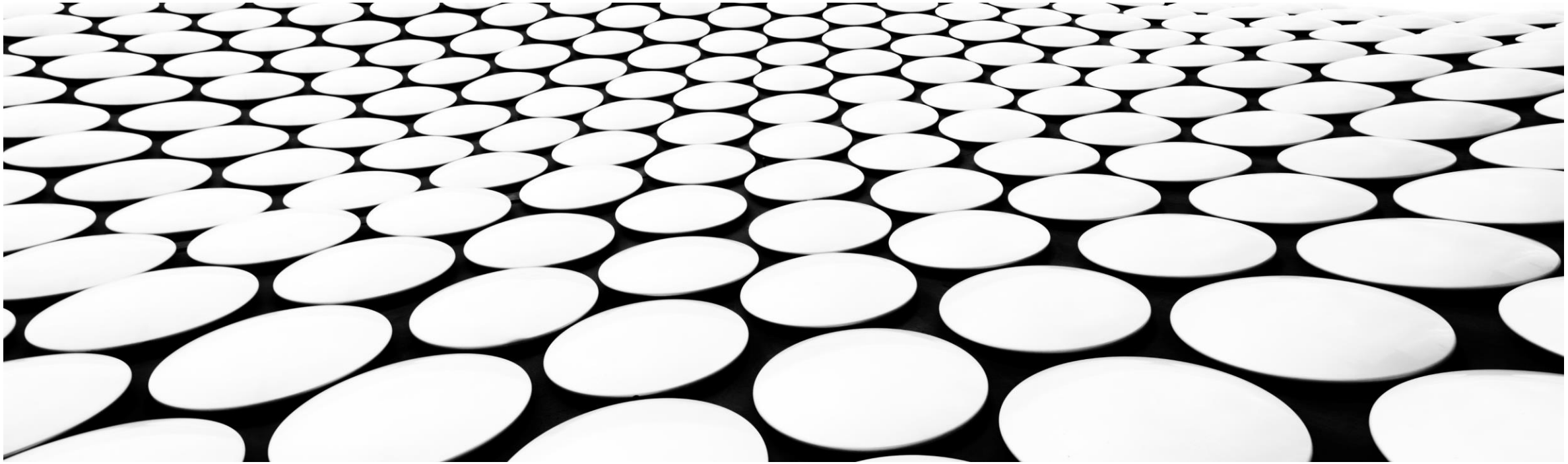
CLUSTERING AND REGRESSION MACHINE LEARNING IN PYTHON

PERTEMUAN - 4

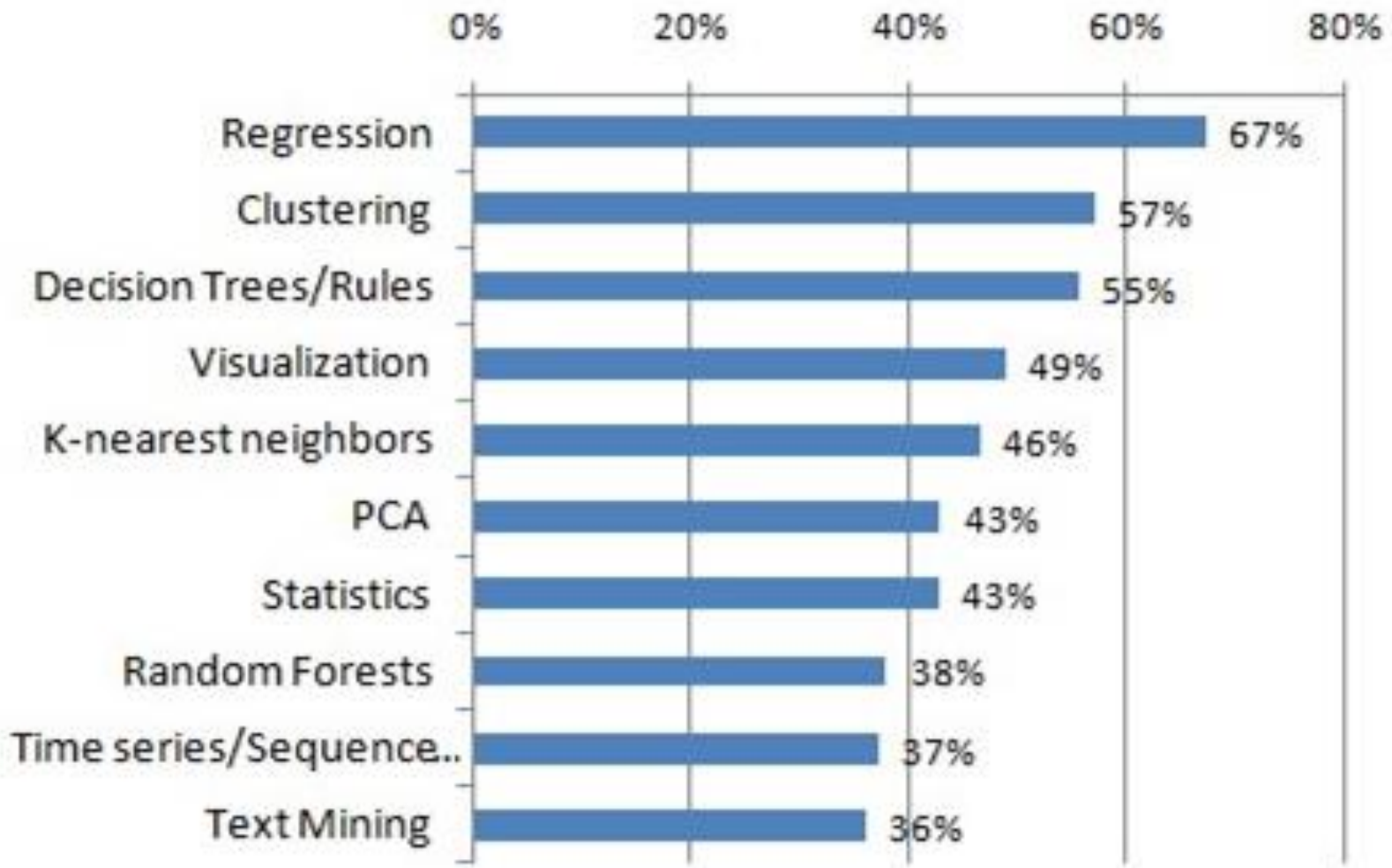
AI ACADEMY 2021

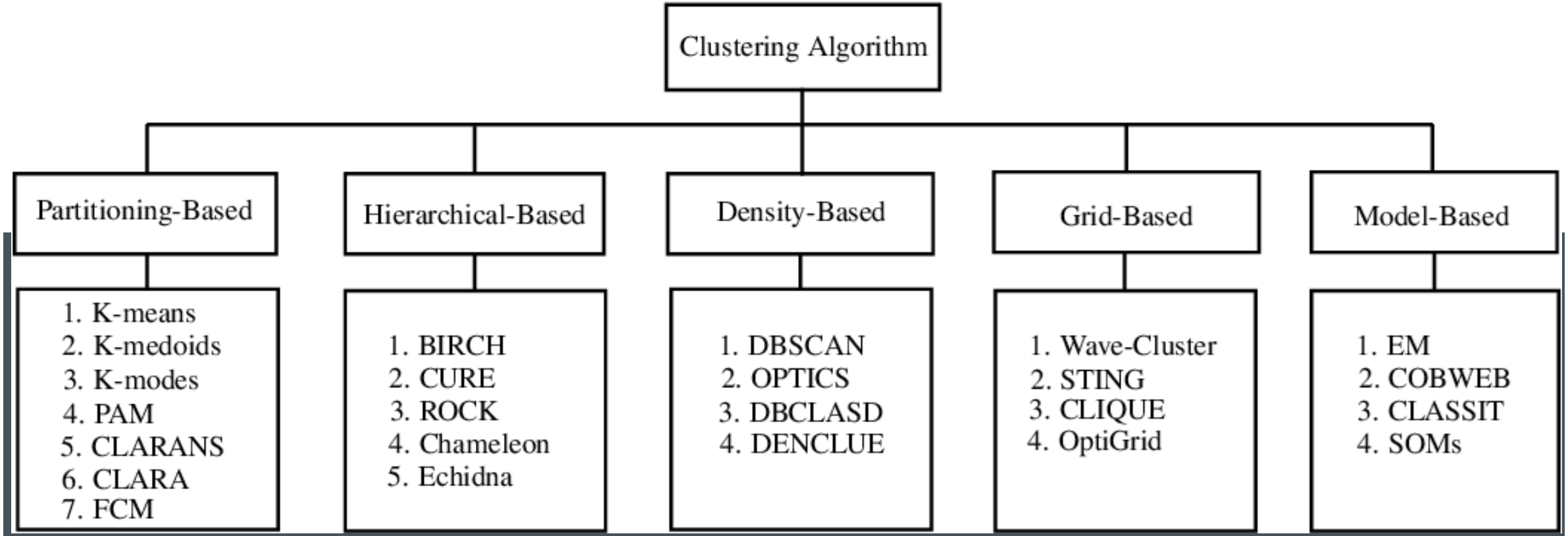


BAHASAN

- Clustering
- Regression

Top 10 Algorithms & Methods used by Data Scientists

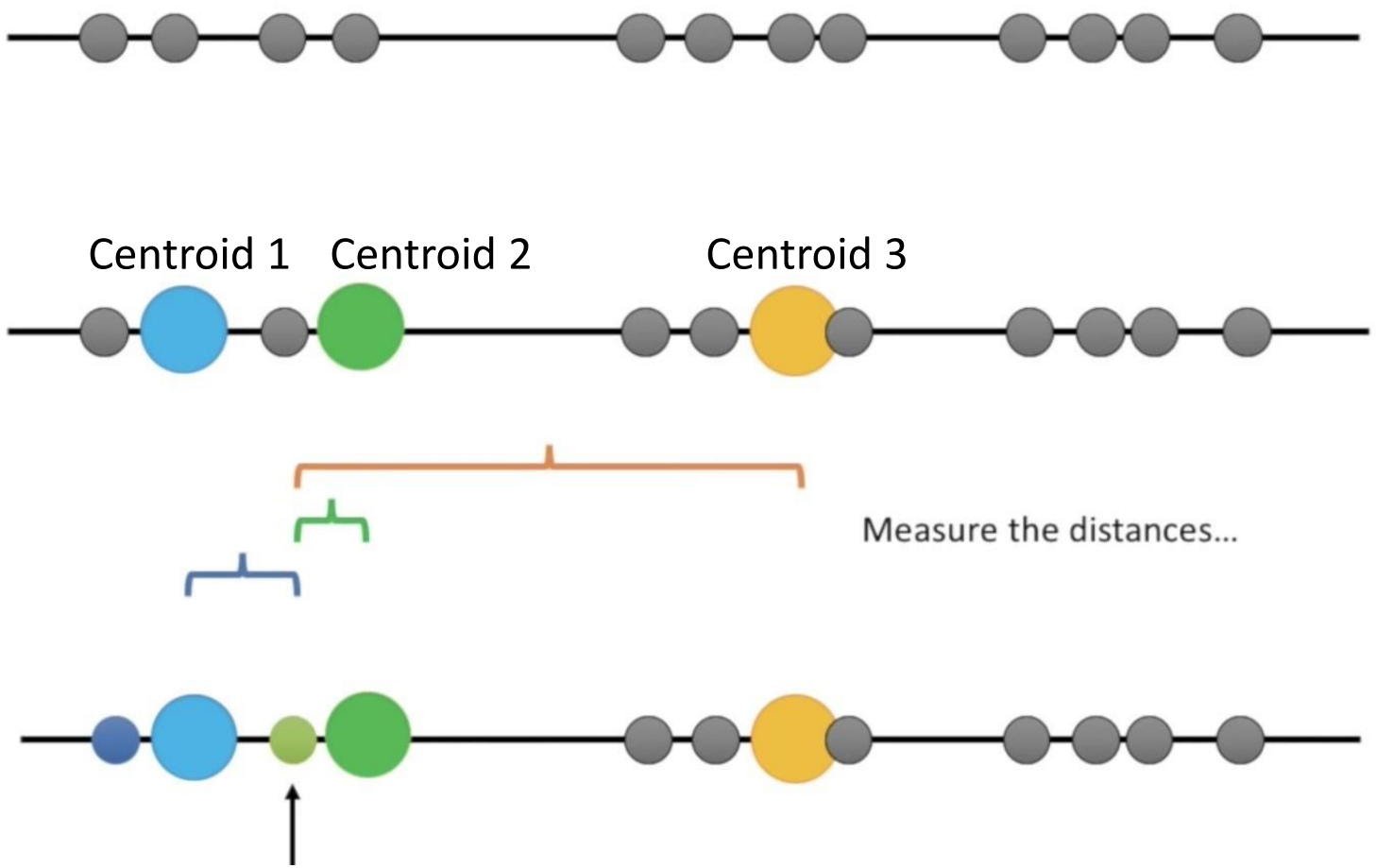



Clustering Algorithm
Partitioning-Based
Hierarchical-Based
Density-Based
Grid-Based
Model-Based
1. K-means
2. K-medoids
3. K-modes
4. PAM
5. CLARANS
6. CLARA
7. FCM
1. BIRCH
2. CURE
3. ROCK
4. Chameleon
5. Echidna
1. DBSCAN
2. OPTICS
3. DBCLASD
4. DENCLUE
1. Wave-Cluster
2. STING
3. CLIQUE
4. OptiGrid
1. EM
2. COBWEB
3. CLASSIT
4. SOMs

