirnya gagal untuk convergence

# Linear Regression - OLS

Equation Line formula:

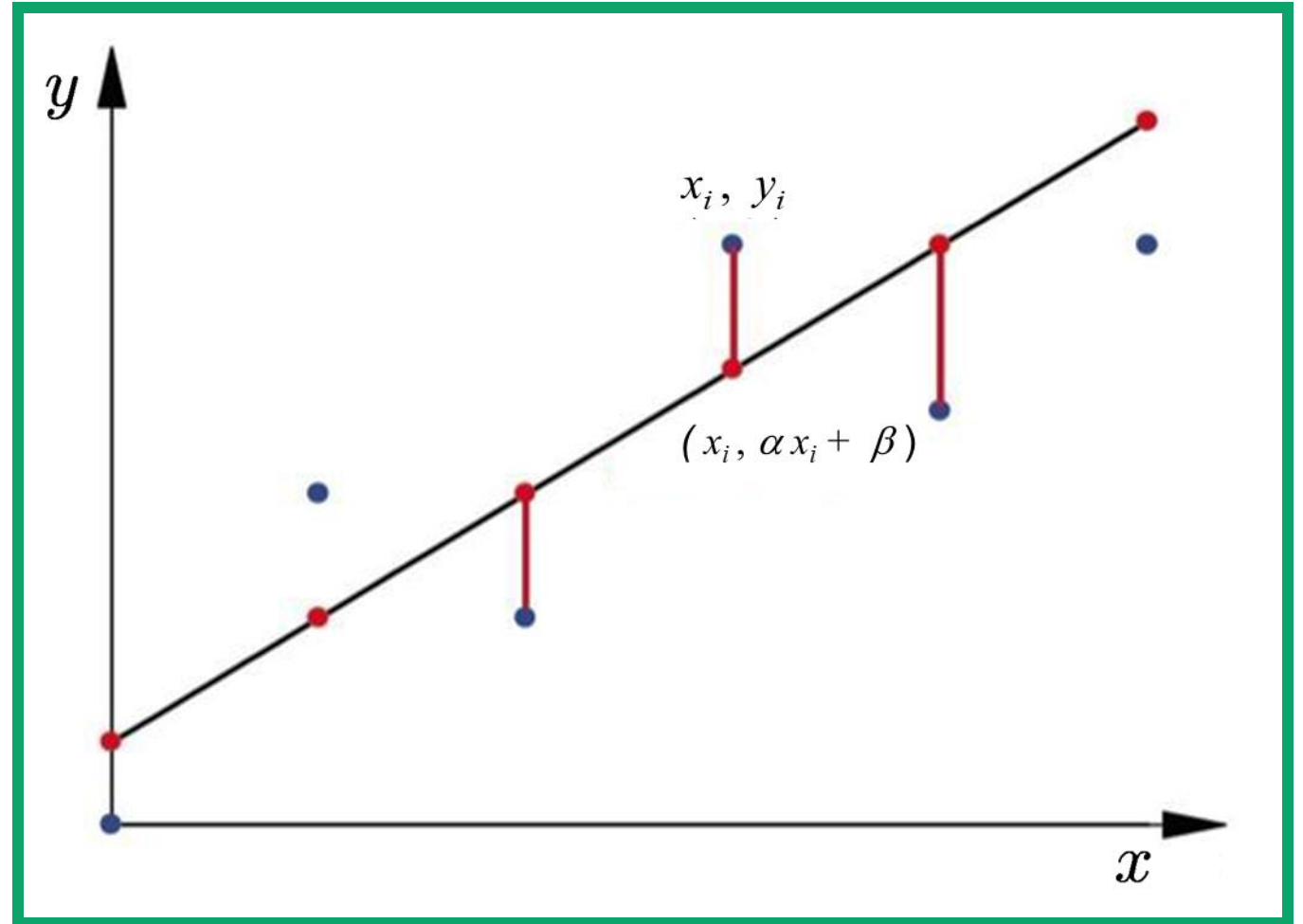
$$y = \alpha * x + \beta$$

Quantity in Minimal

$$E = \sum_{i=1}^n (\alpha x_i + \beta - y_i)^2$$

Sum of the squares of the distances

$$(x_i, \alpha x_i + \beta)$$



