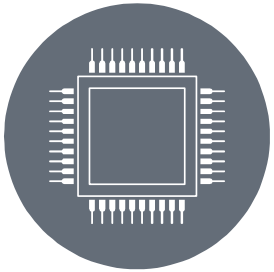
Penelitian Bidang Machine Learning

Materi 6
Metodologi Penelitian
Dosen:
Roni Salambue, S.Kom., M.Si.[1]
DR. Rahmad Kurniawan, ST., MIT., MTA., CISDV.[2]

Capaian Pembelajaran

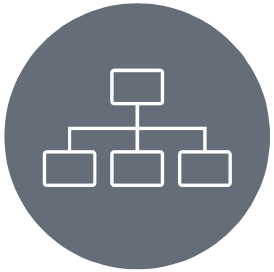
1

Mahasiswa memahami Model Penelitian Machine Learning



2

Mahasiswa mampu membuat Rumusan Masalah dan Hipotesis Penelitian



3

Mahasiswa mampu menyelesaikan masalah penelitian dengan algoritma Machine Learning

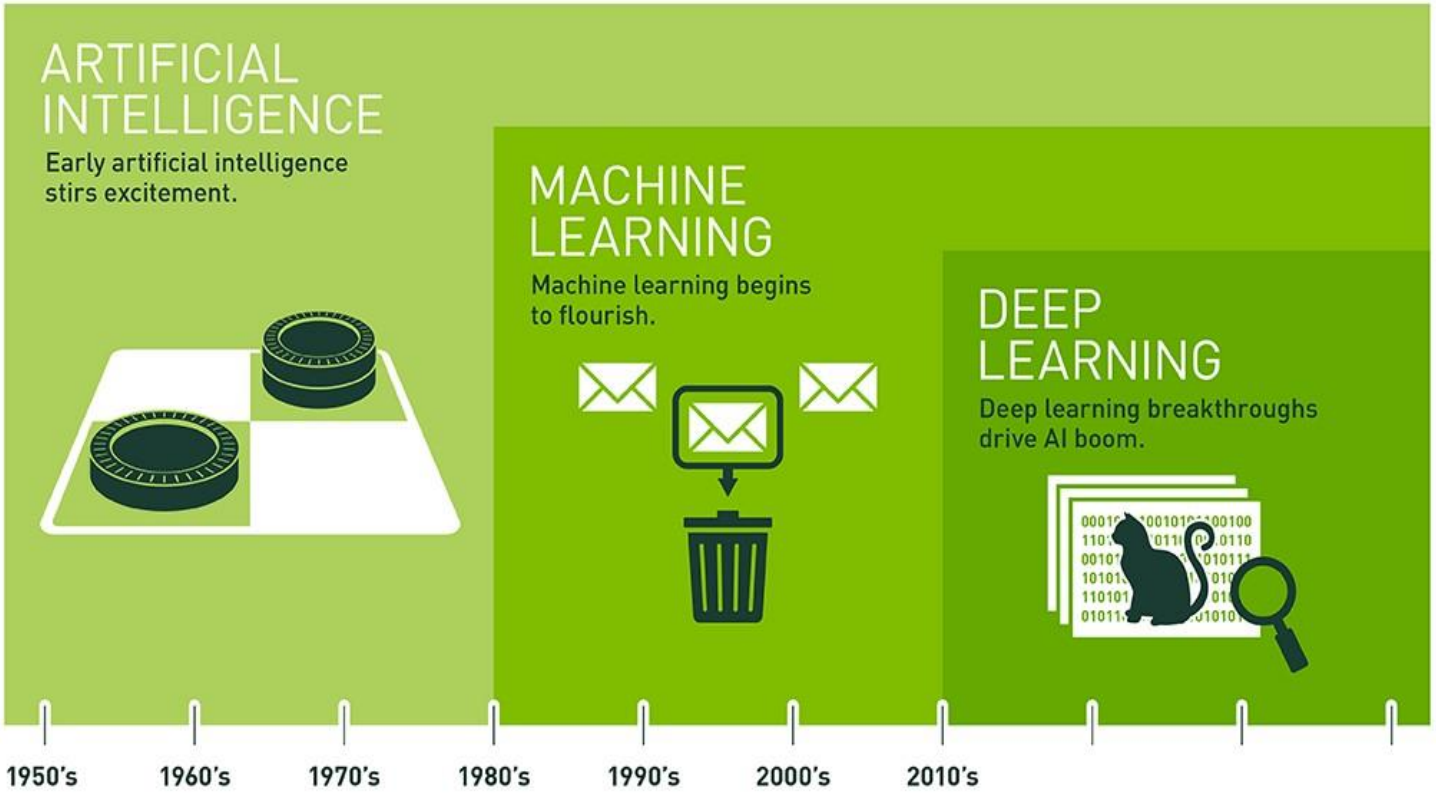
Materi

1. Pengantar Machine Learning
2. Supervised and Unsupervised Learning
3. Rumusan Masalah dan Hipotesis Penelitian
4. Kasus Penelitian Machine Learning
    - Prediksi
    - Clustering

Pengertian Machine Learning

Machine Learning merupakan sub-area dari Ilmu Komputer yang mampu memberikan komputer "**kemampuan untuk belajar tanpa diprogram eksplisit**"

Arthur Samuel
Peneliti IBM di area Computer Gaming dan Artificial Inteligence, 1959



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Artificial Intelegence

Setiap teknik yang membuat komputer dapat memiliki pengetahuan seperti manusia

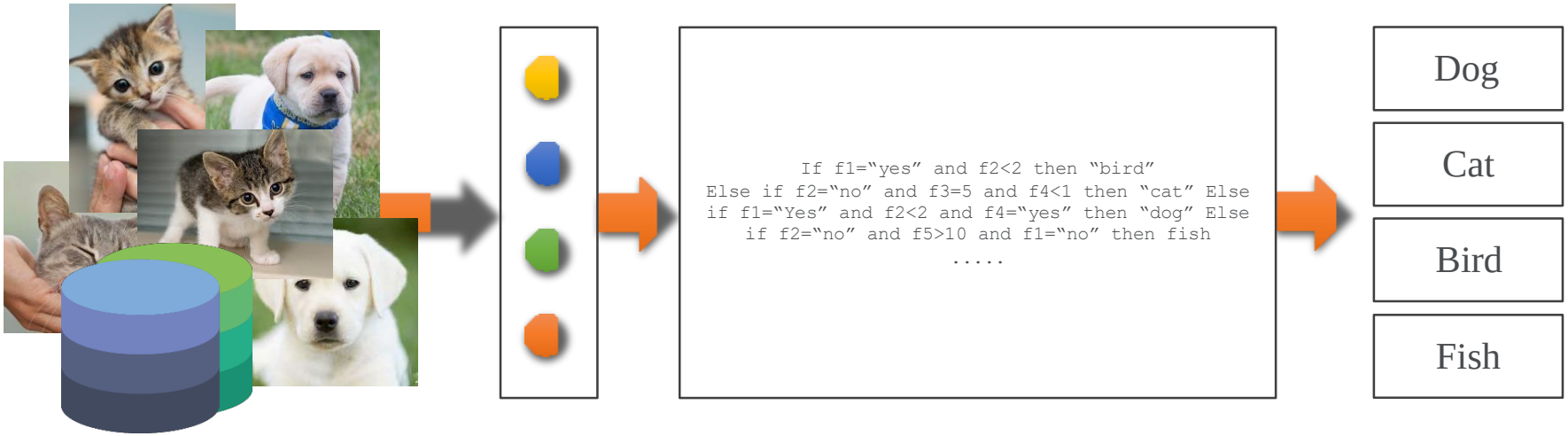
Machine Learning

Teknik untuk mengajari komputer tanpa secara langsung memprogram

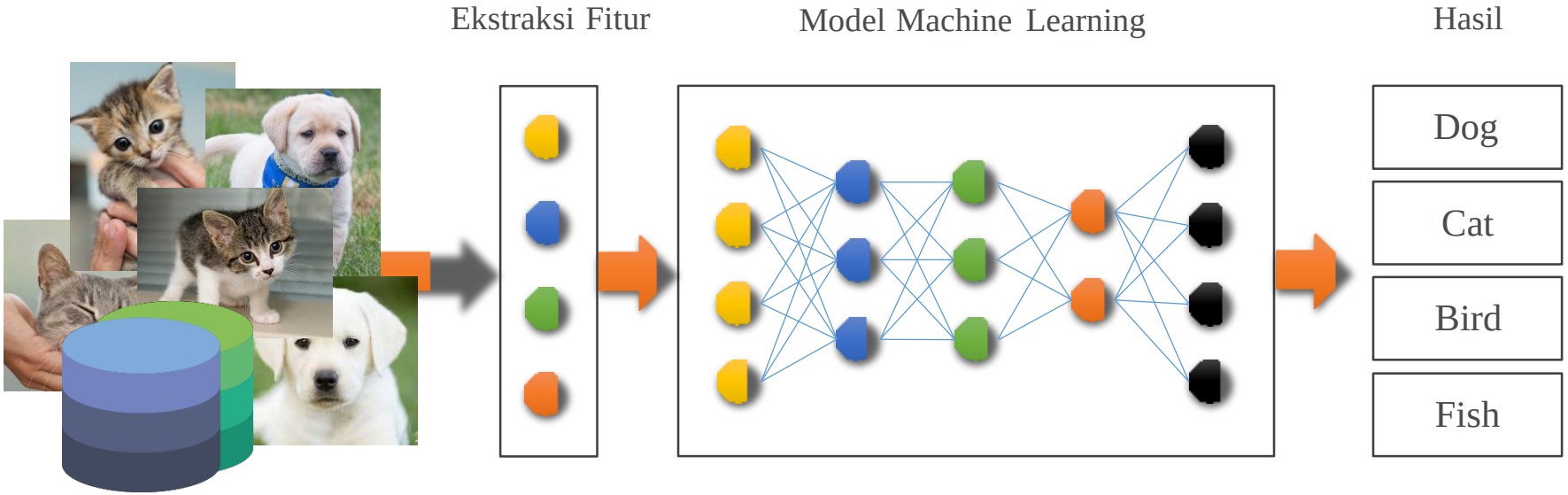
Deep Learning

Belajar untuk memahami fitur dari data dengan menggunakan neural network (Jaringan syaraf tiruan)

Pendekatan Tradisional



Pendekatan Machine Learning



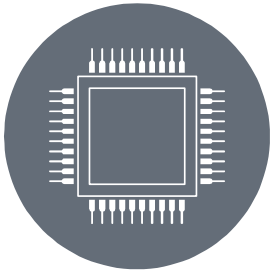
Manfaat Machine Learning

1. Menyederhanakan permasalahan
    - *Traditional approach*: menggunakan berbagai macam rule
    - *Machine learning approach*: menggunakan beberapa baris kode serta dapat diterapkan di problem lain
2. Machine learning dapat beradaptasi dengan data baru sedangkan traditional approach mengharuskan merubah banyak rule
3. Mendapatkan wawasan tentang masalah kompleks dan data dalam jumlah besar