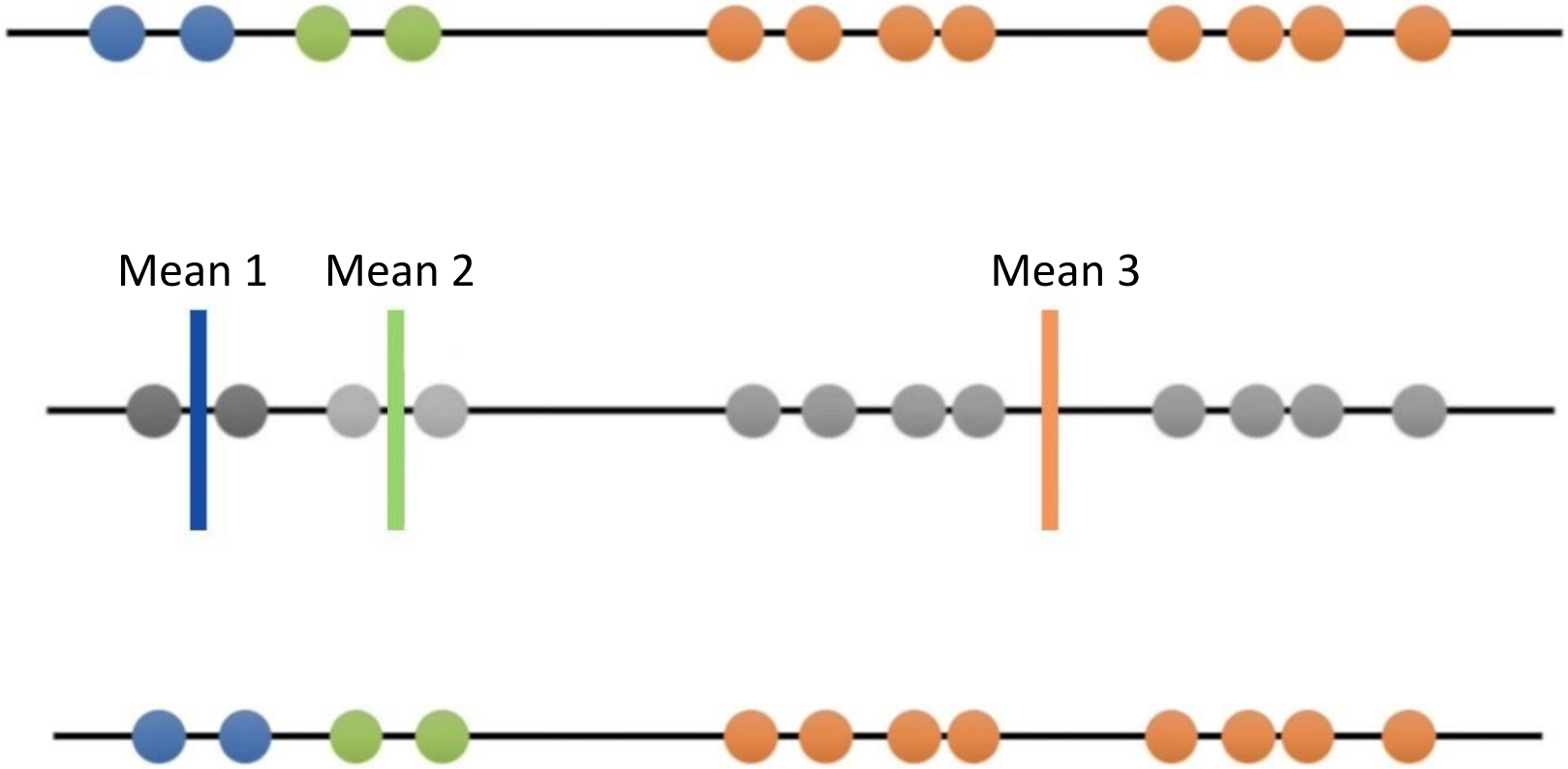
K-MEANS CLUSTERING

PROBLEM

Clustering dataset
Data 3D/2D → 1D



PROBLEM



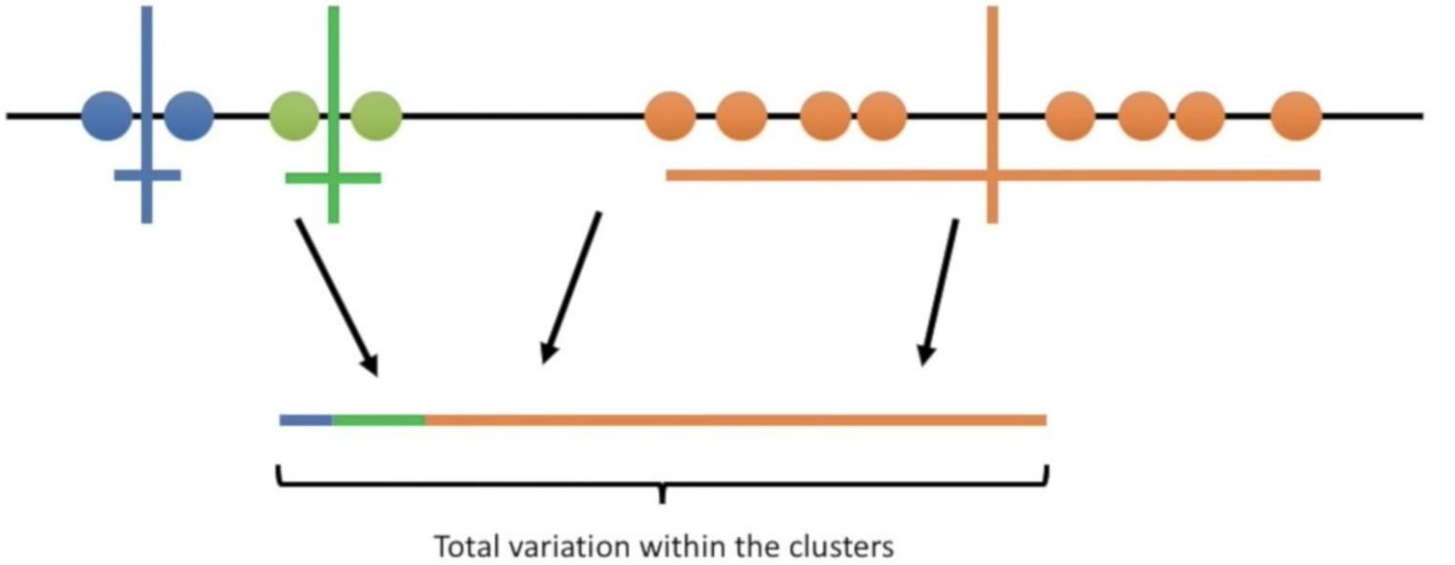
SOLUTION

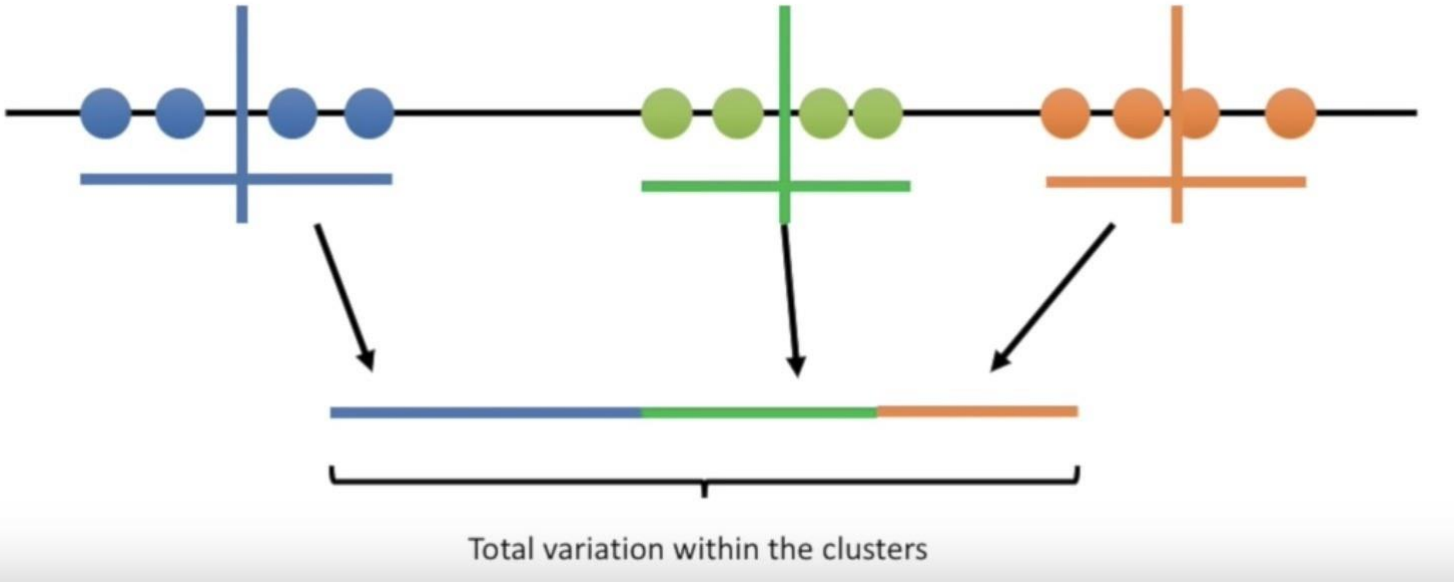
Iterasi centroid ke-2

Iterasi centroid ke-n

HASIL FINAL

SOLUTION


Total variation within the clusters


Total variation within the clusters

K-MEANS

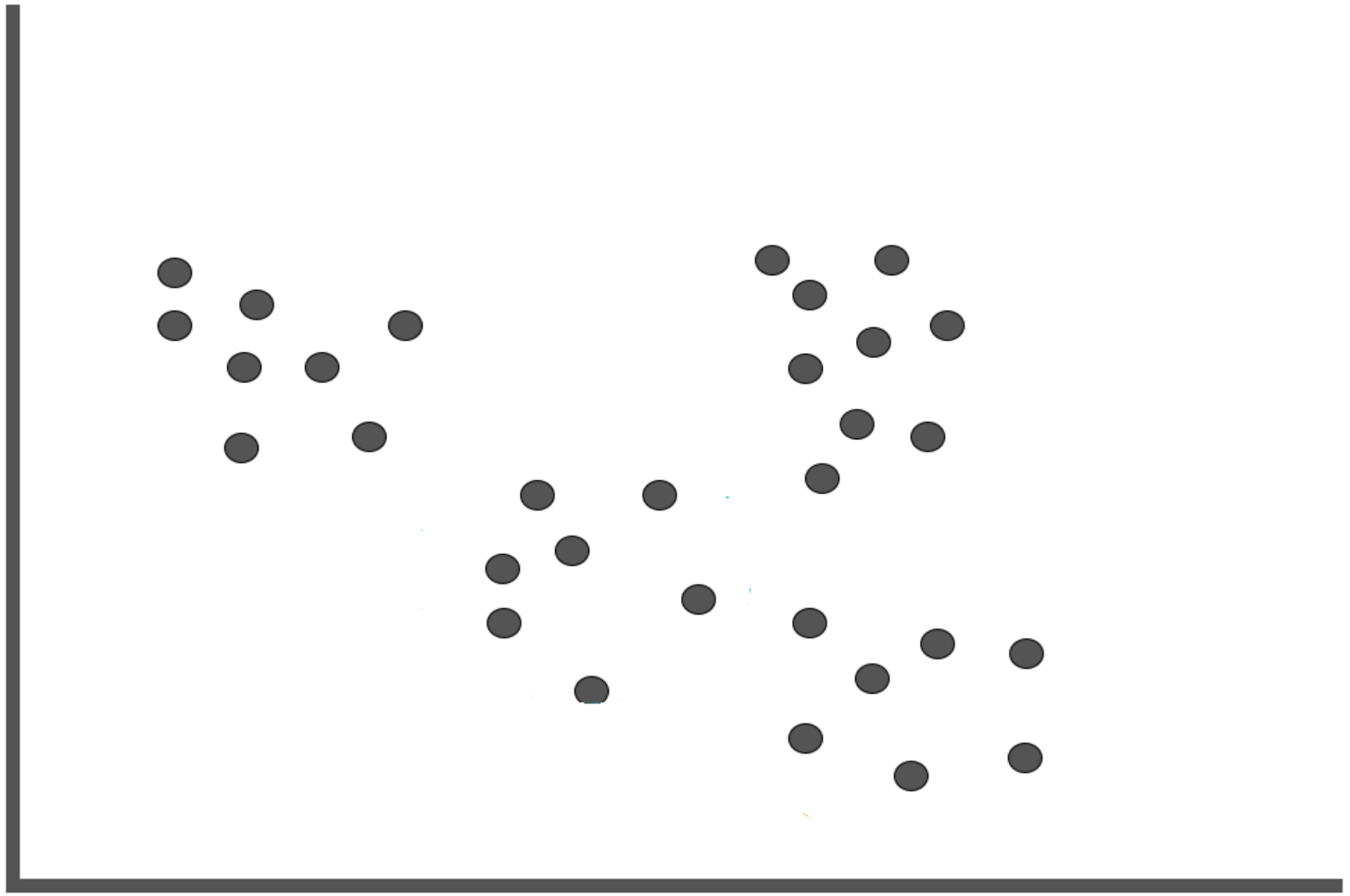