# Linear Regression – R squared

## RSS

$$SS_{res} = \sum_{i=1}^n (y_i - \alpha x_i + \beta)^2$$

## TSS

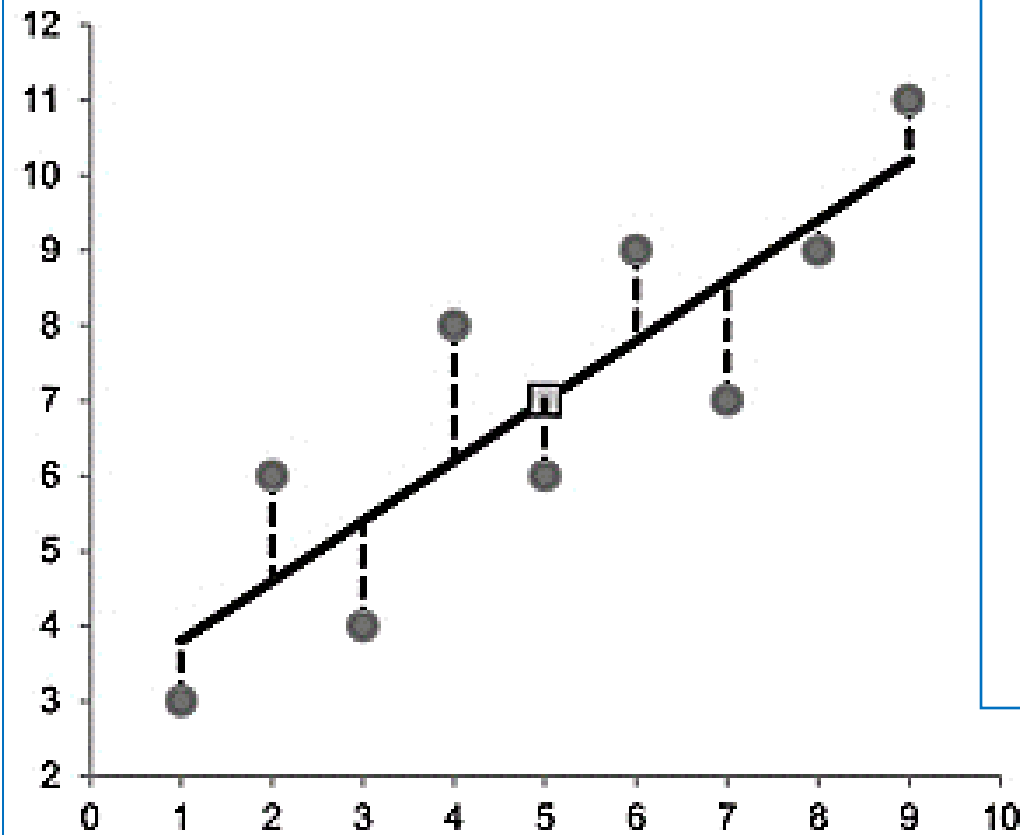
$$SS_{tot} = \sum_{i=1}^n (y_i - y_{mean})^2$$

## R-squared

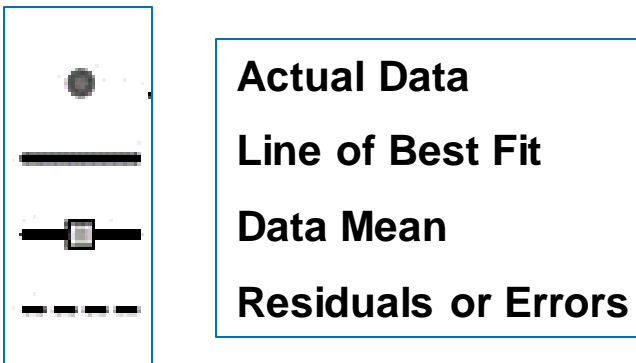
$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \alpha x_i + \beta)^2}{\sum_{i=1}^n (y_i - y_{mean})^2}$$