Model Machine Learning

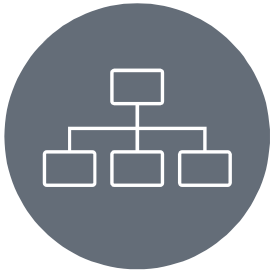
Supervised Learning

Manusia mengajarkan model dengan pengetahuan



Semi Supervised Learning

Peranan manusia dalam mengajarkan model dengan lebih sedikit



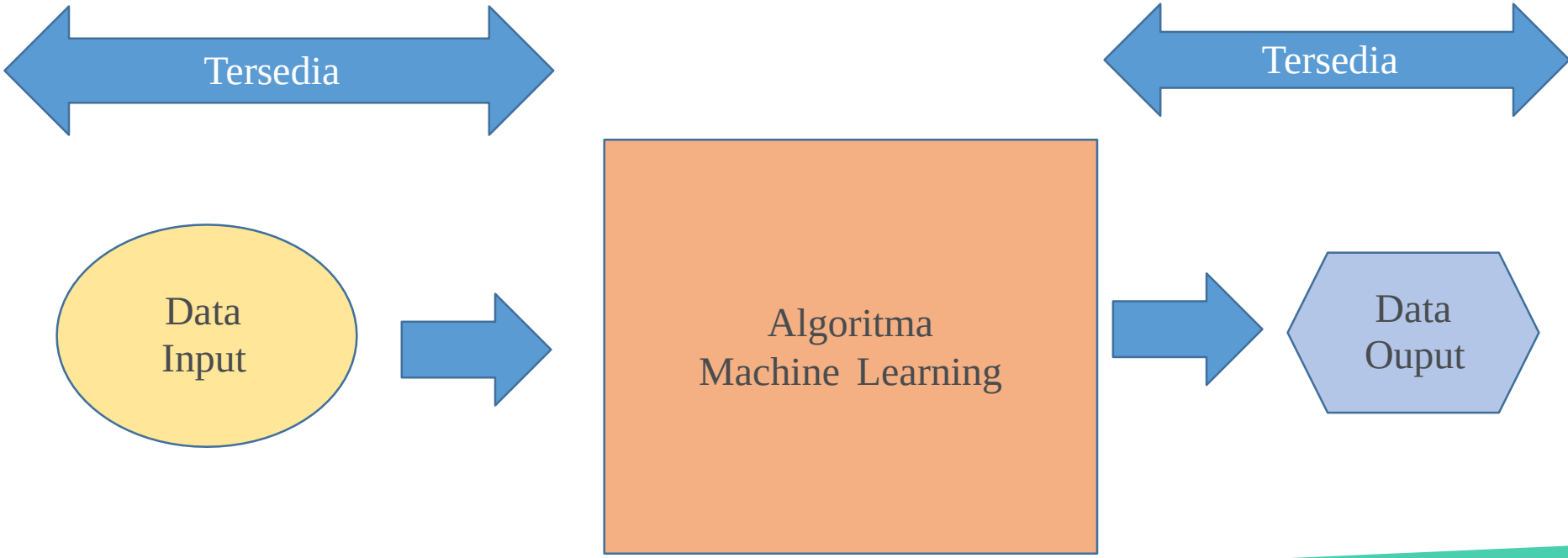
Unsupervised Learning

Pembelajaran tanpa pengawasan, membiarkan model bekerja sendiri untuk menemukan informasi yang mungkin tidak terlihat oleh mata manusia

Supervised Learning

- Mengajarkan model dan melatihnya dengan beberapa data dari dataset yang berlabel

Supervised Learning



Supervised Learning

- Mengajarkan model dan melatihnya dengan beberapa data dari dataset yang berlabel

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Supervised Learning

- Data historis pasien kanker yang terdiri dari komponen clump, unifSize, UnifShape, MargAdh, SingEplSize, bareNuc, BlandChrom, NormNucl, Mit dan Class

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Supervised Learning

- Class adalah data berlabel dari karakteristik data

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Supervised Learning

- Mengajarkan model dengan data berlabel

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Supervised Learning

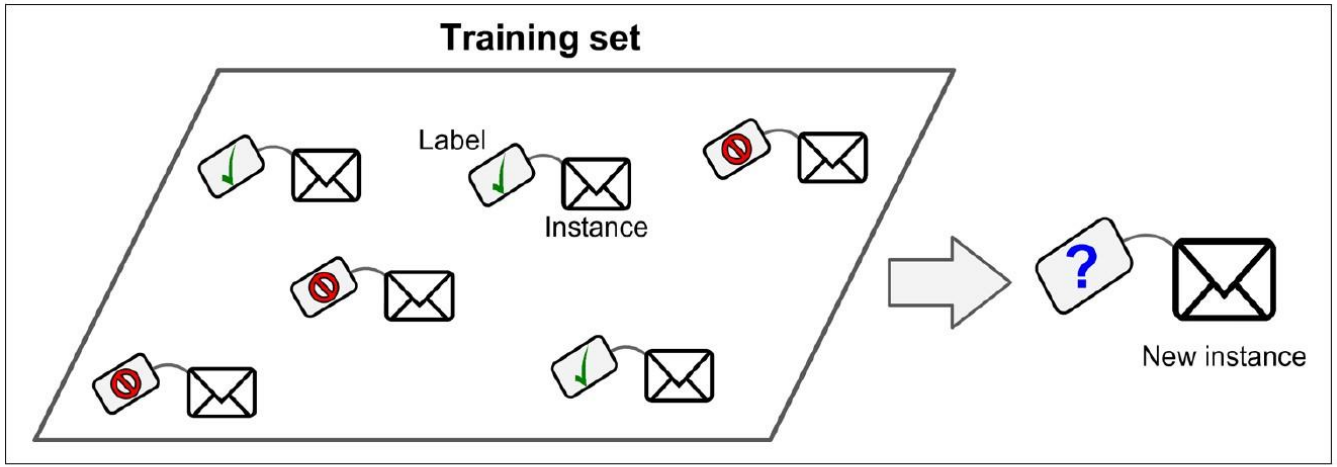
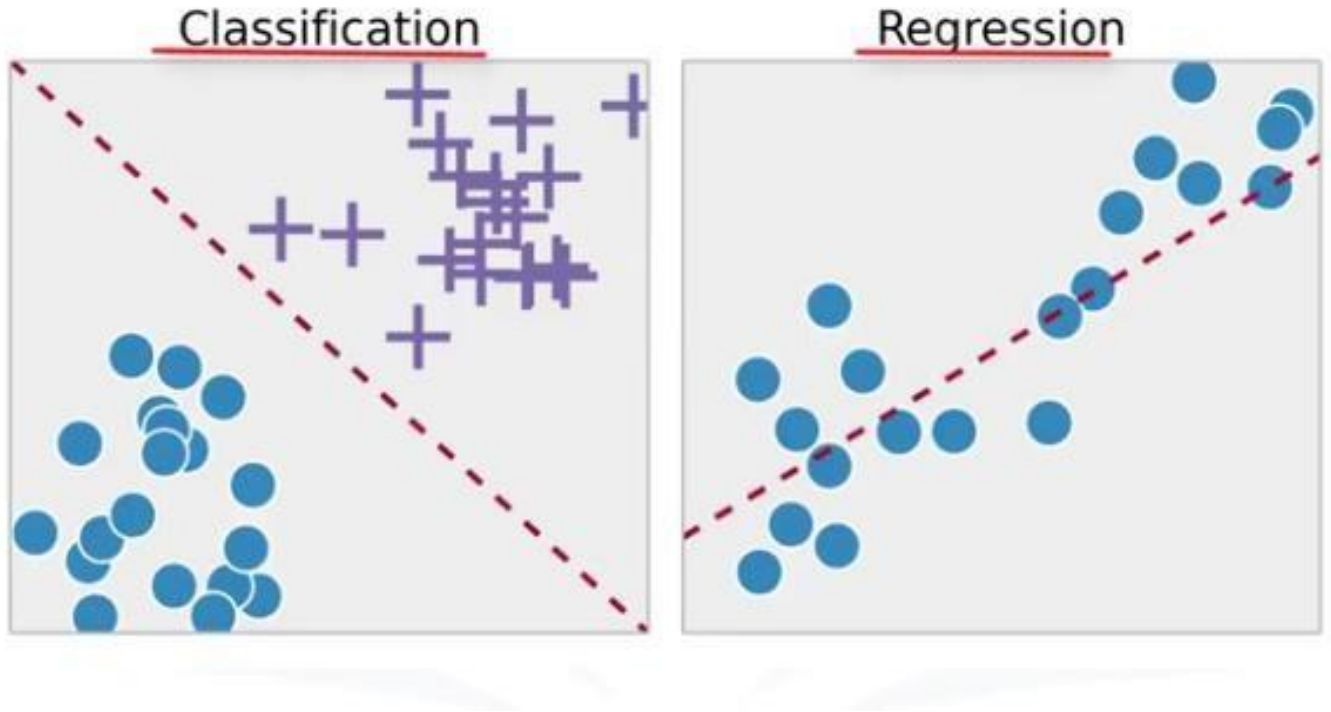


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

- Tugas dari Supervised Learning adalah klasifikasi dan prediksi (regresi)
- Misal memprediksi harga mobil berdasarkan fitur predictor (jarak tempuh, usia, merek, dll.)
- Untuk melatih sistem harus mempunyai banyak contoh mobil