MENENTUKAN NILAI K

MENENTUKAN NILAI K

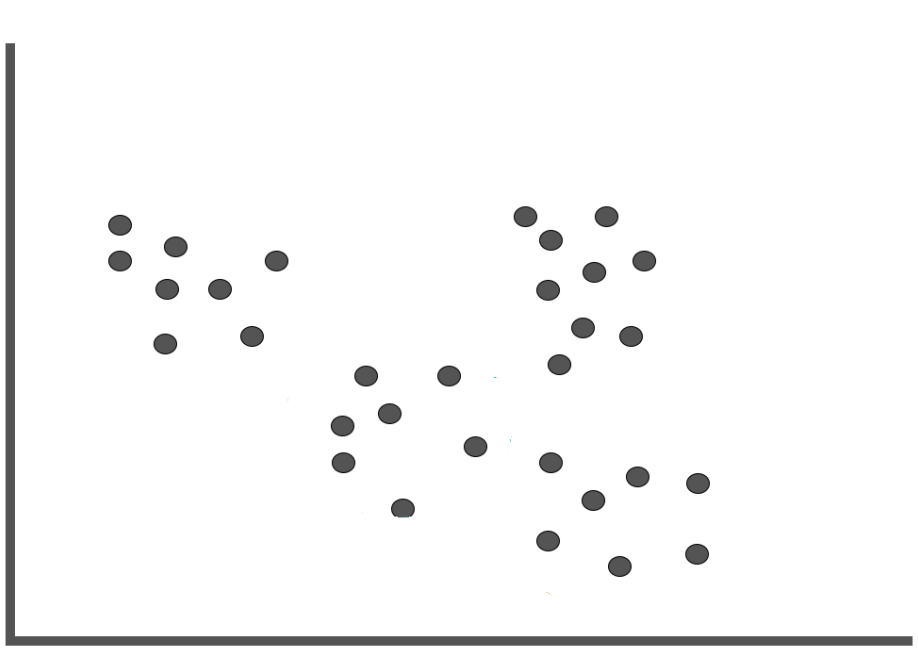
Cara 1. Secara Visual

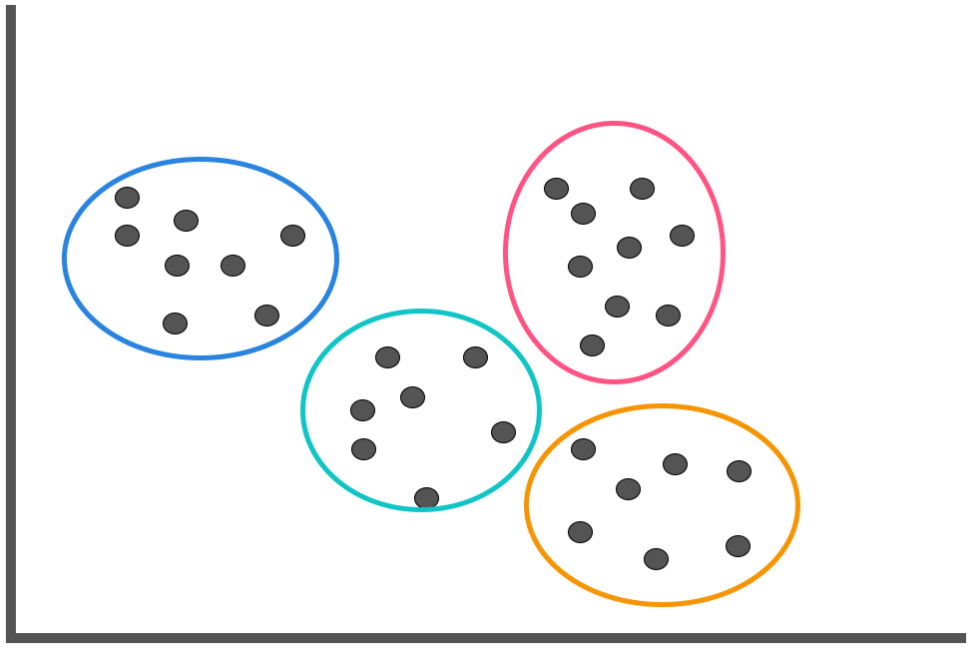
Cara paling mudah untuk menentukan jumlah K atau klaster pada K-means adalah dengan melihat langsung persebaran data. Otak kita bisa mengelompokkan data-data yang berdekatan dengan sangat cepat. Tetapi cara ini hanya bekerja dengan baik pada data yang sangat sederhana.