# Linear Regression – Goodness of Fit (Contd.)

## Least Squares Linear Regression



Meminimalkan jumlah kuadrat dari residu vertikal antara setiap titik dan garis yang berada di antara data





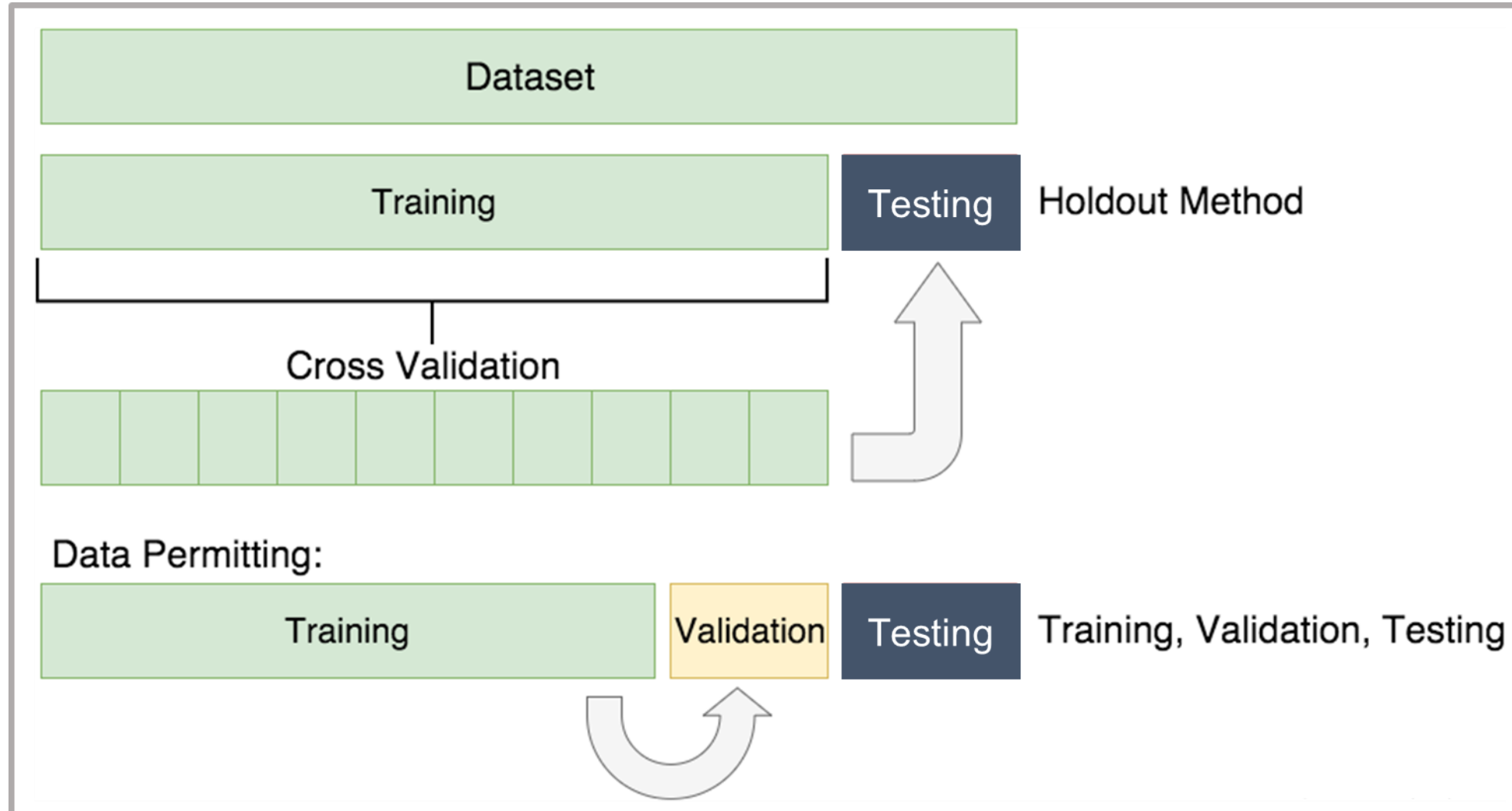


03

K FOLD

- Cross validation
- K fold cross validation

# Testing Model Using Cross Validation



# Cross Validation Types

## K – Fold

Satu subset data digunakan untuk melakukan validasi/pengujian terhadap model sedangkan sisanya digunakan untuk data latih.