Tipe Supervised Learning

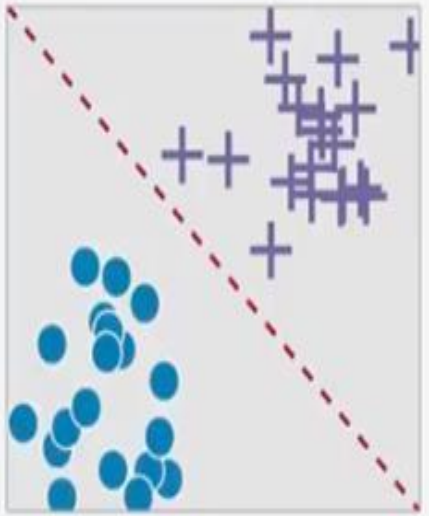


Supervised Learning

- Klasifikasi adalah proses memprediksi label atau kategori kelas diskrit

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Categorical Values

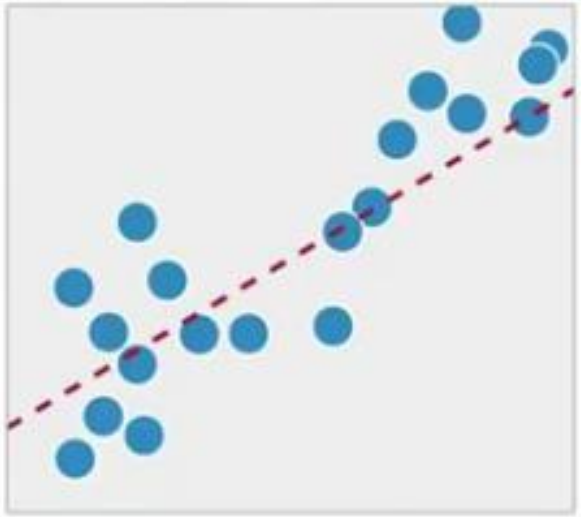


Supervised Learning

- Regresi adalah proses memprediksi nilai kontinu

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values



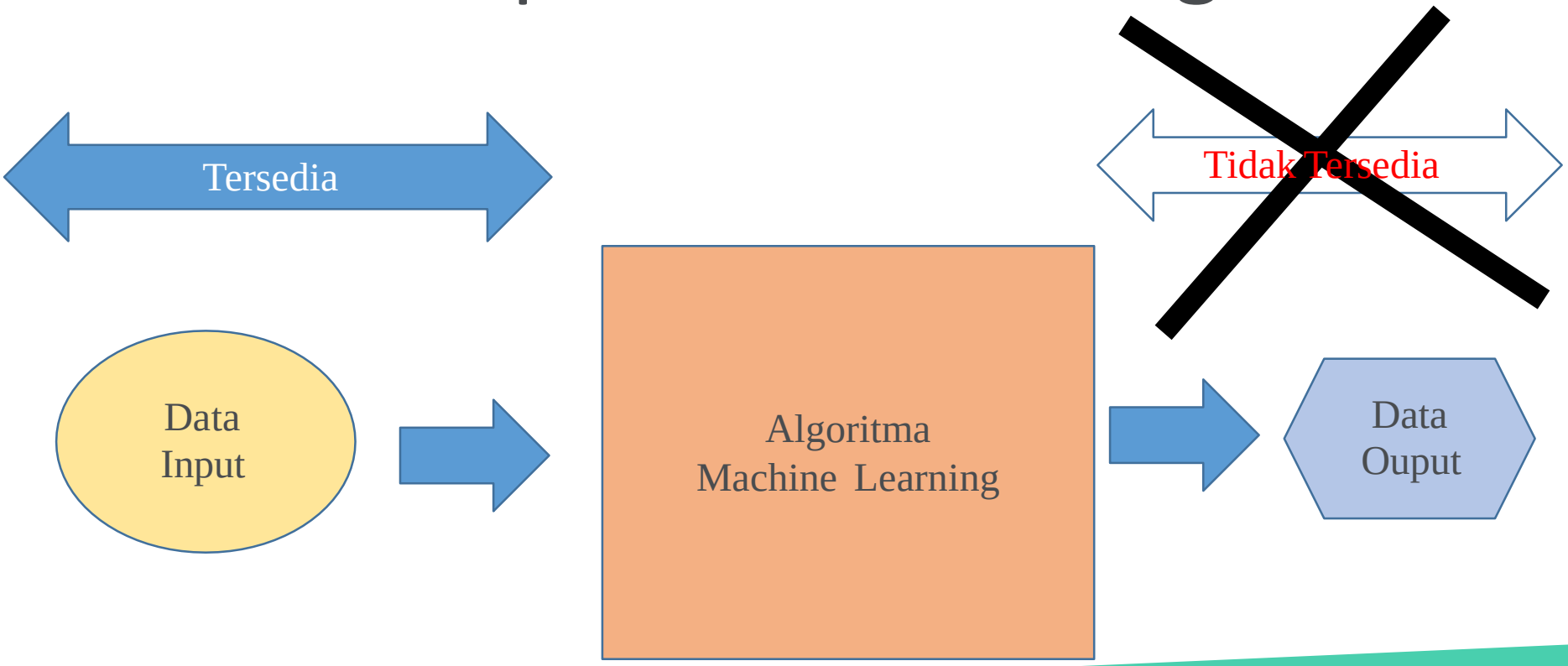
Algoritma Supervised Learning

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural networks

Unsupervised Learning

- Algoritma yang melatih dataset, dan menarik kesimpulan pada data tidak berlabel (*unlabeled*)

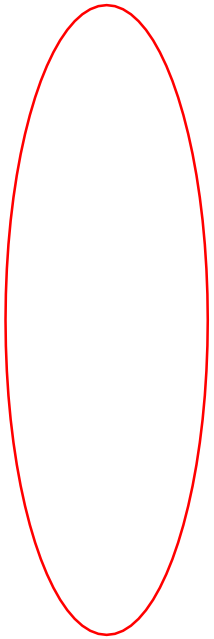
Unsupervised Learning



Unsupervised Learning

- Semua data tidak berlabel

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6



Unsupervised Learning

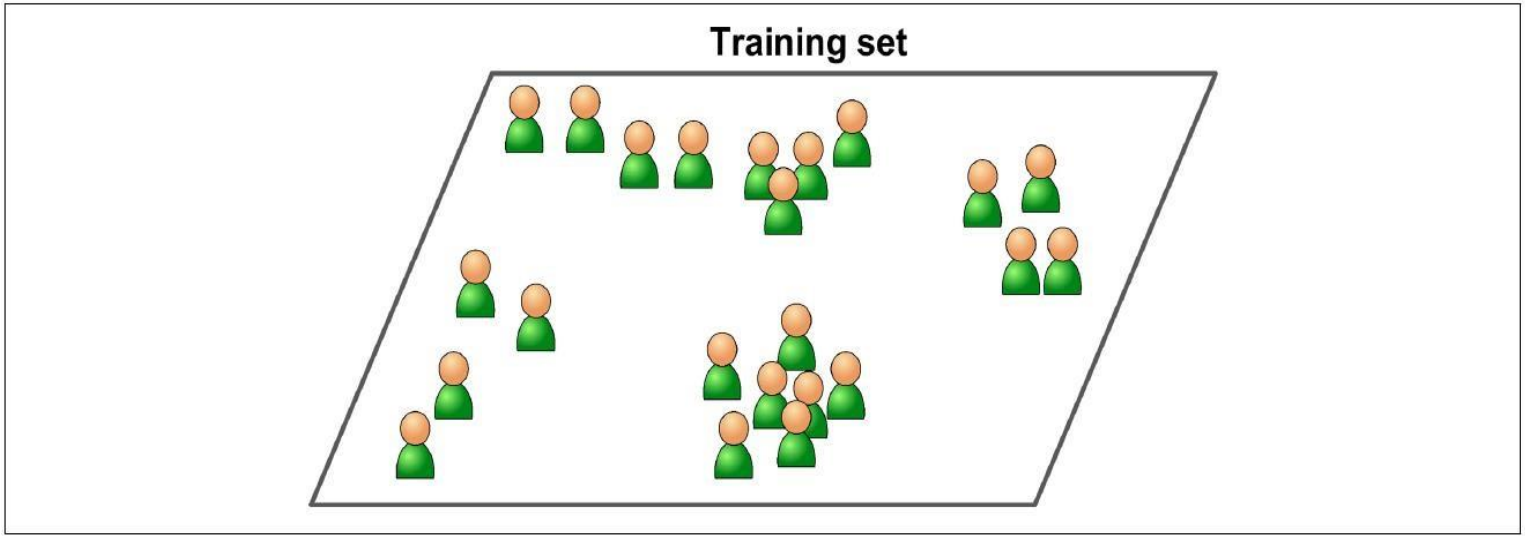
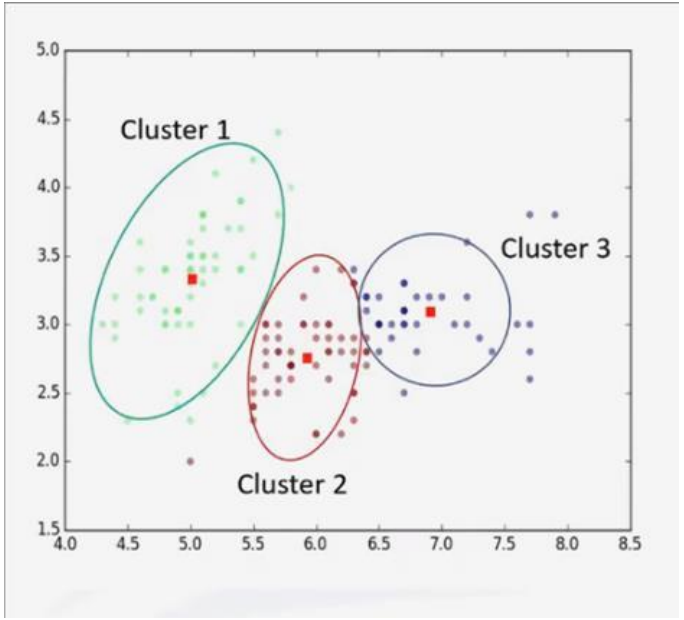


Figure 1-7. An unlabeled training set for unsupervised learning

- Tipe unsupervised learning adalah Cluster
- Contohnya implementasi fraud pada kartu kredit
- Dapat dikombinasikan dengan algortima supervised learning

Unsupervised Learning

- Clustering dianggap sebagai salah satu pembelajaran mesin tanpa pengawasan yang paling populer
- Teknik yang digunakan adalah untuk mengelompokkan titik data atau objek yang serupa
- Clustering digunakan untuk:
  - Menemukan Struktur
  - Peringkasan (summerization)
  - Deteksi anomali



Supervised vs Unsupervised

Supervised:

1. Klasifikasi
   - Klasifikasi data berlabel
2. Regresi
   - Prediksi tren dari data berlabel sebelumnya
3. Memiliki metode evaluasi lebih banyak
4. Lingkungan yang lebih terkontrol

Unsupervised:

1. Clustering
   - Menemukan pola dan mengelompokkan dari data yang tidak berlabel
2. Memiliki sedikit metode evaluasi
3. Lingkungan yang kurang terkontrol

Judul Penelitian: Prediksi emisi Co2 untuk uji kelulusan mobil dalam baku mutu emisi menggunakan Regresi Linier

Rumusan Masalah

Apakah nilai emisi dapat diukur menggunakan metode Regeresi Linier?

Tujuan Penelitian (Hipotesis Penelitian)

Mengukur nilai emisi menggunakan metode Regeresi Linier

Pengantar Regresi

- Penggunaan statistika dalam mengolah data penelitian berpengaruh terhadap tingkat analisis hasil penelitian
- Penelitian dalam bidang ilmu pengetahuan alam (science) yang menggunakan perhitungan statistika akan menghasilkan data yang mendekati benar, jika memperhatikan tata cara analisis data yang digunakan
- Dalam memprediksi dan mengukur nilai dari pengaruh satu variabel (bebas/independent/predictor) terhadap variabel lain (tak bebas/dependent/ response) dapat digunakan uji regresi

Regresi

- Uji regresi adalah kajian dari hubungan antara satu variabel bebas dengan satu atau lebih variabel tak bebas
- Jika variabel bebasnya satu, maka disebut dengan regresi linear sederhana
- Jika variabel bebasnya lebih dari satu, maka dinamakan regresi linear berganda

Regresi

- Analisis/Uji regresi banyak digunakan dalam perhitungan hasil akhir untuk penulisan karya ilmiah/penelitian.
- Hasil perhitungan uji regresi dimuat dalam kesimpulan penelitian dan akan menentukan apakah penelitian yang sedang dilakukan berhasil atau tidak.
- Analisis perhitungan pada uji regresi menyangkut beberapa perhitungan statistika seperti uji signifikansi (uji-t, uji-F), anova dan penentuan hipotesis
- Hasil dari uji regresi berupa suatu persamaan regresi
- Persamaan regresi ini merupakan suatu fungsi prediksi variabel yang mempengaruhi variabel lain

Regresi Linier Sederhana

- Metode statistik untuk menguji sejauh mana hubungan sebab akibat antara variabel Faktor Penyebab terhadap variabel Akibatnya
- Faktor Penyebab (X) disebut dengan Predictor sedangkan variabel Akibat (Y) disebut Response
- Metode ini juga digunakan untuk melakukan peramalan ataupun prediksi tentang karakteristik kualitas maupun kuantitas

Prediksi emisi Co2 untuk uji kelulusan mobil dalam baku mutu emisi menggunakan Regresi Linier

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Prediksi emisi Co2 untuk uji kelulusan mobil dalam baku mutu emisi menggunakan Regresi Linier

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

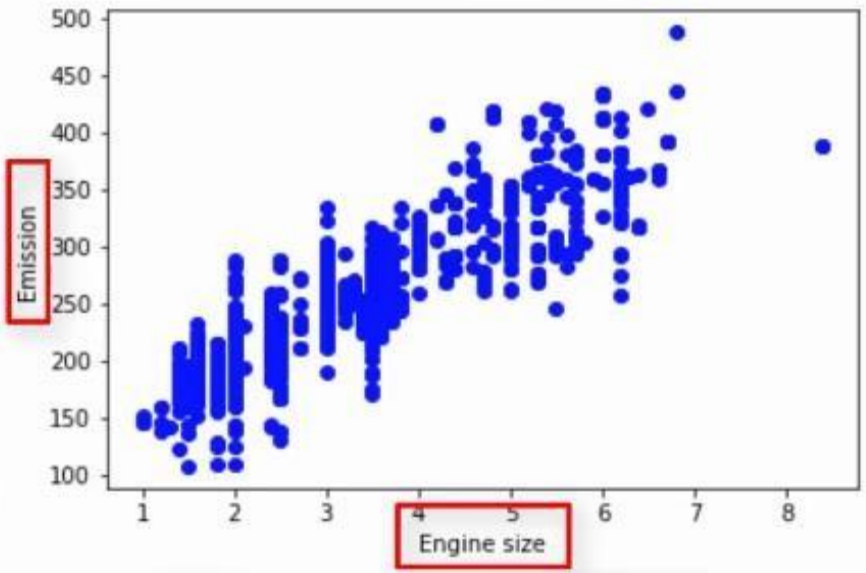
Continuous Values

Topologi Regresi Linear

1. Regresi Linear Sederhana :
- Prediksi emisi Co2 VS Ukuran mesin (engine size)
  - Variabel Independen (X): Ukuran mesin (engine size)
  - Variabel Dependen (Y): Emisi Co2
2. Regresi Linear Berganda:
- Prediksi emisi Co2 VS Ukuran mesin (engine size) dan Silinder
  - Variabel Independen (X): Ukuran mesin (engine size), Silinder
  - Variabel Dependen (Y): Emisi Co2