KASUS SEDERHANA



KASUS SEDERHANA



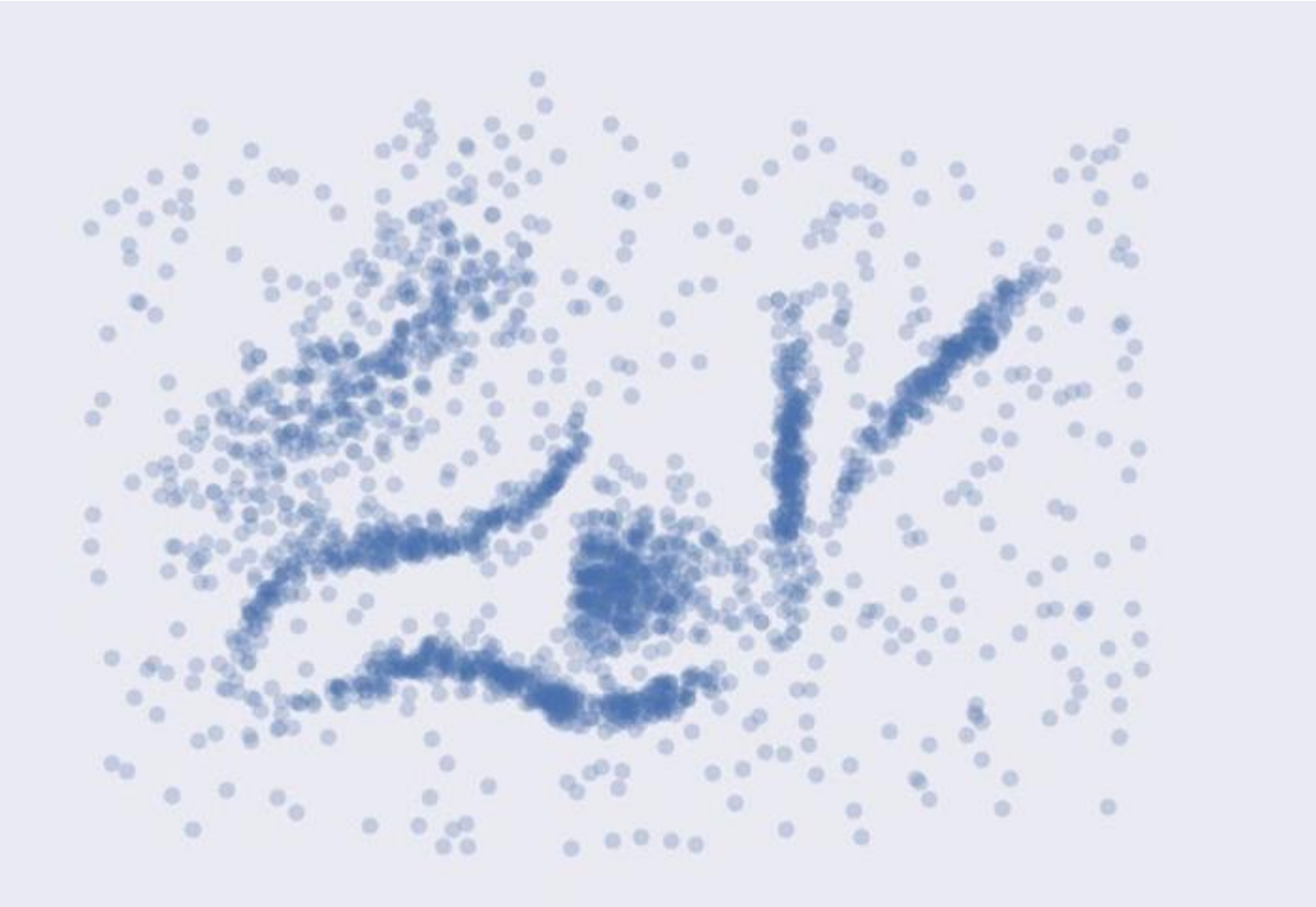


MENENTUKAN NILAI K

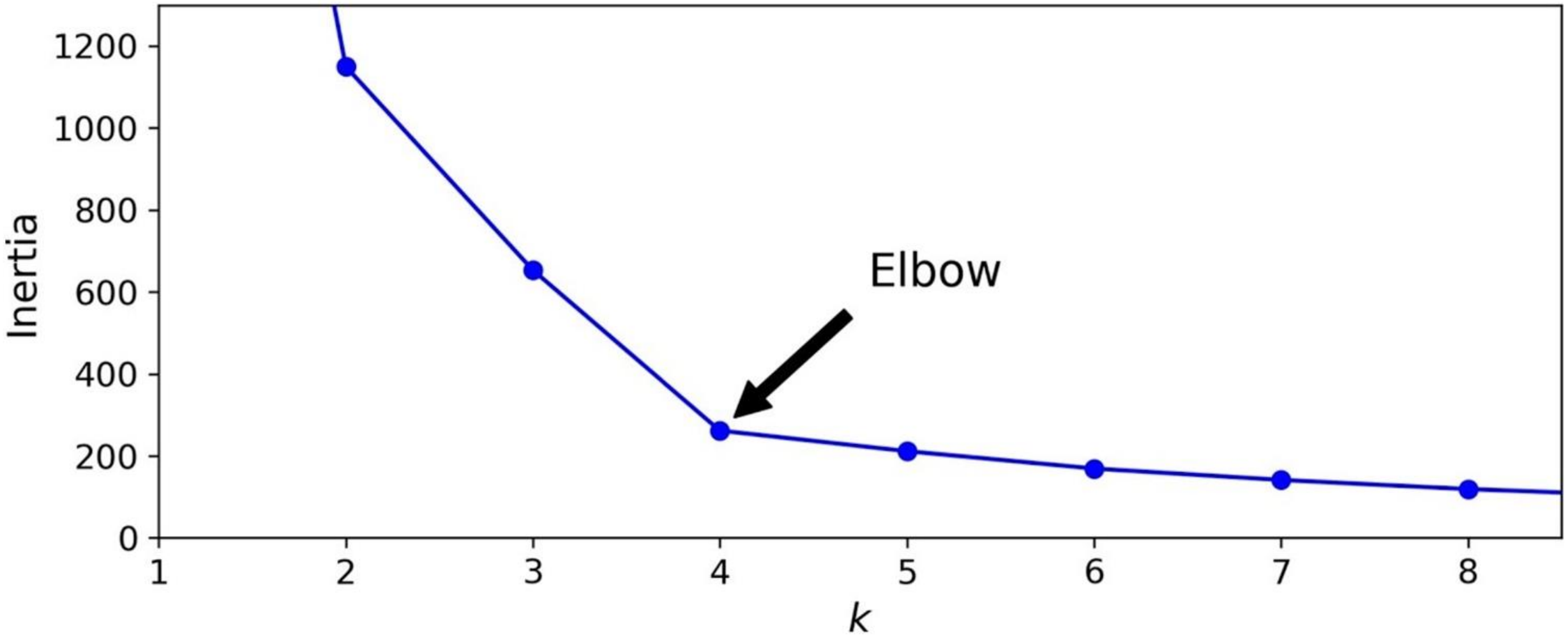
Cara 2. Metode Elbow

Ketika masalah clustering lebih kompleks, kita bisa menggunakan metode Elbow, yaitu menjalankan K-Means pada dataset dengan nilai K pada jarak tertentu (1,2,3, .., N). Kemudian hitung inersia pada setiap nilai K. Inersia memberi tahu seberapa jauh jarak setiap sampel pada sebuah klaster. Semakin kecil inersia maka semakin baik karena jarak setiap sampel pada sebuah klaster lebih berdekatan.

KASUS KOMPLEKS



KASUS KOMPLEKS



HYPERPARAMETER

Hyperparameters available for tuning

1. n_clusters=8 This is what you can and should change
2. max_iter=300 This determines the number of iterations

(Assign & Optimize moving the centroids)

1. n_init=10 Number of times to initialize the algorithm

ALGORITMA

1. Pick k random centroids from the dataset

$$\mu_1, \mu_2, \ldots, \mu_k, i = 1, \ldots, k$$

2. Compute the distances between each data (e.g., Euclidean)

$$d(x, \mu_i) = \|x - \mu_i\|^2$$

3. Assign each data point to the nearest cluster

$$c_i = j : d(x_j, \mu_i) \leq d(x_j, \mu_l), l \neq i, j = 1, 2, \ldots, n$$

4. Reposition the centroids by computing the mean

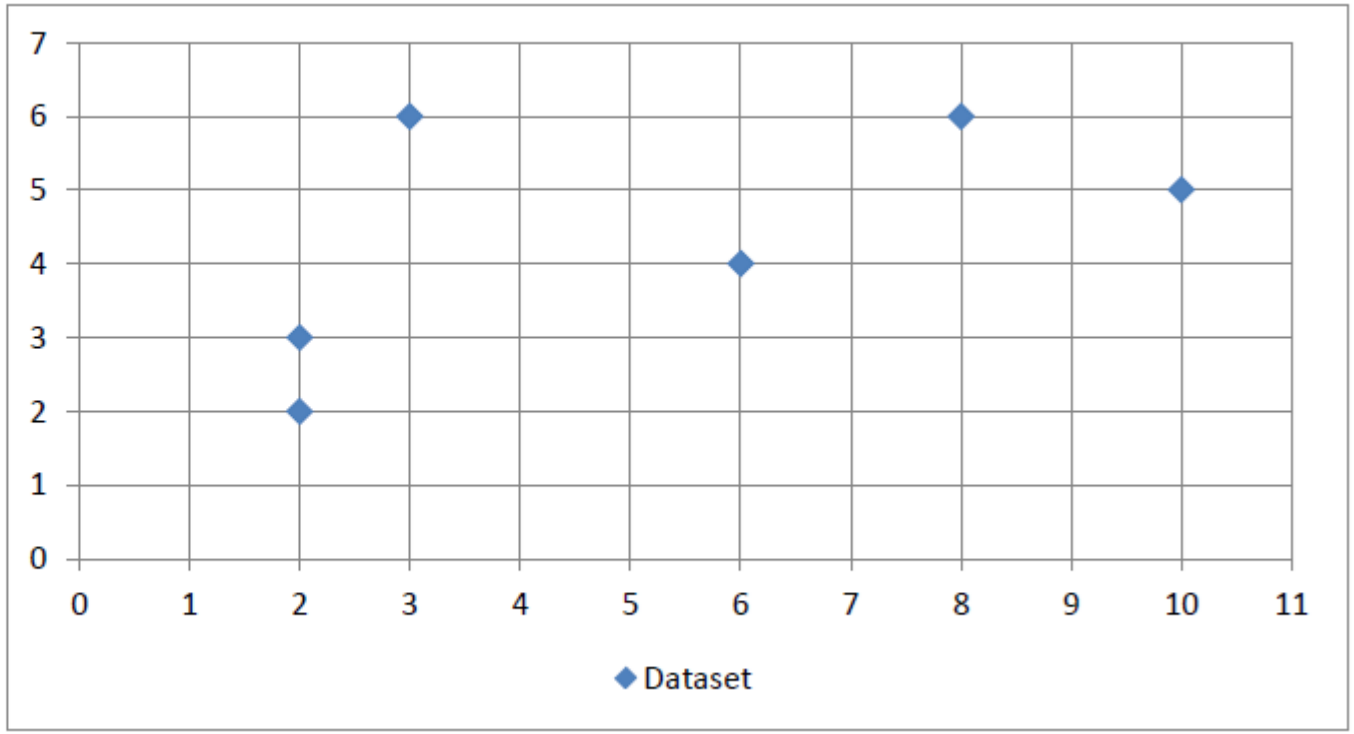
$$\mu_i = \frac{1}{|c_i|} \sum_{j \in c_i} x_j, \forall i$$

5. Repeat (2-4) until clusters become stable

MEMBUAT MODEL K-MEANS SECARA MANUAL

0. MEMVISUALISASIKAN DATASET

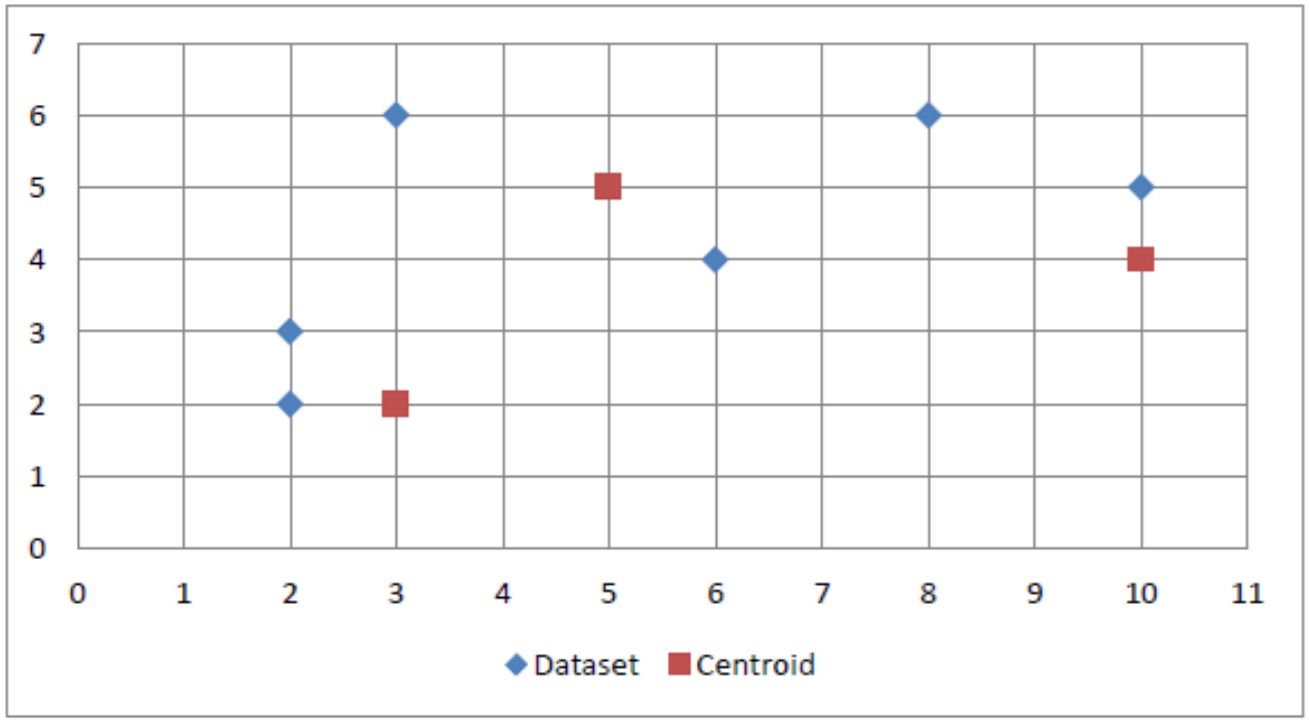
Data ke-	Dataset	
	X	**Y**
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5