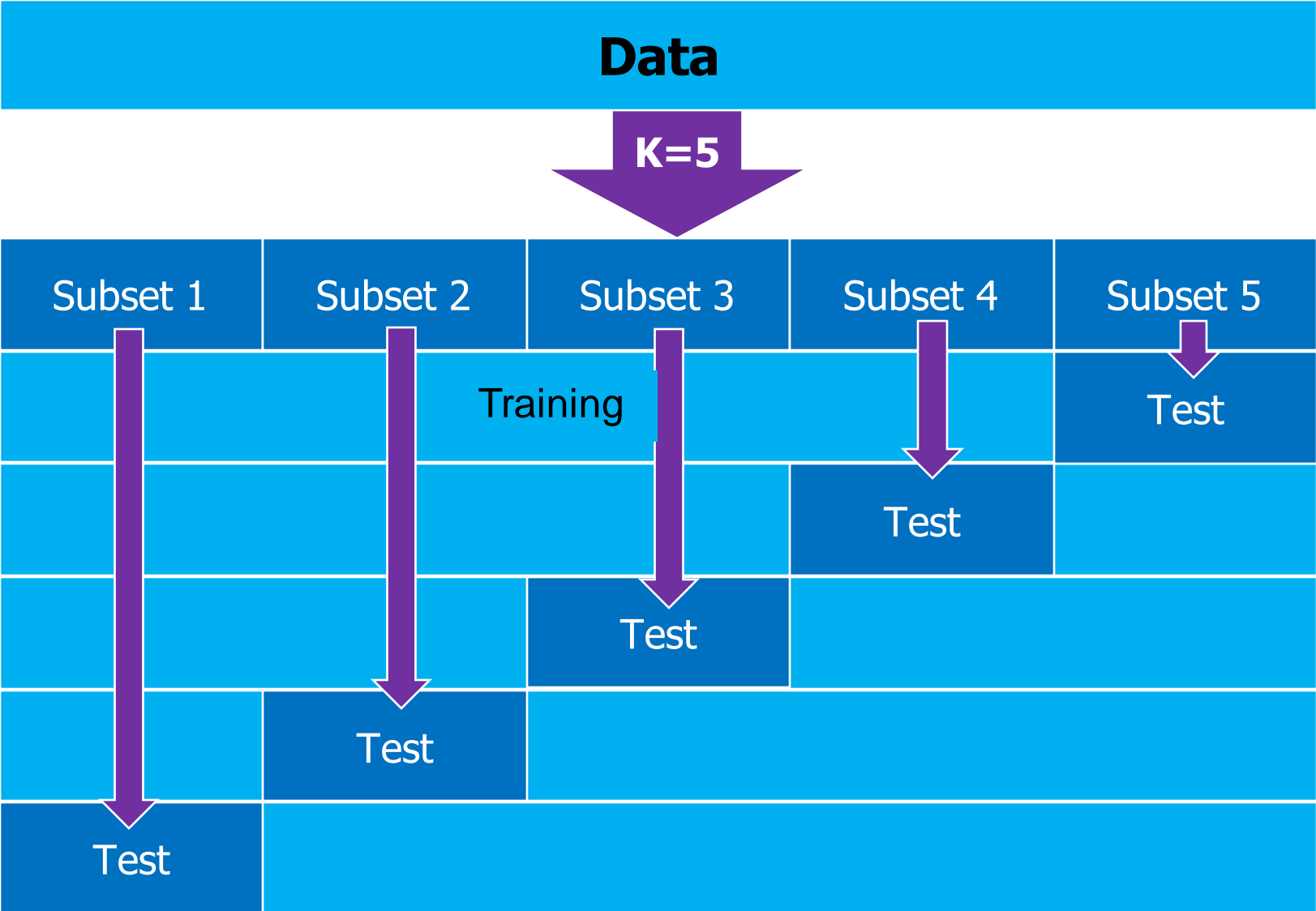
## LeaveOneOut

k sama dengan total observasi yang ada pada data yang mana setiap 1 data observasi digunakan sebagai data validasi atau data uji, sedangkan sisanya digunakan untuk data latih.

## Stratified K– Fold

Lipatan/fold dibuat dengan cara mempertahankan persentase sampel untuk setiap kelas.

# K-Fold Cross Validation



# K-Fold Cross Validation – With an Example

## Import Libraries

```
1 import pandas
2 from sklearn.model_selection import KFold
3 from sklearn.preprocessing import MinMaxScaler
4 from sklearn.model_selection import train_test_split # Import train_test_split function
5 from sklearn.svm import SVR
6 import numpy as np
7 import pandas as pd
```

## Load dataset and set Kfold parameters

```
1 col_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']
2 # load dataset
3 pima = pd.read_csv("pima-indians-diabetes.csv", header=None, names=col_names)
4 kfold = KFold(3, True, 1)
```