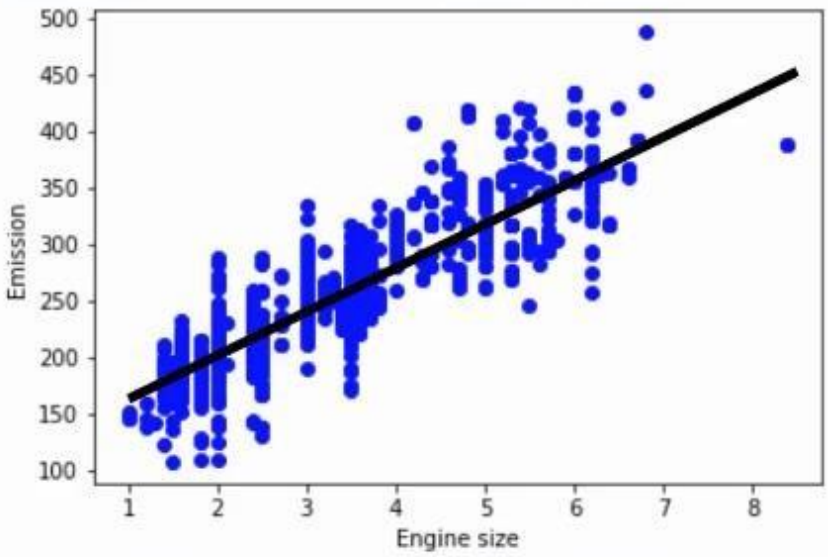
Bagaimana Regresi Linier Bekerja?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



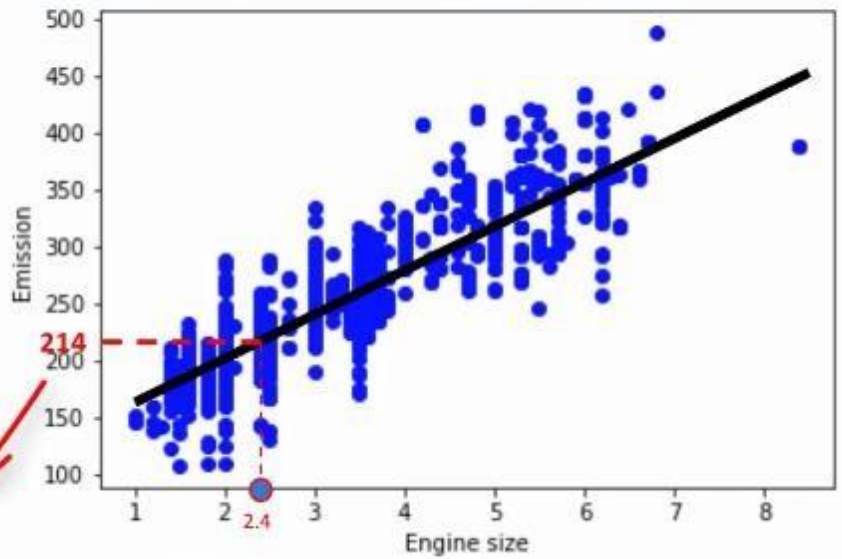
Bagaimana Regresi Linier Bekerja?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

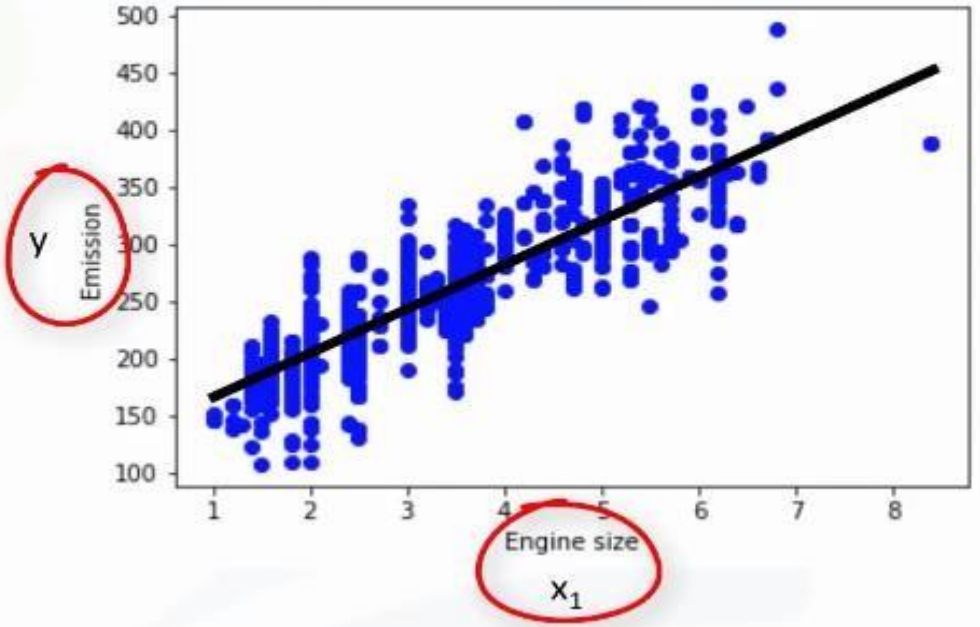


Bagaimana Regresi Linier Bekerja?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