1. MENENTUKAN CENTROID (SECARA RANDOM)

Data ke-	Dataset	
	X	Y
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5

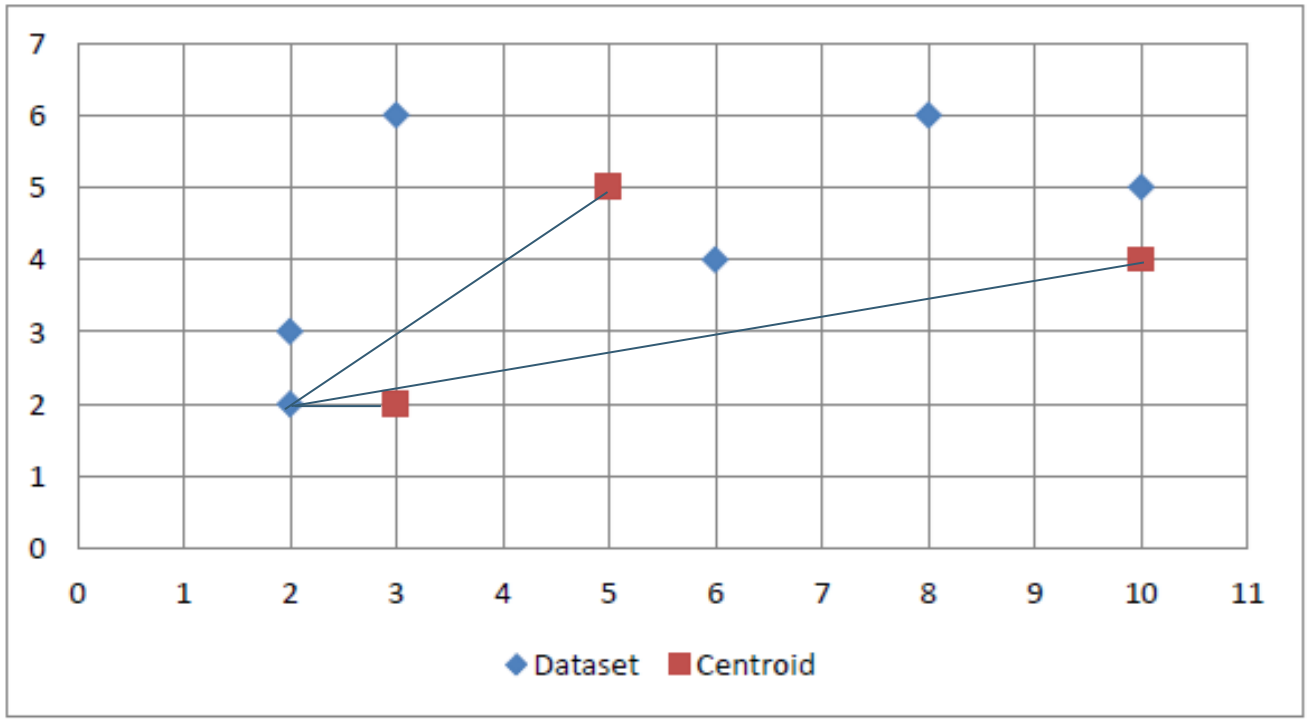
Centroid	
X	Y
3	2
5	5
10	4



2. MENGHITUNG JARAK DATA KE SETIAP CENTROID

Data ke-	Dataset	
	X	Y
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5

Centroid	
X	Y
3	2
5	5
10	4



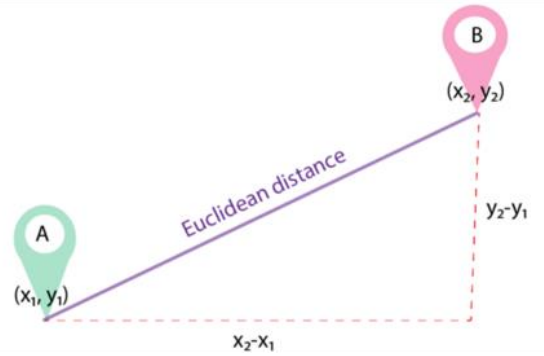
Menghitung jarak ke setiap centroid

Data ke-a

- Jarak ke C1: $\sqrt{(2-3)^2+(2-3)^2} = 1$
- Jarak ke C2: $\sqrt{(2-5)^2+(2-5)^2} = 4.24$
- Jarak ke C3: $\sqrt{(2-10)^2+(2-4)^2} = 8.2$

Euclidean distance (d) :

$$\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$



3. MEMASANG DATA KE CENTROID YANG COCOK

Data ke-	Dataset	
	X	Y
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5

Centroid	
X	Y
3	2
5	5
10	4

Menghitung jarak ke setiap centroid

Data ke-a

- Jarak ke C1: $\sqrt{(2-3)^2+(2-2)^2}=1$
- Jarak ke C2: $\sqrt{(2-5)^2+(2-5)^2}=4.24$
- Jarak ke C3: $\sqrt{(2-10)^2+(2-4)^2}=8.24$

C1

Data ke-b

- Jarak ke C1: $\sqrt{(2-3)^2+(3-2)^2}=1{,}41$
- Jarak ke C2: $\sqrt{(2-5)^2+(3-5)^2}=3.60$
- Jarak ke C3: $\sqrt{(2-10)^2+(3-4)^2}=8.06$

C1

Data ke-c

- Jarak ke C1: $\sqrt{(3-3)^2+(6-2)^2}=4$
- Jarak ke C2: $\sqrt{(3-5)^2+(6-5)^2}=2{,}23$
- Jarak ke C3: $\sqrt{(3-10)^2+(6-4)^2}=7{,}28$

C2

Data ke-d

- Jarak ke C1: $\sqrt{(6-3)^2+(4-2)^2}=3{,}60$
- Jarak ke C2: $\sqrt{(6-5)^2+(4-5)^2}=1{,}41$
- Jarak ke C3: $\sqrt{(6-10)^2+(4-4)^2}=4$

C2

Data ke-e

- Jarak ke C1: $\sqrt{(8-3)^2+(6-2)^2}=6{,}40$
- Jarak ke C2: $\sqrt{(8-5)^2+(6-5)^2}=3{,}16$
- Jarak ke C3: $\sqrt{(8-10)^2+(6-4)^2}=2{,}82$

C3

Data ke-f

- Jarak ke C1: $\sqrt{(10-3)^2+(5-2)^2}=7{,}61$
- Jarak ke C2: $\sqrt{(10-5)^2+(5-5)^2}=5$
- Jarak ke C3: $\sqrt{(10-10)^2+(5-4)^2}=1$

C3

4. MEMPERBARUI CENTROID (ITERASI 2)

Data ke-	Dataset	
	X	Y
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5

Centroid	
X	Y
3	2
5	5
10	4

New Centroid	
X	Y
2	2.5
4.5	5
9	5.5

NC 1

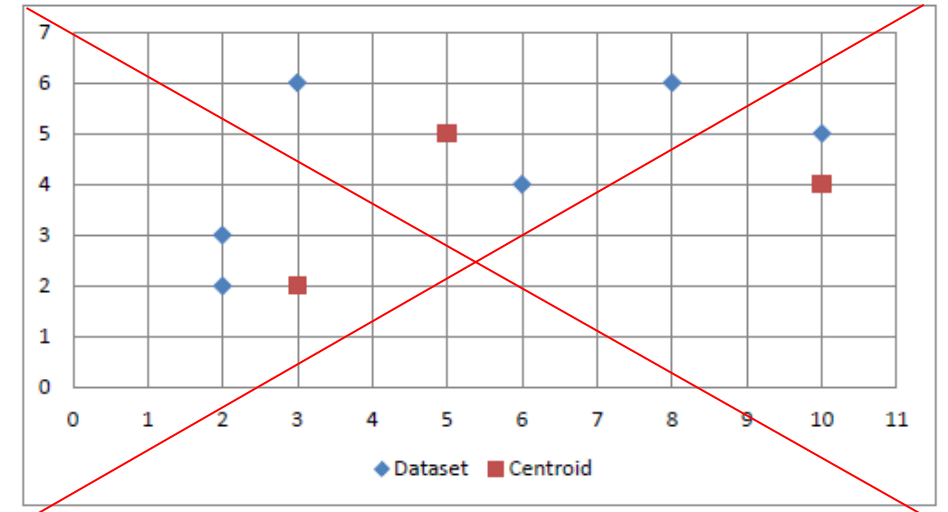
$$X = \frac{\sum X}{N} = \frac{2+2}{2} = 2$$
$$Y = \frac{\sum Y}{N} = \frac{2+3}{2} = 2{,}5$$

NC 2

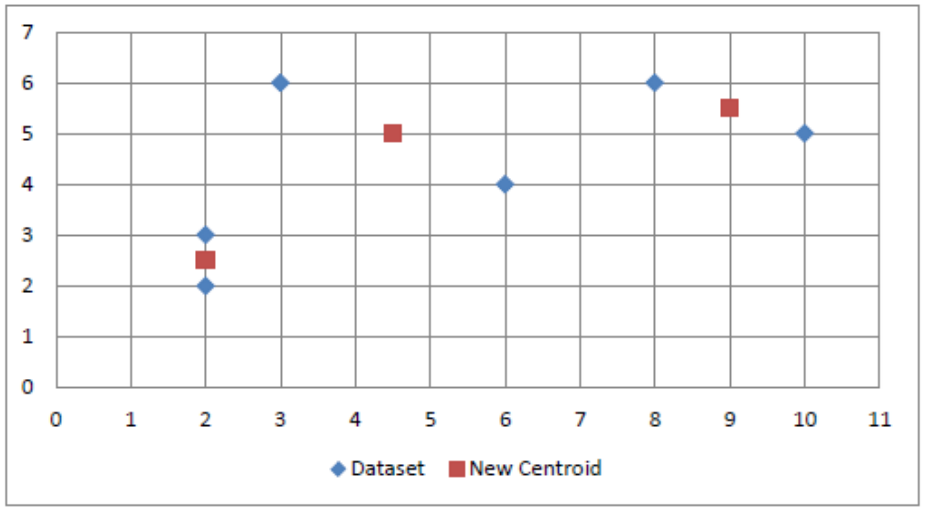
$$X = \frac{\sum X}{N} = \frac{3+6}{2} = 4{,}5$$
$$Y = \frac{\sum Y}{N} = \frac{6+4}{2} = 5$$

NC 3

$$X = \frac{\sum X}{N} = \frac{8+10}{2} = 9$$
$$Y = \frac{\sum Y}{N} = \frac{6+5}{2} = 5{,}5$$



OLD



NEW

5. MENENTUKAN CENTROID YANG COCOK (LAGI)
. MEMPERBARUI CENTROID (LAGI)
....

C1 = data a, data b	OC 1 to NC 1
C2 = data c, data d	OC 2 to NC 2
C3 = data e, data f	OC 3 to NC 3

. MENENTUKAN CENTROID YANG COCOK (ITERASI 2)

Data ke-	Dataset	
	X	Y
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5

New Centroid	
X	Y
2	2.5
4.5	5
9	5.5

Pembagian Klaster Sama Saja

Menghitung jarak ke setiap centroid

Data ke-a

- Jarak ke C1: $\sqrt{(2-2)^2+(2-2.5)^2}=0.50$
- Jarak ke C2: $\sqrt{(2-4.5)^2+(2-5)^2}=3.90$
- Jarak ke C3: $\sqrt{(2-9)^2+(2-5.5)^2}=7.82$

C1

Data ke-b

- Jarak ke C1: $\sqrt{(2-2)^2+(3-2.5)^2}=0.50$
- Jarak ke C2: $\sqrt{(2-4.5)^2+(3-5)^2}=3.20$
- Jarak ke C3: $\sqrt{(2-9)^2+(3-5.5)^2}=7.43$