04

## REGULARIZATION

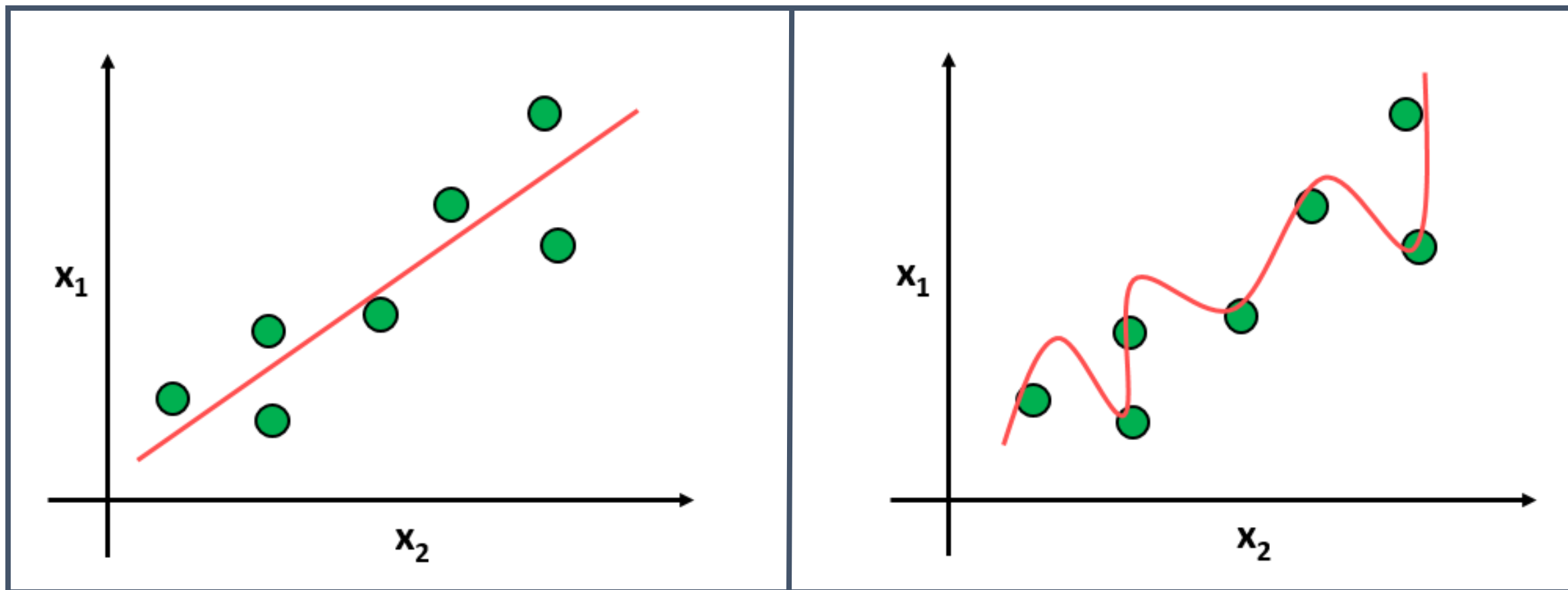
- LASSO
- RIDGE



# Overfitting & Generalisation

- Ketika kita melatih sebuah model dengan lebih banyak data, model tersebut kemungkinan akan mulai semakin akurat terhadap data latih, namun menjadi lebih buruk dalam menangani data uji yang akan diberikan kemudian
- Kondisi ini dikenal dengan istilah “over-fitting” yang pada akhirnya akan mempengaruhi performa model (generalization error).
- Coefficients yang bernilai besar berpotensi terjadi overfitting
- Teknik untuk menyelesaikan masalah large coefficients: Regularization

# Good fit v/s overfitting





# How to minimize overfitting?

- Untuk meminimalisir generalization error (overfitting) kita harus mengumpulkan data sampel sebanyak mungkin
- Gunakan random subset dari data sampel yang kita miliki untuk proses latih (training)
- Gunakan data sampel yang tersisa untuk melakukan validasi/pengujian seberapa baik performa suatu model ketika diberikan data yang tidak pernah digunakan selama proses latih

# Lasso (L1) Regression

- Least Absolute Shrinkage and Selection Operator (LASSO)
- Memiliki jumlah sample yang sangat besar ( $n$ ) sehubungan dengan jumlah dimensi ( $d$ ) akan meningkatkan kualitas dari suatu model
- Satu cara untuk mengurangi jumlah dimensi adalah dengan memanfaatkan data sampel yang memiliki pengaruh yang paling besar dan mengabaikan data sampel yang tidak berpengaruh (noise)
- L1 regularization bisa bekerja secara demikian dengan menerapkan penalty yang akan menghilangkan bobot dari dimensi yang berperilaku sebagai noise
- L1 regularization memanfaatkan sparse vector (vector dengan banyak nilai 0)