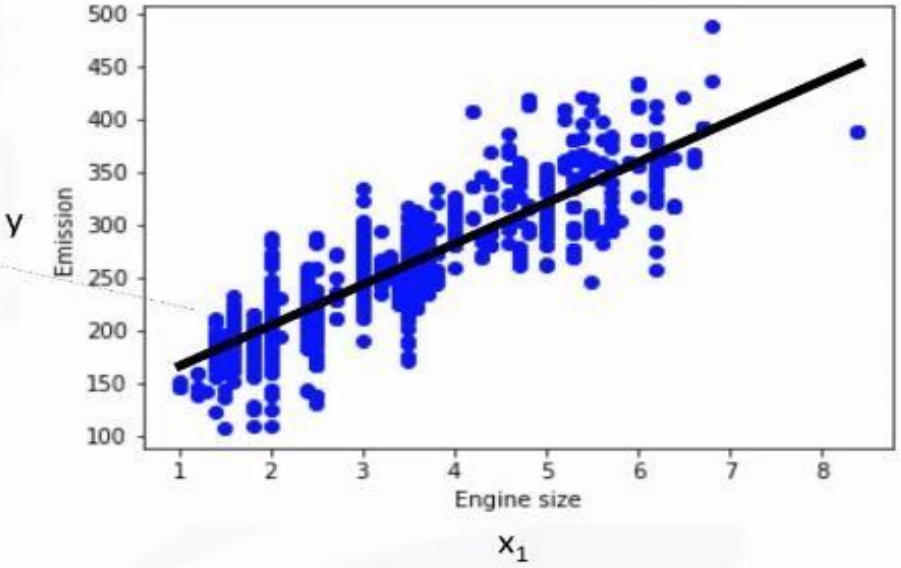


Representasi Model Regresi Linear



Representasi Model Regresi Linear

$$\hat{y} = \theta_0 + \theta_1 x_1$$



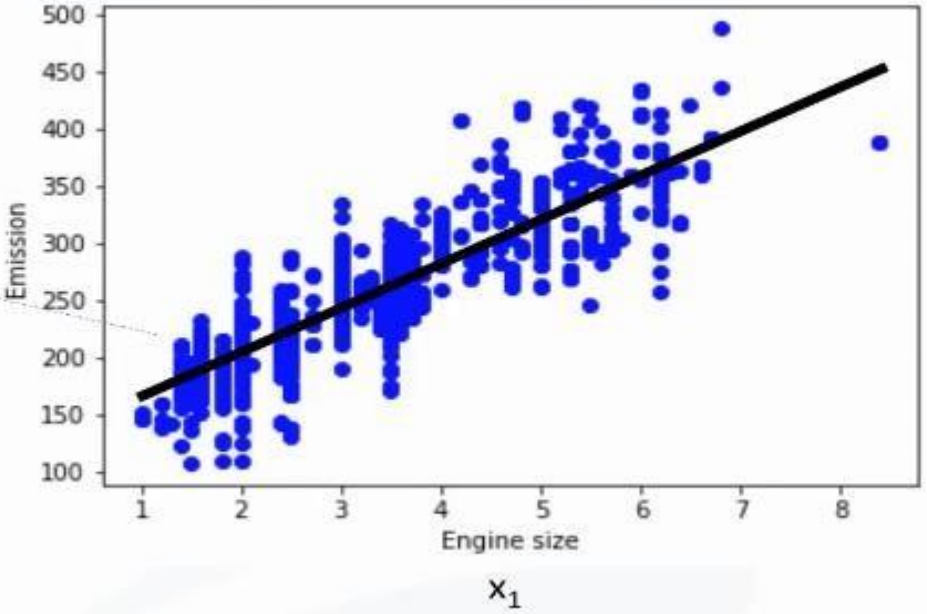
Representasi Model Regresi Linear

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable

a single predictor

y



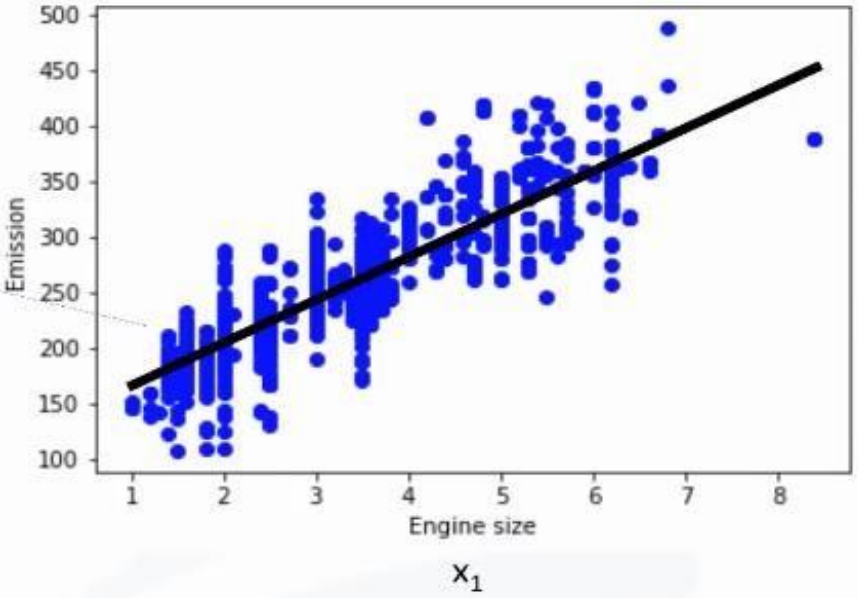
Representasi Model Regresi Linear

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable

a single predictor

y



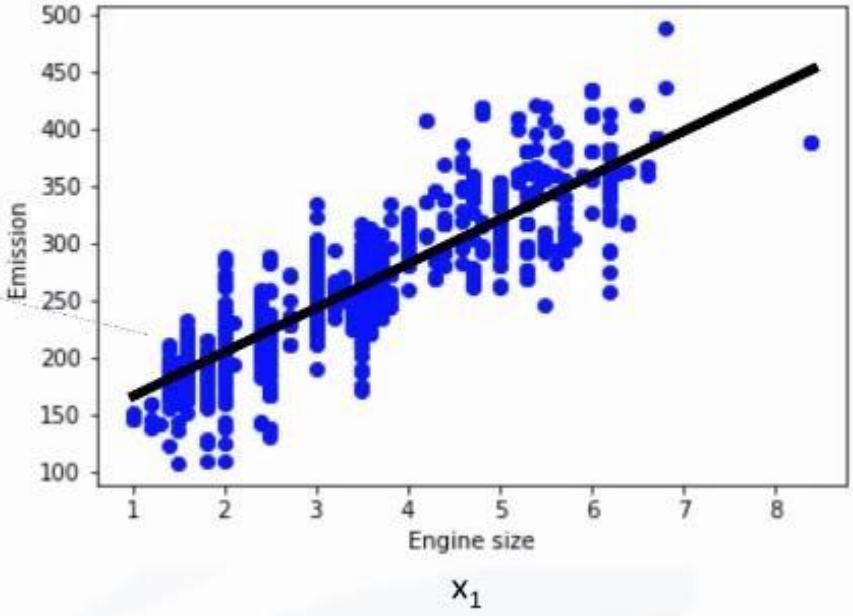
Representasi Model Regresi Linear

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable

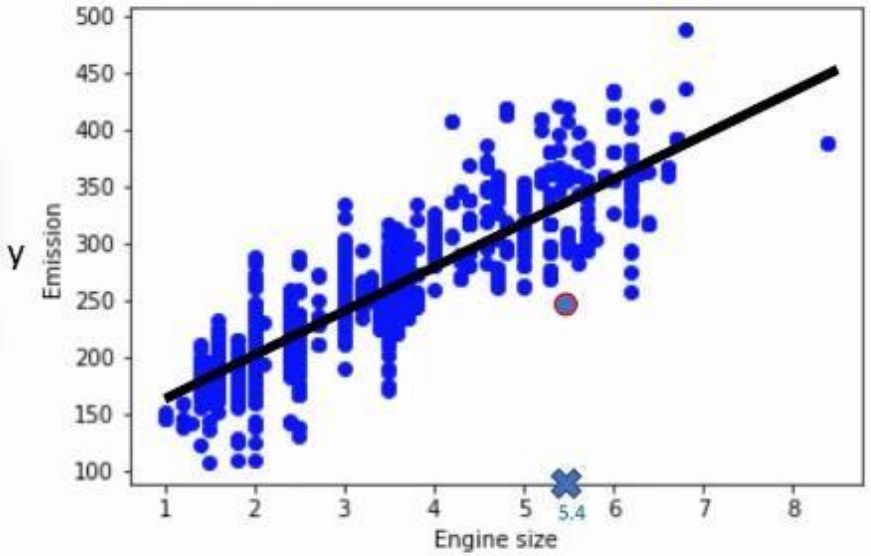
a single predictor

y



x_1

Bagaimana Mencari Model yang Fit?



Bagaimana Mencari Model yang Fit?

$x_1 = 2.4$ independent variable

y= 250 actual Co2 emission of x1

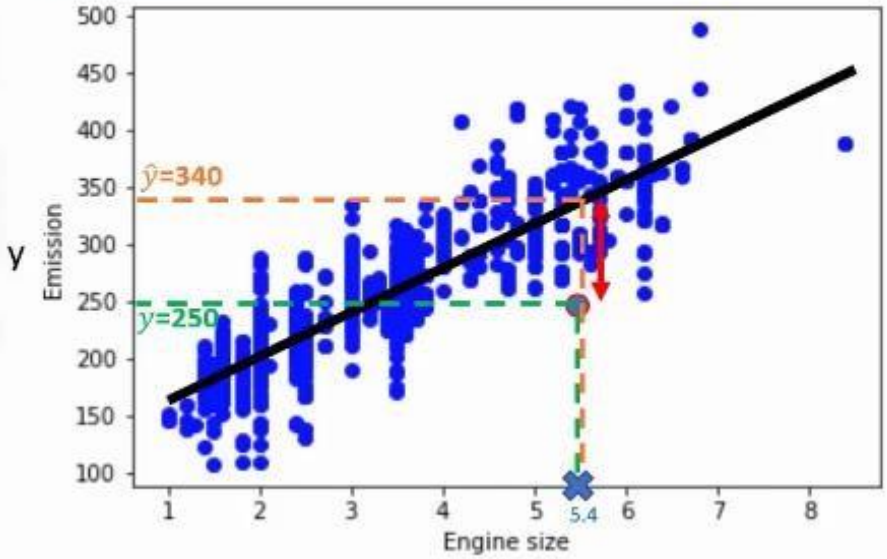
$\hat{y} = \theta_0 + \theta_1 x_1$

$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$

= 250 – 340

= -90



Bagaimana Mencari Model yang Fit?

$x_1 = 2.4$ independent variable

y= 250 actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$

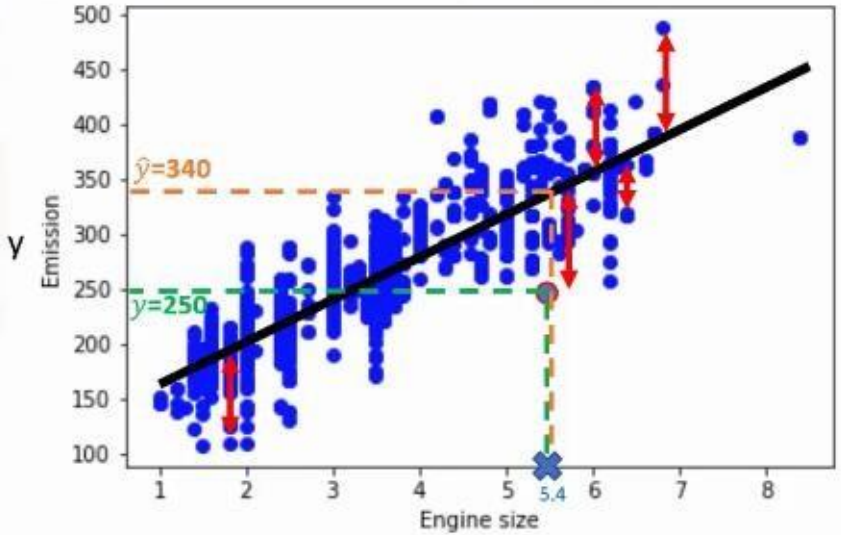
$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$

= 250 – 340

= -90

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



Bagaimana Mencari Model yang Fit?

$x_1 = 2.4$ independent variable

y= 250 actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$

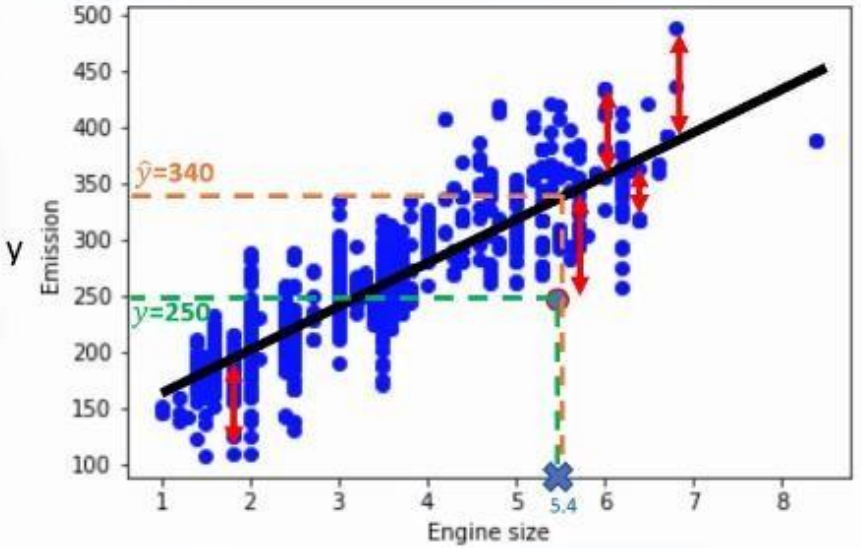
$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$

= 250 – 340

= -90

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



Estimasi Parameter-Parameter

$$\hat{y} = \theta_0 + \theta_1 x_1$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 mencakup kolom ENGINESIZE; y mencakup kolom CO2EMISSIONS.

Estimasi Parameter-Parameter

$$\hat{y} = \theta_0 + \theta_1 x_1$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

x_1 (ENGINESIZE) $\rightarrow$ y (CO2EMISSIONS)

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE

y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 (ENGINESIZE) y (CO2EMISSIONS)

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.34

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 256

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.34

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 256

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.34

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 256

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.34

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 256

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \ldots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \ldots}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ...)/9 = 3.34

$\bar{y}$ = (196 + 221 + 136 + ...)/9 = 256

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \ldots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \ldots}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.34$

$\bar{y} = (196 + 221 + 136 + \dots)/9 = 256$

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \dots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \dots}$$

$\theta_1 = 39$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 (ENGINESIZE) $\quad$ y (CO2EMISSIONS)

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.34$

$\bar{y} = (196 + 221 + 136 + \dots)/9 = 256$

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \dots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \dots}$$

$\theta_1 = 39$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$\theta_0 = 256 - 39 * 3.34$

$\theta_0 = 125.74$

Estimasi Parameter-Parameter

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.34$

$\bar{y} = (196 + 221 + 136 + \dots)/9 = 256$

$$\theta_1 = \frac{(2.0 - 3.34)(196 - 256) + (2.4 - 3.34)(221 - 256) + \dots}{(2.0 - 3.34)^2 + (2.4 - 3.34)^2 + \dots}$$

$\theta_1 = 39$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$\theta_0 = 256 - 39 * 3.34$

$\theta_0 = 125.74$

$$\hat{y} = 125.74 + 39 x_1$$

Prediksi dengan Garis Model

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Prediksi dengan Garis Model

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1 \, EngineSize$$

$$Co2Emission = 125 + 39 \, EngineSize$$

Prediksi dengan Garis Model

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1 \, EngineSize$$

$$Co2Emission = 125 + 39 \, EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

Judul Penelitian: Clustering Customer Potensial menggunakan k-Means Clustering

Rumusan Masalah

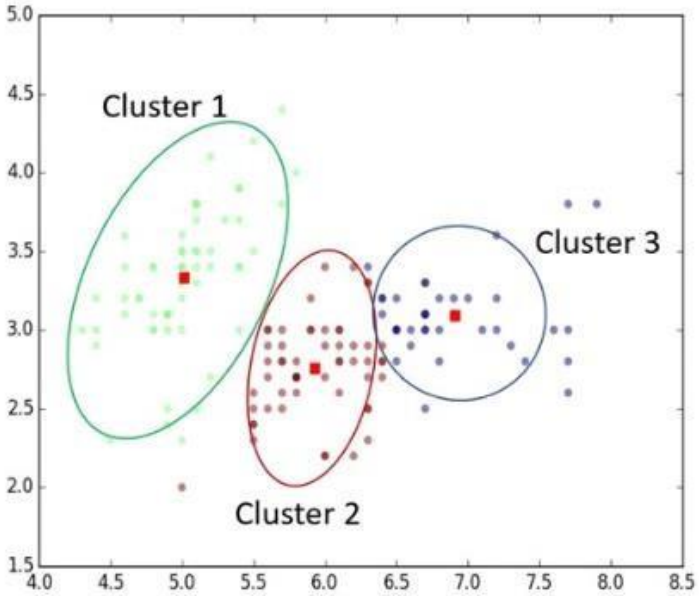
Bagaimana mencari customer potensial menggunakan metode k-Means Clustering?