C1

Data ke-c

- Jarak ke C1: $\sqrt{(3-2)^2+(6-2.5)^2}=3.64$
- Jarak ke C2: $\sqrt{(3-4.5)^2+(6-5)^2}=2.23$
- Jarak ke C3: $\sqrt{(3-9)^2+(6-5.5)^2}=6.18$

C2

Data ke-d

- Jarak ke C1: $\sqrt{(6-2)^2+(4-2.5)^2}=3.60$
- Jarak ke C2: $\sqrt{(6-4.5)^2+(4-5)^2}=2.46$
- Jarak ke C3: $\sqrt{(6-9)^2+(4-5.5)^2}=4.27$

C2

Data ke-e

- Jarak ke C1: $\sqrt{(8-2)^2+(6-2.5)^2}=6.94$
- Jarak ke C2: $\sqrt{(8-4.5)^2+(6-5)^2}=3.64$
- Jarak ke C3: $\sqrt{(8-9)^2+(6-5.5)^2}=1.80$

C3

Data ke-f

- Jarak ke C1: $\sqrt{(10-2)^2+(5-2.5)^2}=8.73$
- Jarak ke C2: $\sqrt{(10-4.5)^2+(5-5)^2}=5.50$
- Jarak ke C3: $\sqrt{(10-9)^2+(5-5.5)^2}=1.11$

C3

. MEMPERBARUI CENTROID (ITERASI 3)

Data ke-	Dataset	
	X	Y
a	2	2
b	2	3
c	3	6
d	6	4
e	8	6
f	10	5

New Centroid	
X	Y
2	2.5
4.5	5
9	5.5

New Centroid	
X	Y
2	2.5
4.5	5
9	5.5

NC 1

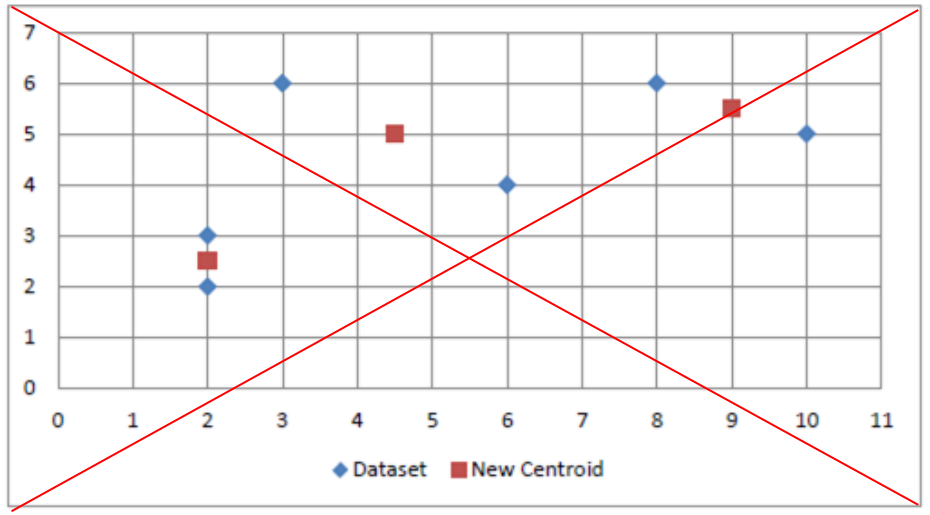
$$X = \frac{\sum X}{N} = \frac{2+2}{2} = 2$$
$$Y = \frac{\sum Y}{N} = \frac{2+3}{2} = 2{,}5$$

NC 2

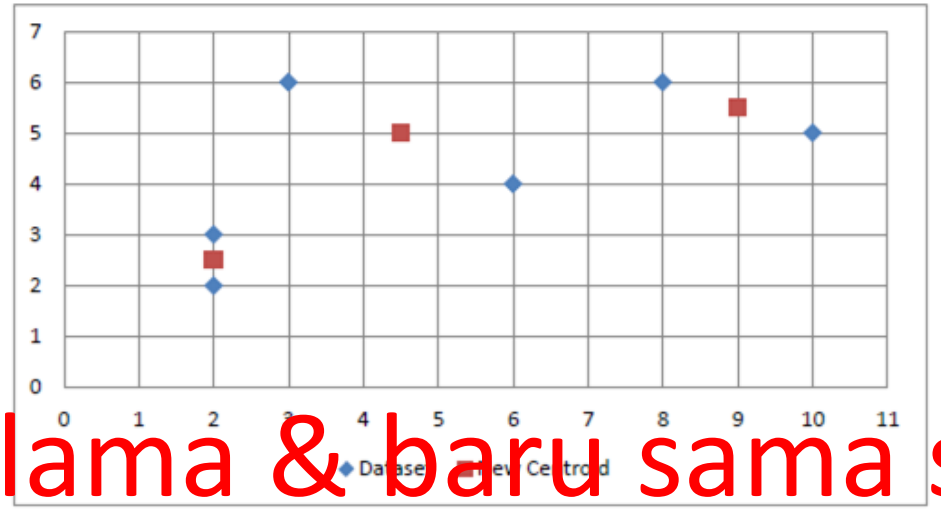
$$X = \frac{\sum X}{N} = \frac{3+6}{2} = 4{,}5$$
$$Y = \frac{\sum Y}{N} = \frac{6+4}{2} = 5$$

NC 3

$$X = \frac{\sum X}{N} = \frac{8+10}{2} = 9$$
$$Y = \frac{\sum Y}{N} = \frac{6+5}{2} = 5{,}5$$



OLD



NEW

Centroid lama & baru sama saja

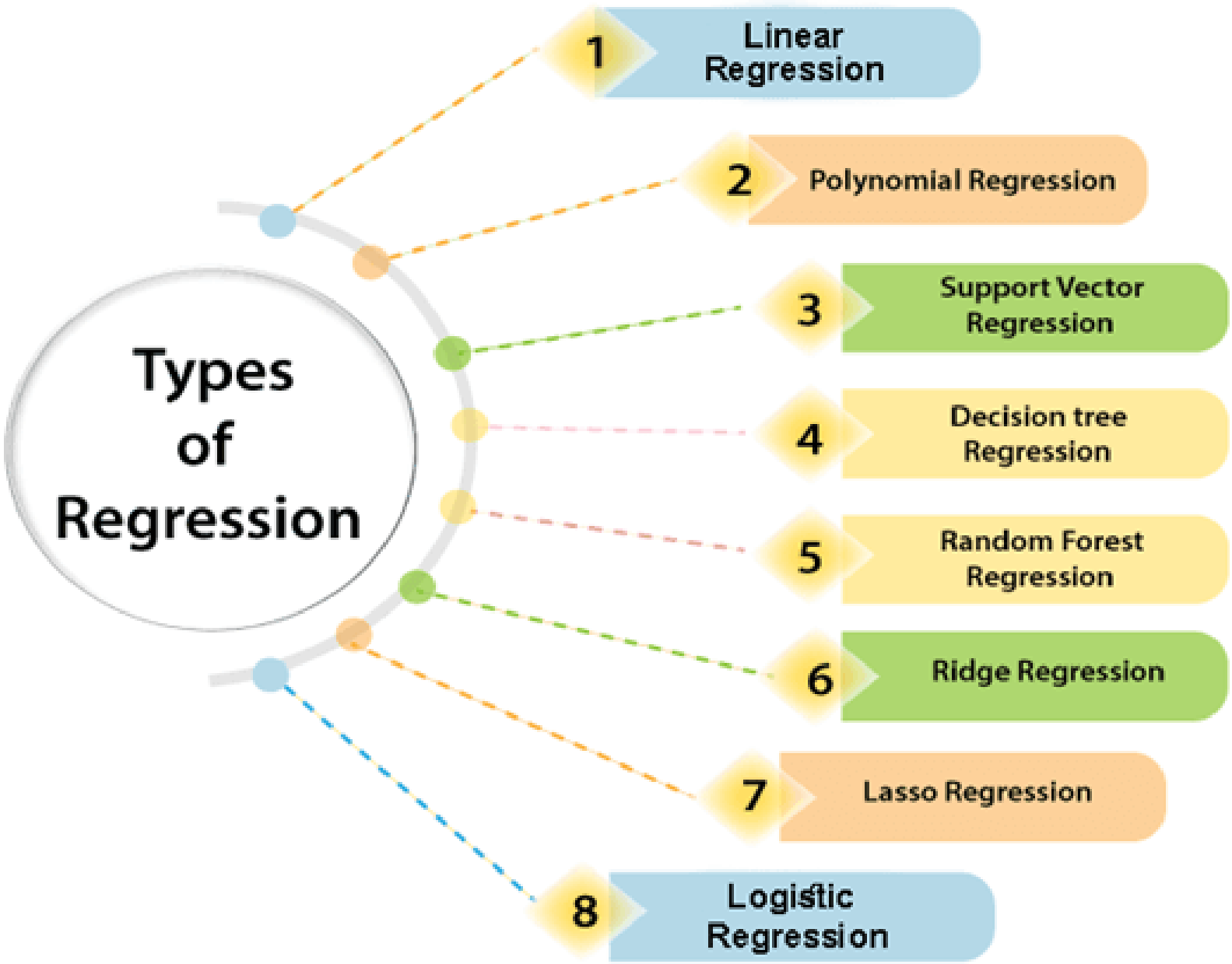
MEMBUAT MODEL K-MEANS DENGAN PYTHON

MEMBUAT MODEL ML

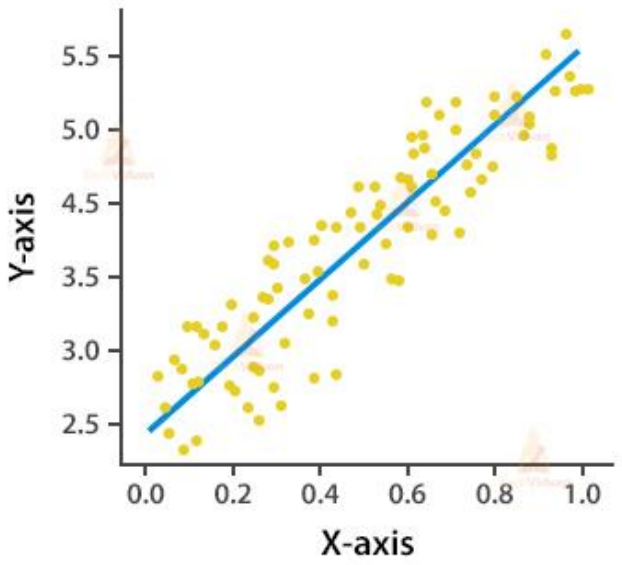
```
from sklearn.decomposition import PCA

pca = PCA(n_components = 2)
X2D = pca.fit_transform(X)
```

```
from sklearn.cluster import KMeans
km5 = KMeans(n_clusters=5)
km5.fit(X)
```

LINEAR REGRESSION & LOGISTIC REGRESSION

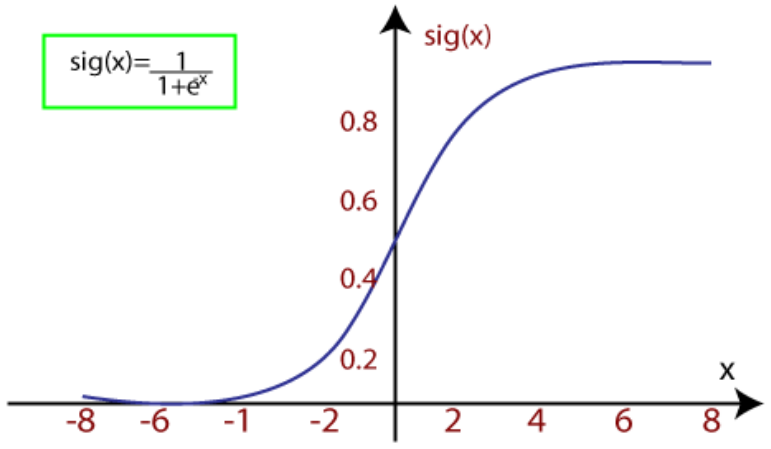

Types of Regression
1 Linear Regression
2 Polynomial Regression
3 Support Vector Regression
4 Decision tree Regression
5 Random Forest Regression
6 Ridge Regression
7 Lasso Regression
8 Logistic Regression