# Lasso (L1) Regression

- Tergantung dari seberapa besar regularization strength, weights tertentu pada akhirnya bisa bernilai zero, yang mana membuat LASSO juga yang mana membuat LASSO juga bermanfaat untuk teknik feature selection:

$$j(w_i) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \|w_i\|$$

# Ridge (L2) Regression

- Cara lain untuk mengurangi kompleksitas dari suatu model dan mencegah overfitting terhadap outliers adalah dengan menggunakan L2 regression yang juga dikenal ridge regression.
- Di L2 Regularization kita melibatkan term tambahan untuk cost function yang memiliki efek penalty terhadap large weights yang pada akhirnya akan meminimalisir skew.

# Ridge (L2) Regression

- Ridge regression adalah L2 model yang mendapatkan penalty dimana secara sederhana menambahkan squared sum of the weights pada least-squares cost function:

$$j(w_i) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \|w_i\|^2$$

- Dengan meningkatkan nilai dari hyperparameter  $\lambda$ , kita bisa meningkatkan kekuatan dari regularization strength dan mengurangi weights dari suatu model

# L1 & L2 Regularisation (Elastic Net)

- L1 Regularisation meminimalisir dampak dari dimensi yang memiliki pengaruh yang lemah atau dianggap sebagai “noise”
- L2 Regularisation meminimalisir dampak dari outliers dari data latih yang ada
- L1 & L2 Regularisation bisa digunakan secara bersama-sama dan kombinasi ini dikenal dengan istilah Elastic Net regularisation
- Karena differential of the error function memuat sigmoid yang tidak memiliki inverse, kita tidak bisa menemukan  $w$  dan harus menggunakan gradient descent