Tujuan Penelitian (Hipotesis Penelitian)

Mencari customer potensial menggunakan metode k-Means Clustering

Clustering

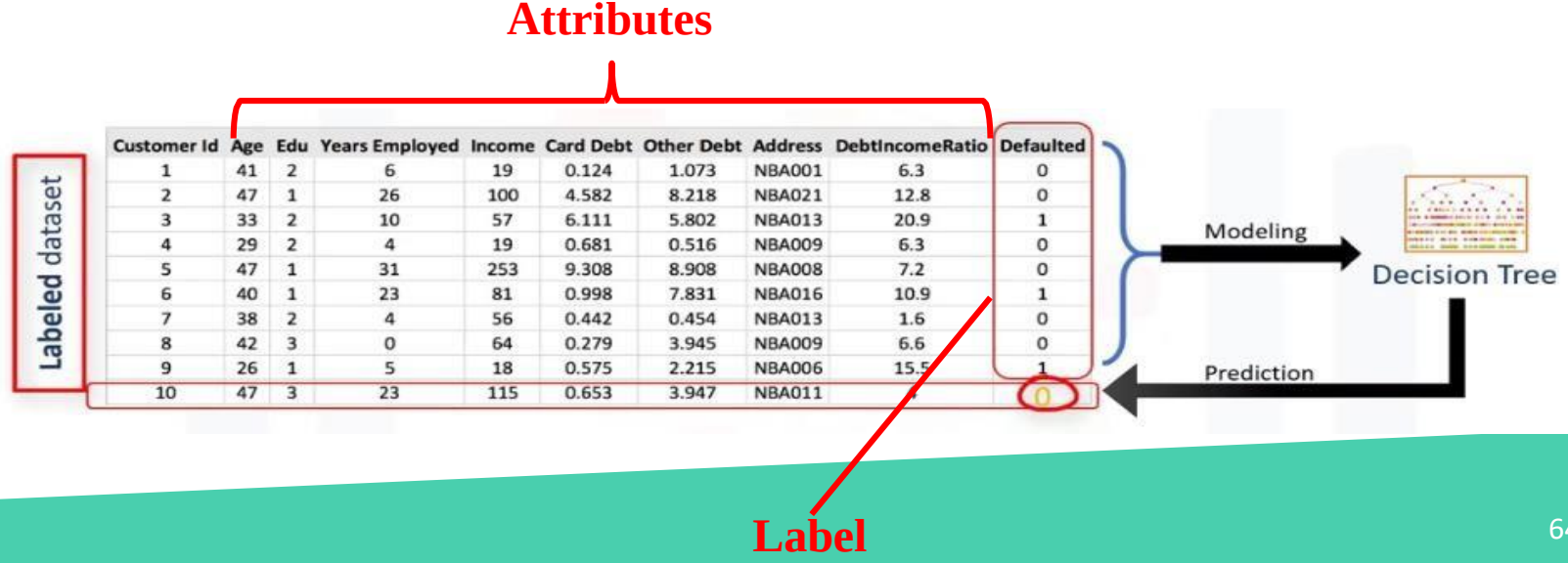
- Menemukan cluster pada dataset tanpa pengawasan (unsupervised)
- Sebuah grup objek yang yang memiliki kesamaan (*similar*) diantara objek didalam cluster objek, dan memiliki ketidaksamaan (*dissimilar*) dengan objek di cluster lainnya



Classification vs Clustering

Classification:

- Dibimbing/diawasi menggunakan set data berlabel saat dilakukan training (proses pembelajaran)
- Training (proses pembelajaran/pembentukan model) menggunakan atribut dan label



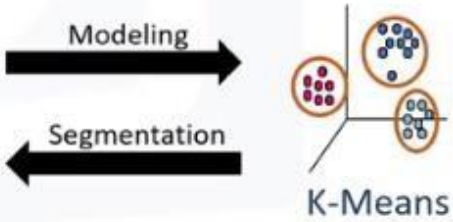
Classification vs Clustering

Clustering:

- Proses pemodelan tidak diawasi dengan menggunakan label dataset
- Label hanya digunakan untuk validasi model

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Classification vs Clustering

Labeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1
10	47	3	23	115	0.653	3.947	NBA011	4	0

Modeling

Decision Tree

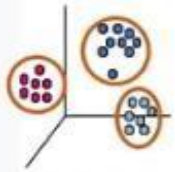
Prediction

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Modeling

Segmentation



K-Means

Penggunaan Clustering

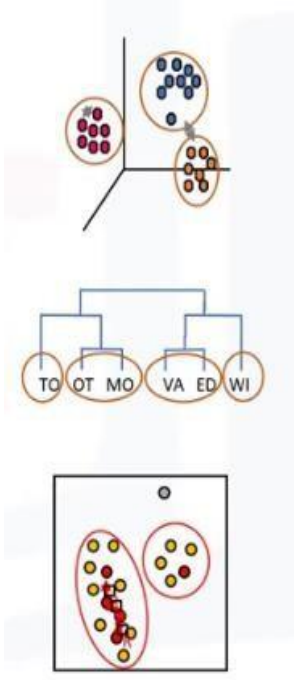
1. Retail Marketing
   - Mengidentifikasi pola pembelian pelanggan
   - Merekomendasikan buku atau film baru kepada pelanggan baru
2. Banking
   - Deteksi penipuan dalam penggunaan kartu kredit
   - Mengidentifikasi kelompok pelanggan (misalnya: Loyal/Tidak Loyal)
3. Insurance
   - *Fraud detection* (deteksi penipuan) dalam analisis klaim ansuransi
   - Risiko asuransi pelanggan

Penggunaan Clustering

4. Publication
   - Mengelompokkan berita secara otomatis berdasarkan kontennya
   - Merekomendasikan artikel berita serupa
5. Medicine
   - Mengkarakterisasi perilaku pasien
6. Biology
   - Mengelompokkan penanda genetik (*genetic markers*) untuk mengidentifikasi ikatan keluarga

Algoritma Clustering

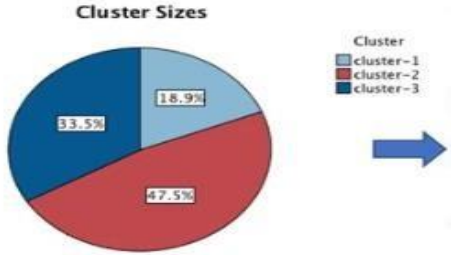
1. Partitioned-based Clustering
   - Relatively efficient
   - E.g., k-Means, k-Median, Fuzzy c-Means
2. Hierarchical Clustering
   - Produces trees of clusters
   - E.g. Agglomerative, Divisive
3. Density-based Clustering
   - Produces arbitrary shaped clusters
   - E.g. DBSCAN



k-Means Clustering

- Clustering bekerja pada data yang tidak diawasi berdasarkan kesamaan setiap dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

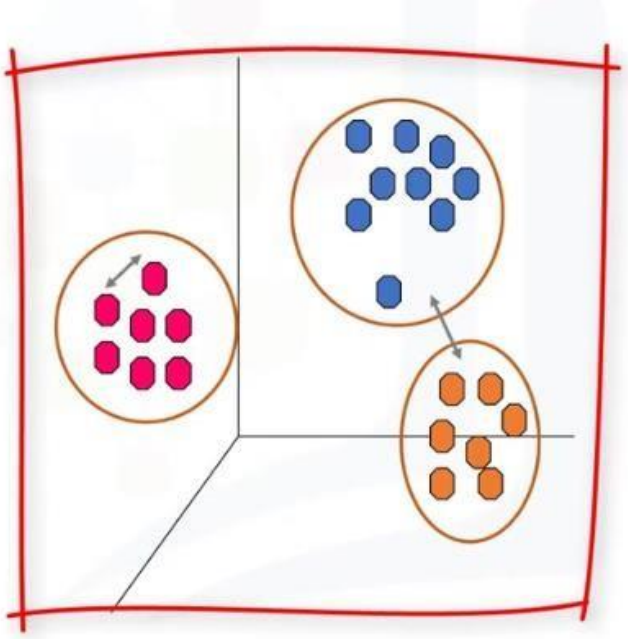


Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFULUENT AND MIDDLE AGED
3	AFFULUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFULUENT AND MIDDLE AGED
6	AFFULUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFULUENT AND MIDDLE AGED

Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

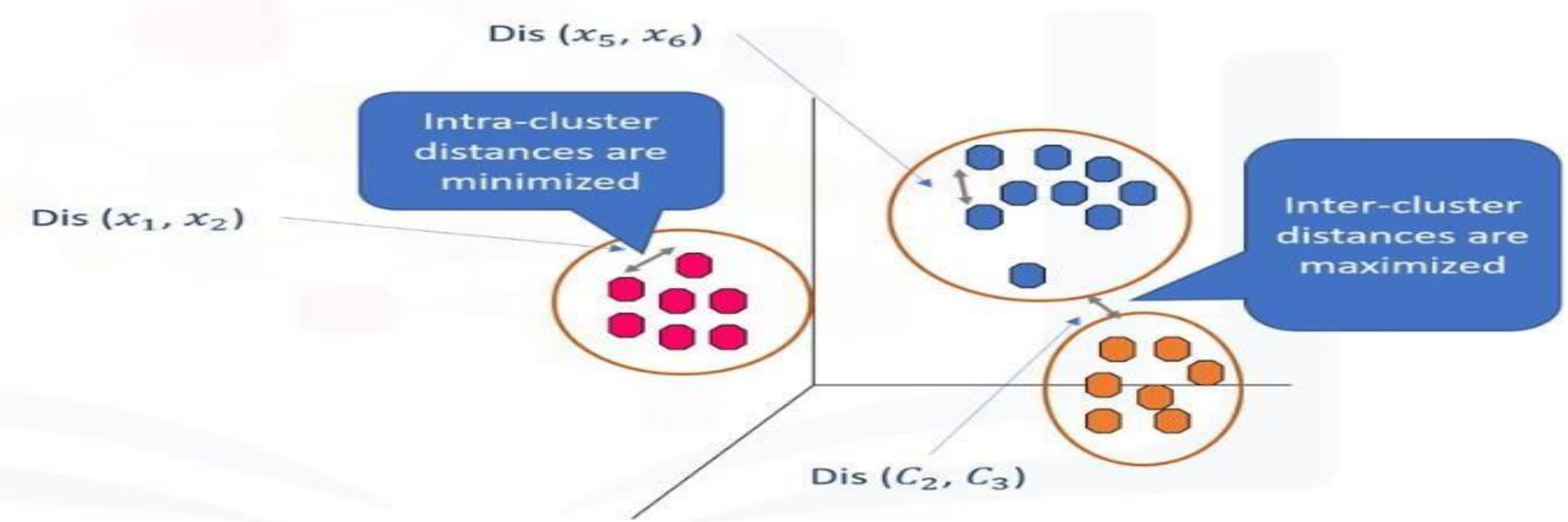
Algoritma K-Means

- K-Means termasuk dalam Partitioning Clustering
- K-Means membagi data menjadi subset (cluster) yang tidak tumpang tindih
- Data dalam sebuah cluster sangat mirip
- Data antar kelompok sangat berbeda



Menentukan similarity atau dissimilarity

- Similarity digunakan untuk dataset dalam internal satu cluster
- Disimilarity digunakan untuk dataset antar cluster



1-dimentional similarity/distance

- 1-dimentional similarity dapat digunakan untuk mengukur jarak dua titik menggunakan 1 nilai
- Rumus Euclidean Distance dapat digunakan untuk mengukur similarity