Linear Regression

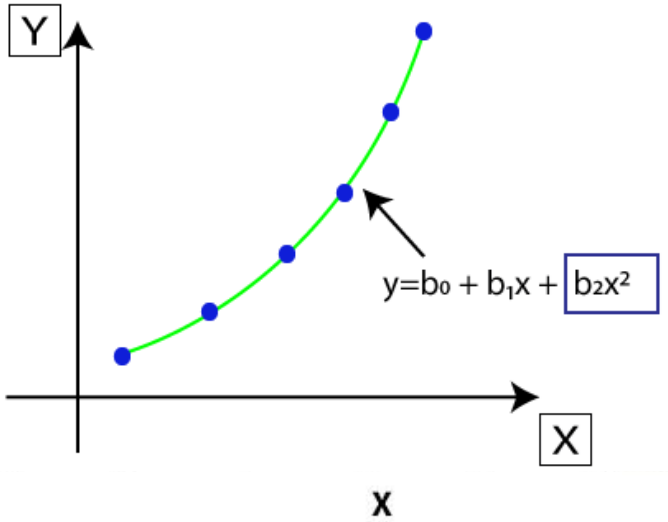


Logistic Regression

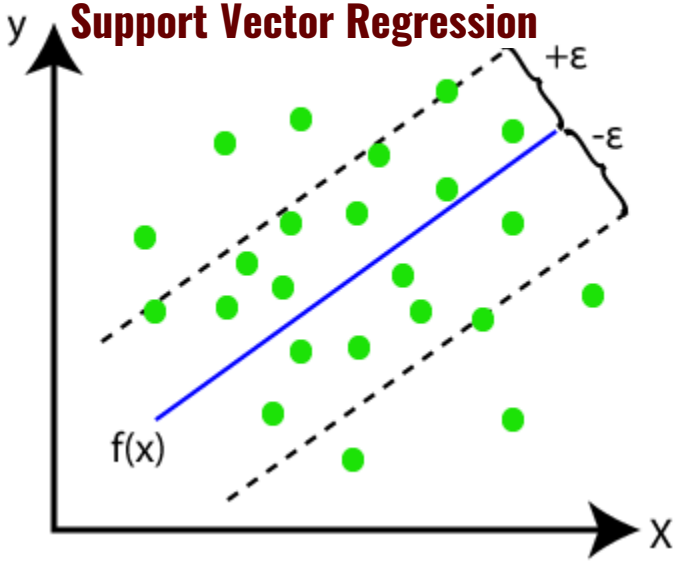
$$sig(x) = \frac{1}{1+e^{x}}$$



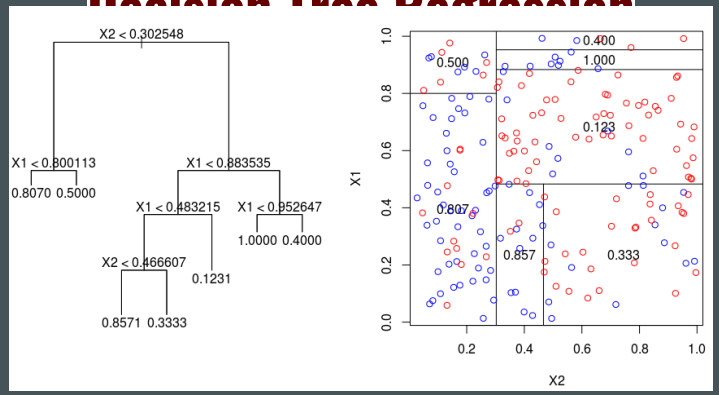
Polynomial Regression



Support Vector Regression



Decision Tree Regression



Lasso Regression

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

ElasticNet Regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

Lama Bekerja	Industri	Tingkat Pendidikan	Gaji
6 tahun	Marketing	SMA	8.000.000
12 tahun	IT	S1	16.000.000
8 tahun	Kesehatan	S2	20.000.000
5 tahun	IT	SMK	?
6 tahun	Marketing	S2	14.000.000
21 tahun	Perbankan	S3	35.000.000
3 tahun	IT	S1	10.000.000

REGRESI

Lama Bekerja	Industri	Tingkat Pendidikan	Gaji
6 tahun	Marketing	SMA	8.000.000
12 tahun	IT	S1	16.000.000
8 tahun	Kesehatan	S2	20.000.000
5 tahun	IT	SMK	?
6 tahun	Marketing	S2	14.000.000
21 tahun	Perbankan	S3	35.000.000
3 tahun	IT	S1	10.000.000

Regresi adalah salah satu teknik ML yang mirip dengan klasifikasi. Bedanya pada klasifikasi, sebuah model ML memprediksi sebuah kelas, sedangkan model regresi memprediksi bilangan kontinu. Bilangan kontinu adalah bilangan numerik.

Jadi model klasifikasi memprediksi kelas atau kategori dan model regresi memprediksi sebuah nilai berdasarkan atribut yang tersedia. Agar lebih paham, perhatikan contoh di samping.

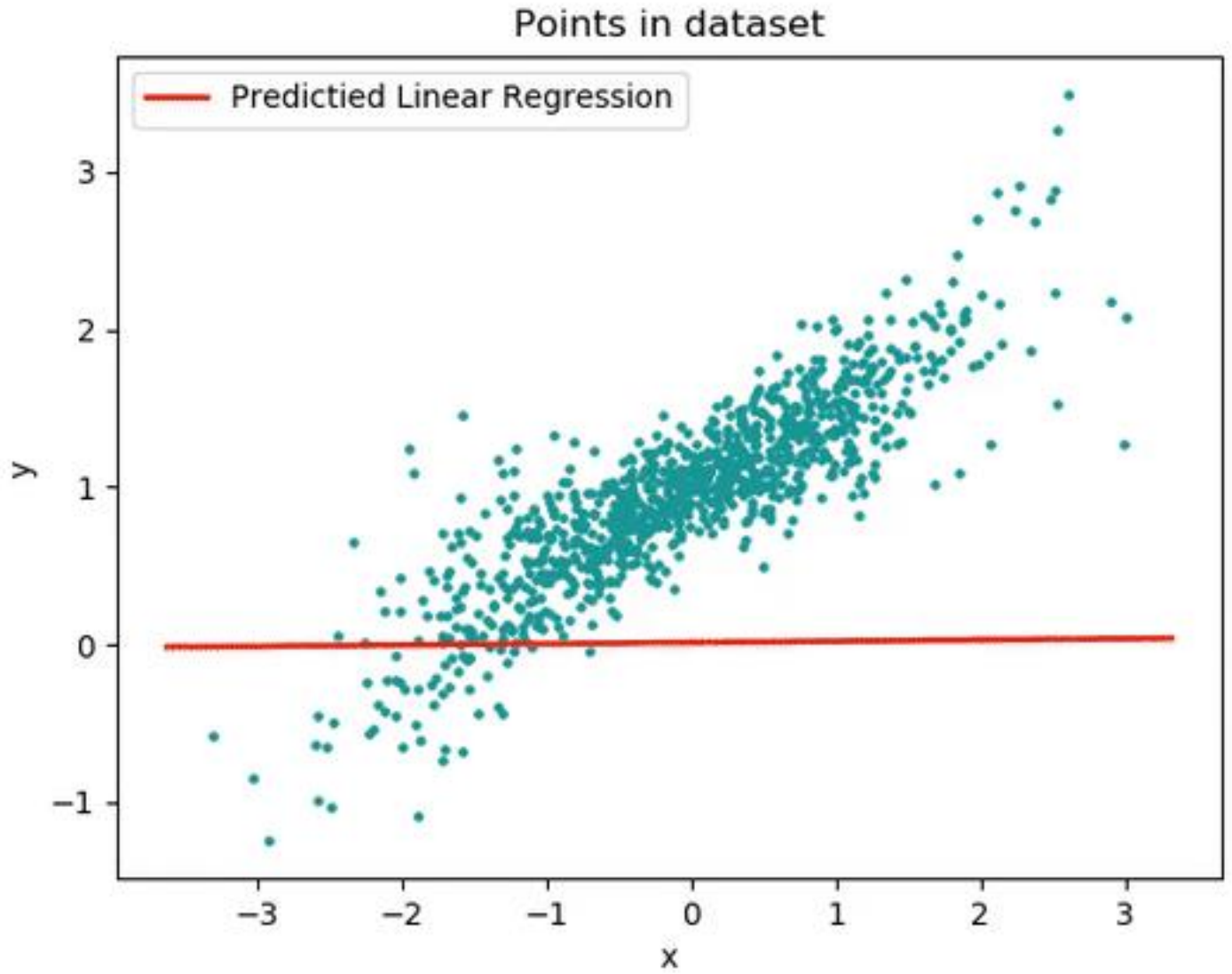
Pada contoh data di samping, model regresi akan memprediksi gaji berdasarkan atribut lama bekerja, industri, dan tingkat pendidikan. Gaji adalah contoh dari bilangan kontinu, di mana gaji tak memiliki kategori-kategori yang terbatas.

REGRESI

Lama Bekerja	Industri	Tingkat Pendidikan	Gaji
6 tahun	Marketing	SMA	8.000.000
12 tahun	IT	S1	16.000.000
8 tahun	Kesehatan	S2	20.000.000
5 tahun	IT	SMK	?
6 tahun	Marketing	S2	14.000.000
21 tahun	Perbankan	S3	35.000.000
3 tahun	IT	S1	10.000.000

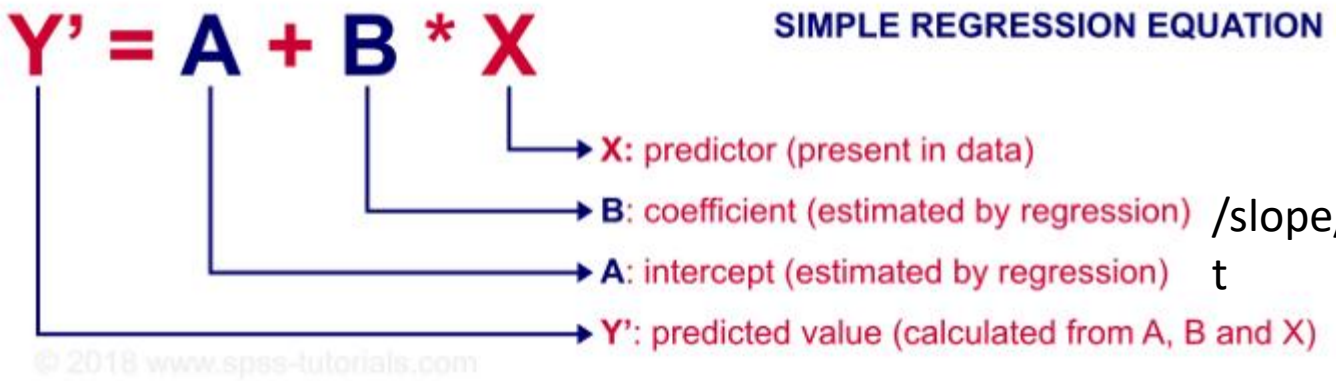
Regresi linier adalah salah satu metode supervised yang masuk dalam golongan regression, sesuai namanya. Contoh paling terkenal dari regresi linier adalah memperkirakan harga rumah berdasarkan fitur yang terdapat pada rumah seperti luas rumah, jumlah kamar tidur, lokasi dan sebagainya. Ini adalah model paling sederhana yang perlu diketahui guna memahami metode machine learning lain yang lebih kompleks.