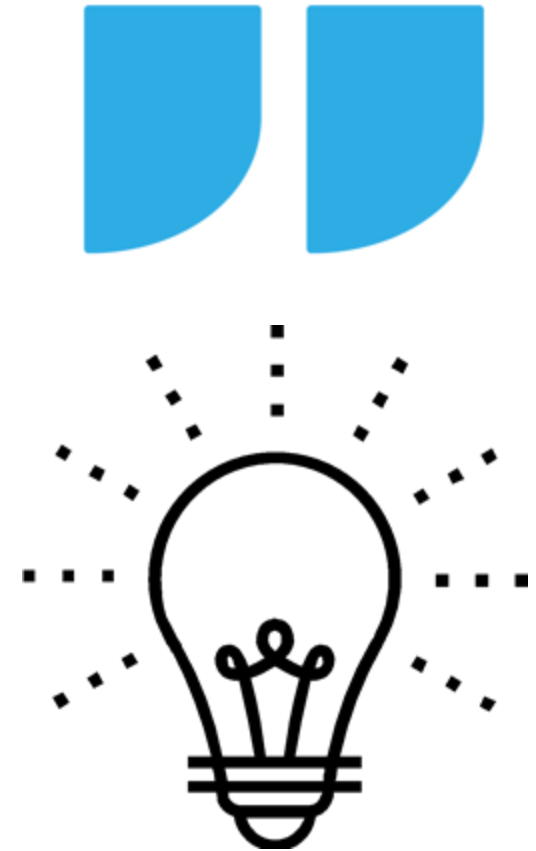
05

## CONCLUSION

- Summary

# Summary

- Regression technique dan use cases
- Gradient descent dengan linear regression
- Fungsi dari Cost function
- Melakukan training model untuk memahami gradient descent di Linear Regression
- Cross validation
- K-fold cross validation
- Regularisation pada regresi yang terdiri dari 2 tipe: Lasso and Ridge Regression





# Quiz

## Question

Di persoalan linear regression, kita menggunakan "R-squared" untuk mengukur goodness-of-fit. Ketika kita menambahkan feature di linear regression model tersebut dan melakukan retraining dengan menggunakan model yang sama, mana dari pernyataan berikut ini yang benar?

- A. Jika R Squared meningkat, maka variable tersebut memiliki pengaruh signifikan
- B. Jika R Squared berkurang, maka variable tersebut tidak memiliki pengaruh signifikan
- C. R squared tidak bisa memberi tahu tentang pengaruh variable
- D. Tidak ada yang benar

# Quiz

## Question

Di persoalan linear regression, kita menggunakan “R-squared” untuk mengukur goodness-of-fit. Ketika kita menambahkan feature di linear regression model tersebut dan melakukan retraining dengan menggunakan model yang sama, mana dari pernyataan berikut ini yang benar?

- A. Jika R Squared meningkat, maka variable tersebut memiliki pengaruh signifikan
- B. Jika R Squared berkurang, maka variable tersebut tidak memiliki pengaruh signifikan
- C. R squared tidak bisa memberi tahu tentang pengaruh variable
- D. Tidak ada yang benar

**Jawaban: C**



## Orbit Future Academy

PT Orbit Ventura Indonesia  
Center of Excellence (Jakarta Selatan)  
Gedung Veteran RI, Lt.15  
Unit Z15-002, Plaza Semanggi  
Jl. Jenderal Sudirman Kav.50, Jakarta  
12930, Indonesia

- 📖 Jakarta Selatan/Pusat
- 📖 Jakarta Barat/BSD
- 📖 Kota Bandung
- 📖 Kab. Bandung
- 📖 Jawa Barat

## Hubungi Kami

Director of Sales & Partnership  
[ira@orbitventura.com](mailto:ira@orbitventura.com)  
**+62 858-9187-7388**

## Social Media



# TERIMA KASIH