Customer 1
Age
54

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

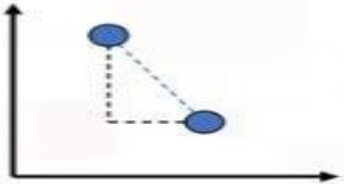
$$\text{Dis}(x_1, x_2) = \sqrt{(54 - 50)^2} = 4$$

Customer 2
Age
50

2-dimentional similarity/distance

- 2-dimentional similarity dapat digunakan untuk mengukur jarak dua titik menggunakan 2 nilai atau 2D *matrix space*

Customer 1	
Age	Income
54	190



Customer 2	
Age	Income
50	200

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54 - 50)^2 + (190 - 200)^2} = 10.77$$

Multi-dimentional similarity/distance

- Dengan demikian metode Euclidean Distance dapat digunakan untuk multidimensi dengan menambahkan titik pada rumus

Customer 1		
Age	Income	education
54	190	3

Customer 2		
Age	Income	education
50	200	8

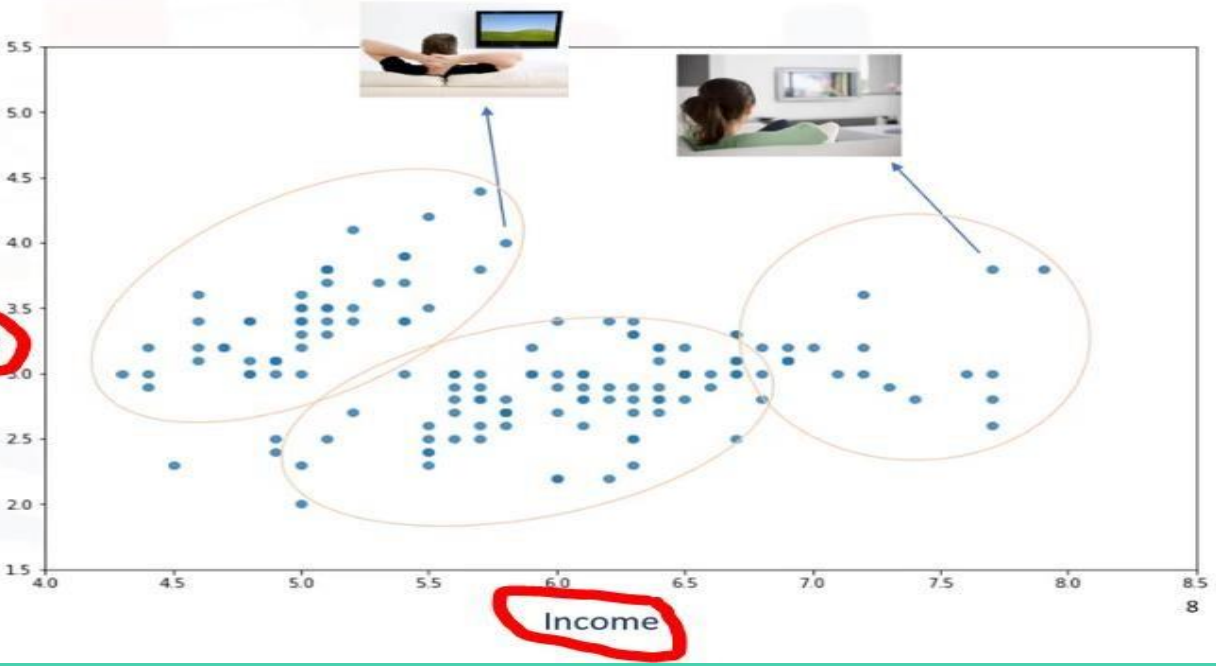
$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

Bagaimana k-Means Clustering bekerja?

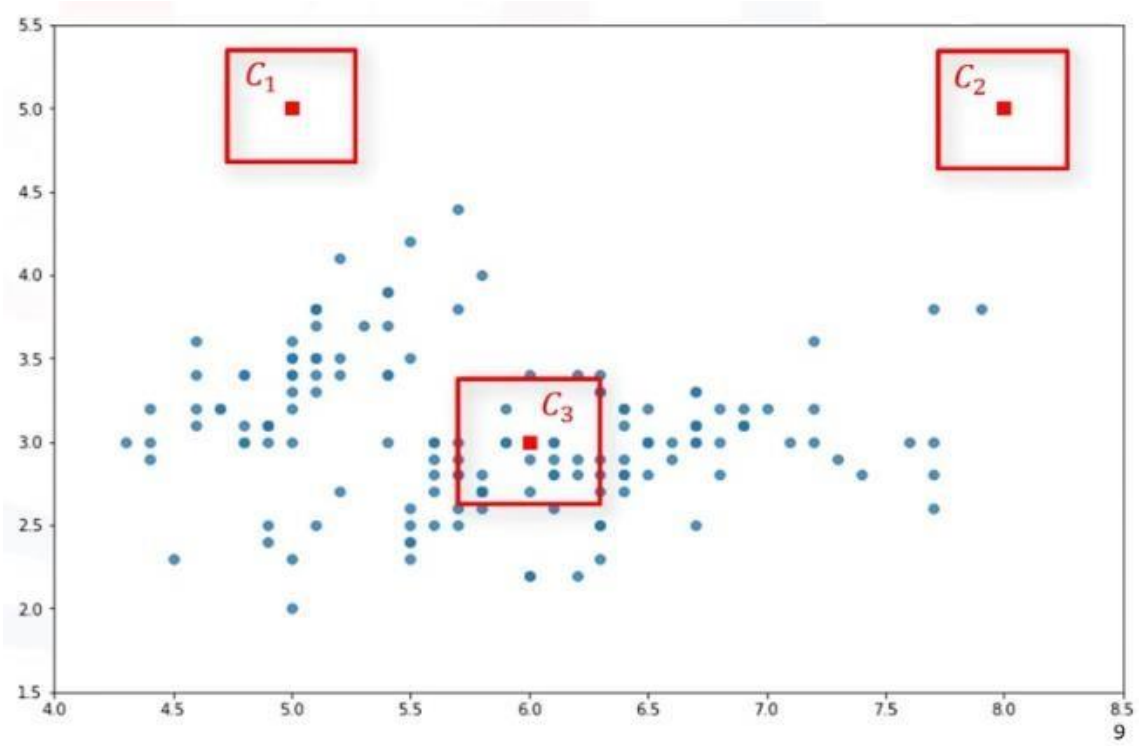
- Misalkan terdapat data yang memiliki attributes age dan income, dan tersebar dalam matrix 2D space (dapat digambarkan dalam diagram Cartesian)

Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...	...	..



k-Means Clustering - inisialisasi k

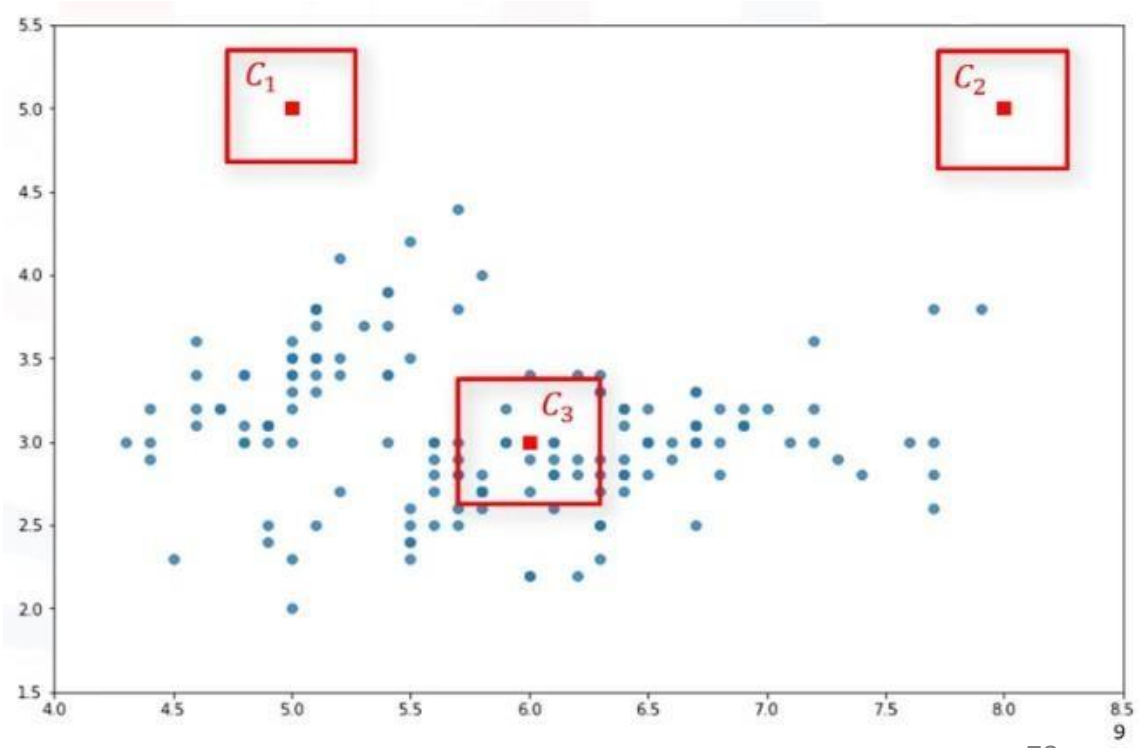
1. Misal diinisialisasi k=3, k sebagai centroids yang dipilih secara random



k-Means Clustering - inisialisasi k

1. Misal diinisialisasi k=3, k sebagai centroids yang dipilih secara random

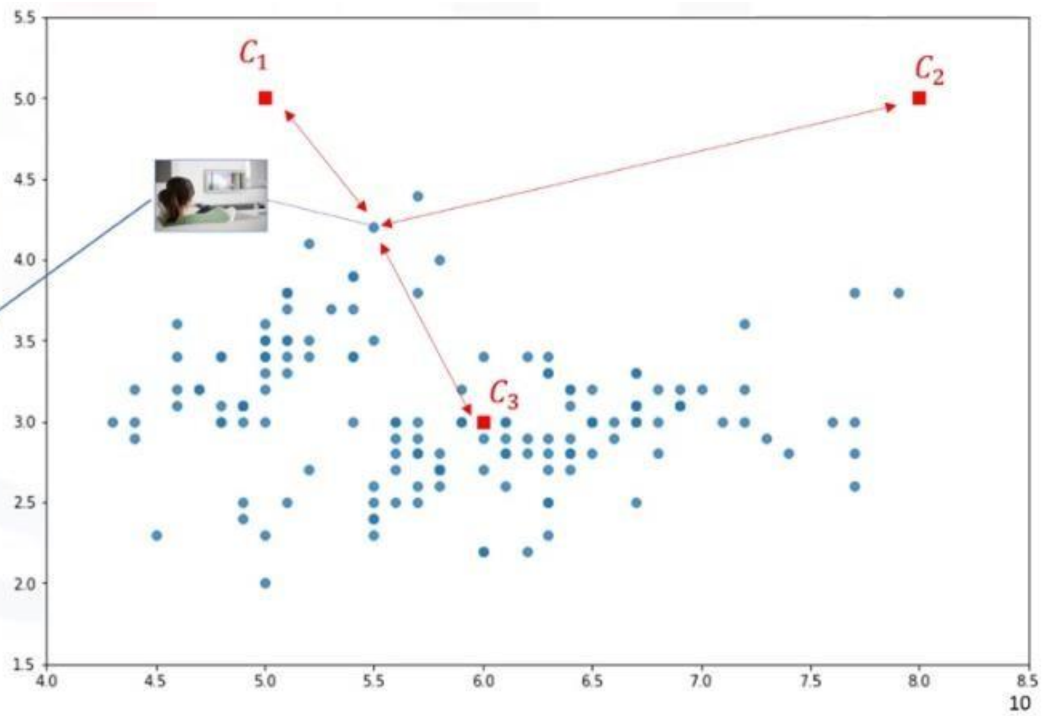
C1 = [5., 5.]
C2 = [8., 5.]
C3 = [6., 3.]