Regresi linier cocok dipakai ketika terdapat hubungan linear pada data. Namun untuk implementasi pada kebanyakan kasus, ia kurang direkomendasikan. Sebabnya, regresi linier selalu mengasumsikan ada hubungan linier pada data, padahal tidak.


Points in dataset
Predictied Linear Regression
y
x
3
2
1
0
−1
−3
−2
−1
0
1
2
3

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$



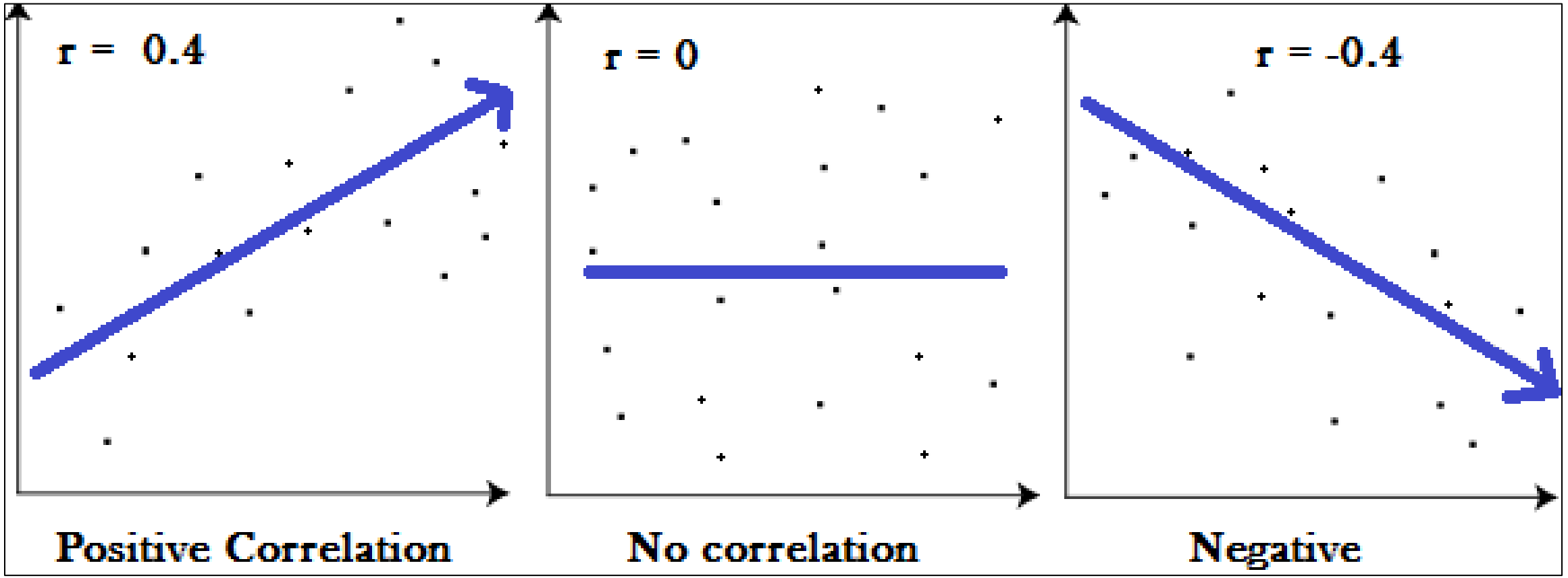
/slope/gradien t

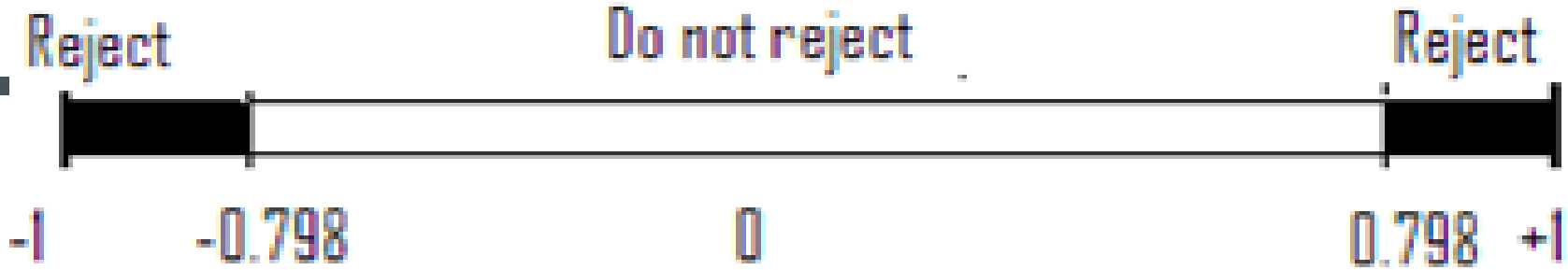
$$Y(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$$

CORRELATION

Pearson correlation coefficient

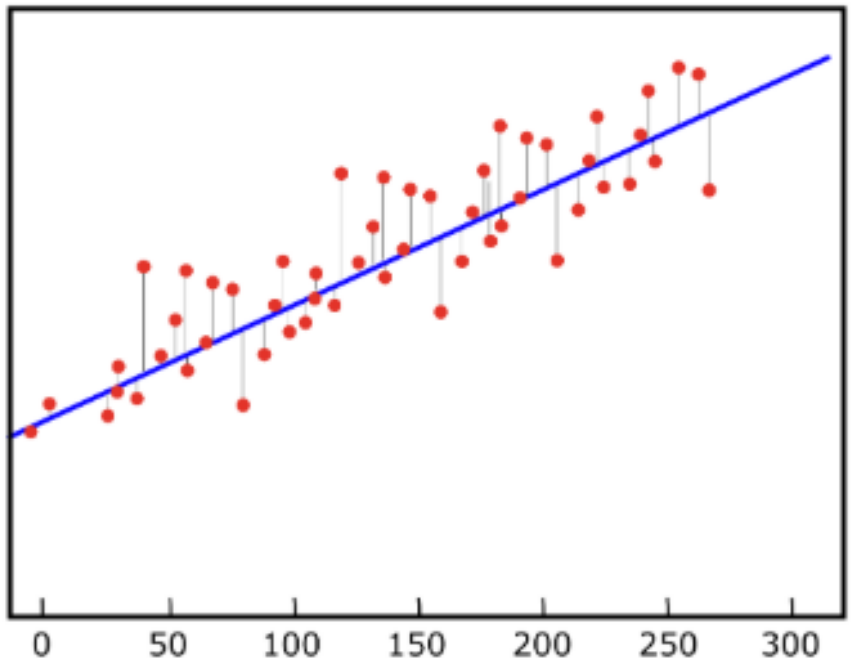
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$





r value =	
+0.70 or higher	Very strong positive relationship
+0.40 to +0.69	Strong positive relationship
+0.30 to +0.39	Moderate positive relationship
+0.20 to +0.29	weak positive relationship
+0.01 to +0.19	No or negligible relationship
0	No relationship [zero correlation]
-0.01 to -0.19	No or negligible relationship
-0.20 to -0.29	weak negative relationship
-0.30 to -0.39	Moderate negative relationship
-0.40 to -0.69	Strong negative relationship
-0.70 or higher	Very strong negative relationship

A. REGRESI LINEAR

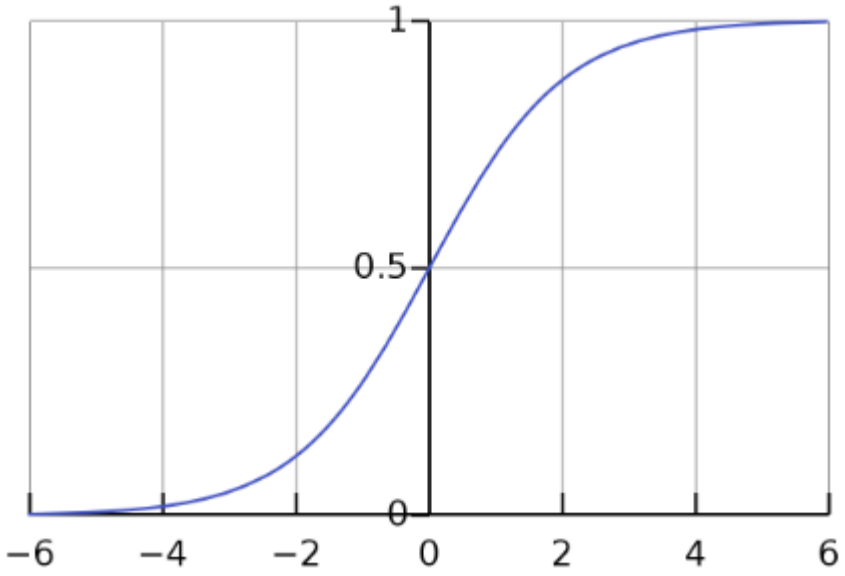


1. Secara sederhana regresi linear adalah teknik untuk memprediksi sebuah nilai dari variable Y (variabel dependen) berdasarkan beberapa variabel tertentu X (variabel independen) jika terdapat hubungan linier antara X dan Y.

1. Hubungan antara hubungan linier dapat direpresentasikan dengan sebuah garis lurus (disebut garis regresi).

1. Sebab garis regresi adalah sebuah model probabilistik dan prediksi kita adalah perkiraan,maka tentu akan ada eror/penyimpangan terhadap nilai asli dari variabel Y. Pada gambar di bawah, garis merah yang menghubungkan data-data ke garis regresi merupakan eror. Semakin banyak eror artinya model regresi itu belum optimal.

B. REGRESI LOGISTIK



Sesuai namanya, logistic regression menggunakan fungsi logistik untuk menghitung probabilitas kelas dari sebuah sampel.

Contohnya sebuah email memiliki probabilitas 78% merupakan spam maka email tersebut termasuk dalam kelas spam. Dan jika sebuah email memiliki <50% probabilitas merupakan spam, maka email tersebut diklasifikasikan bukan spam.

GOOGLE COLAB LINK LIST

- https://colab.research.google.com/drive/1kJtcN1YWT110XwJO9MyNmg-AiI3xZ_Be?usp=sharing
- https://colab.research.google.com/drive/1H-AXNiicrZW37W4Yh7vzZESm53YDtkvi?usp=sharing

REFERENSI

Acton, F. S. Analysis of Straight-Line Data. New York: Dover, 1966.

Edwards, A. L. “The Correlation Coefficient.” Ch. 4 in An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, pp. 33-46, 1976.

Gonick, L. and Smith, W. “Regression.” Ch. 11 in The Cartoon Guide to Statistics. New York: Harper Perennial, pp. 187-210, 1993

https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1

https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

https://dev.to/ashuto7h/p3-linear-regression-242

https://techvidvan.com/tutorials/types-of-regression/

https://www.javatpoint.com/regression-analysis-in-machine-learning

https://gdcoder.com/decision-tree-regressor-explained-in-depth/

TUGAS MINGGU 4

Buatlah program singkat regresi di google colaboratory(Kasus dataset bebas)

Setiap baris statement diberi komentar

DL: 8 Oktober Pukul 23.59

TERIMA KASIH