k-Means Clustering - hitung jarak (distance)

2. Hitung jarak setiap titik dataset dengan 3 centroid yang telah ditentukan secara random

C_1	C_2	C_3
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p.., c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p.., c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p.., c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$

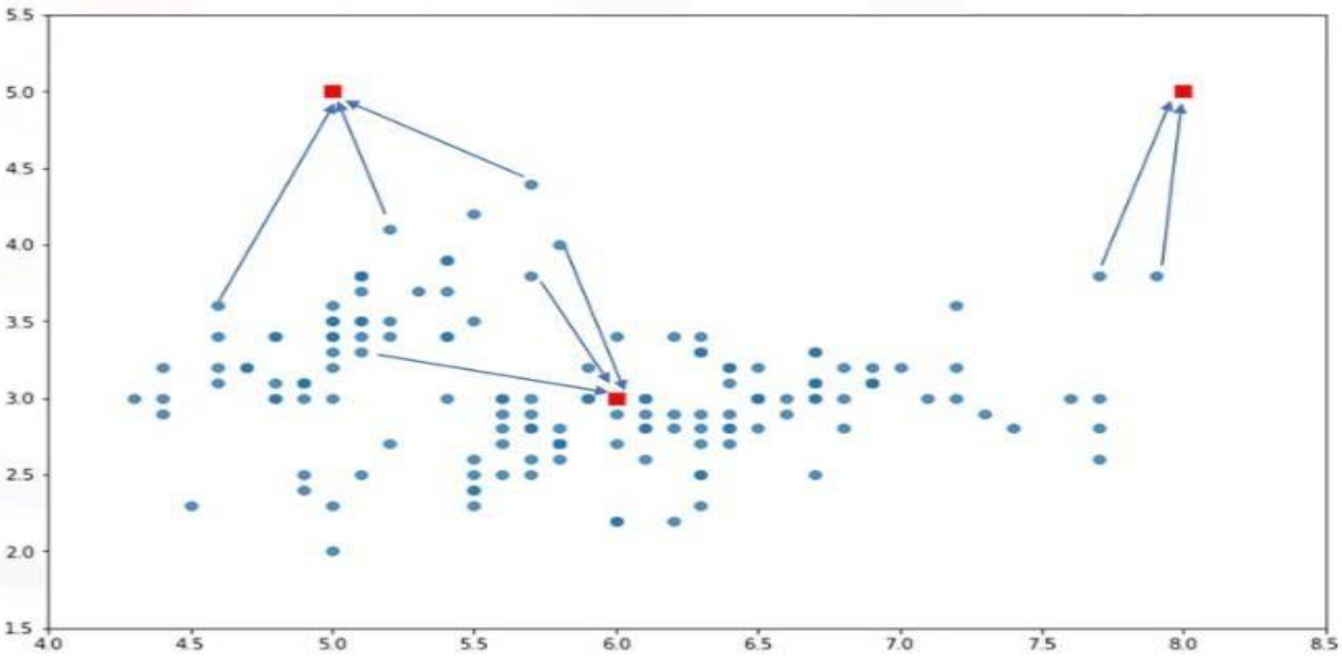


k-Means Clustering – tetapkan ke centroid

3. Tetapkan setiap titik dataset ke centroid terdekat

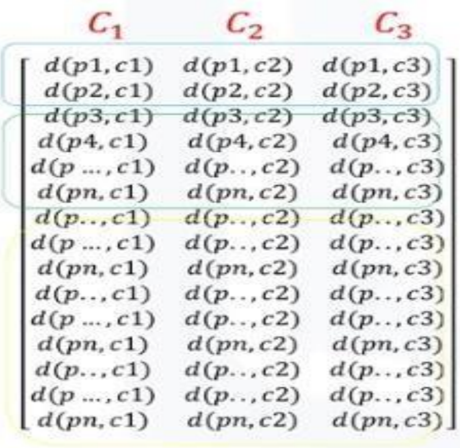
$$
\begin{array}{ccc}
C_1 & C_2 & C_3
\end{array}
$$

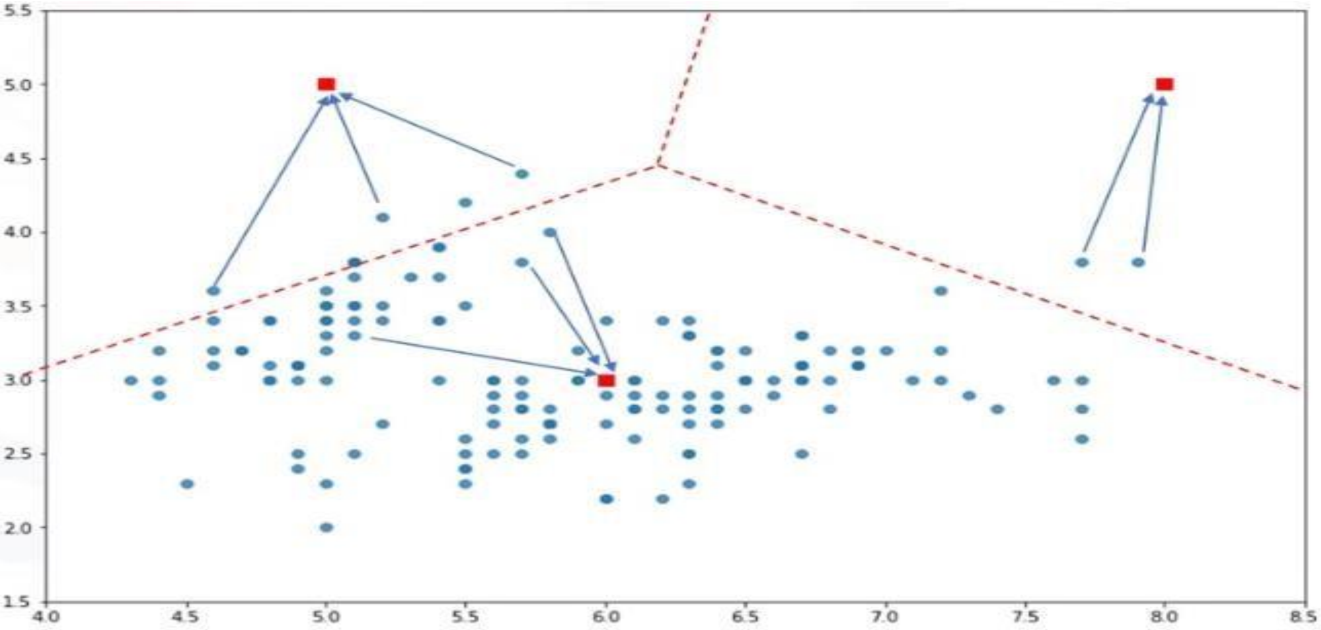
$$
\begin{bmatrix}
d(p1,c1) & d(p1,c2) & d(p1,c3) \\
d(p2,c1) & d(p2,c2) & d(p2,c3) \\
d(p3,c1) & d(p3,c2) & d(p3,c3) \\
d(p4,c1) & d(p4,c2) & d(p4,c3) \\
d(p\ ...,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ ...,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ ...,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ ...,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3)
\end{bmatrix}
$$



k-Means Clustering – tetapkan titik ke satu centroid

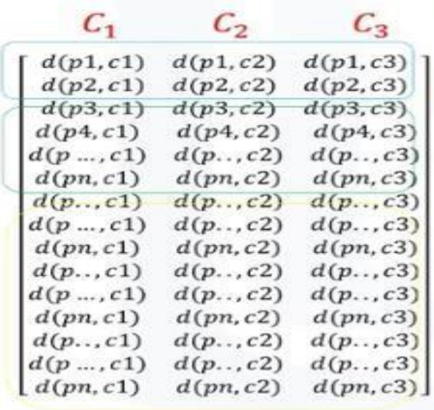
3. Tetapkan setiap titik dataset ke centroid terdekat, sehingga terbentuk voronoi diagram yang menunjukkan pembatas antar cluster

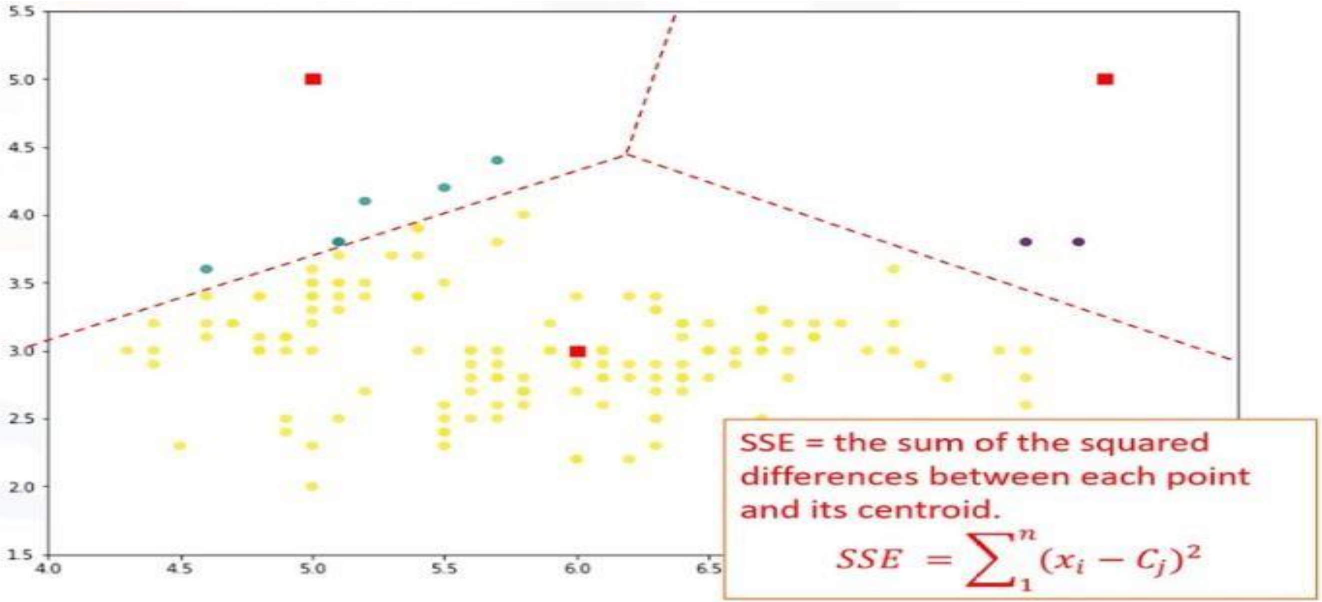




k-Means Clustering – tetapkan titik ke satu centroid

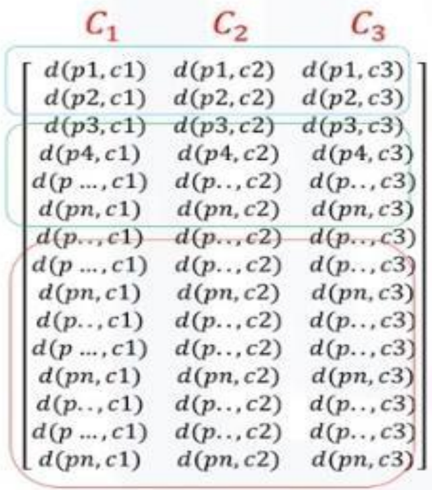
3. Tetapkan setiap titik dataset ke centroid terdekat dan hitung SSE
   SSE merupakan jumlah error yang terjadi antara titik dataset dengan centroid
   Tugas k-Means adalah mengoptimalkan (memperkecil) nilai SSE disetiap iterasi

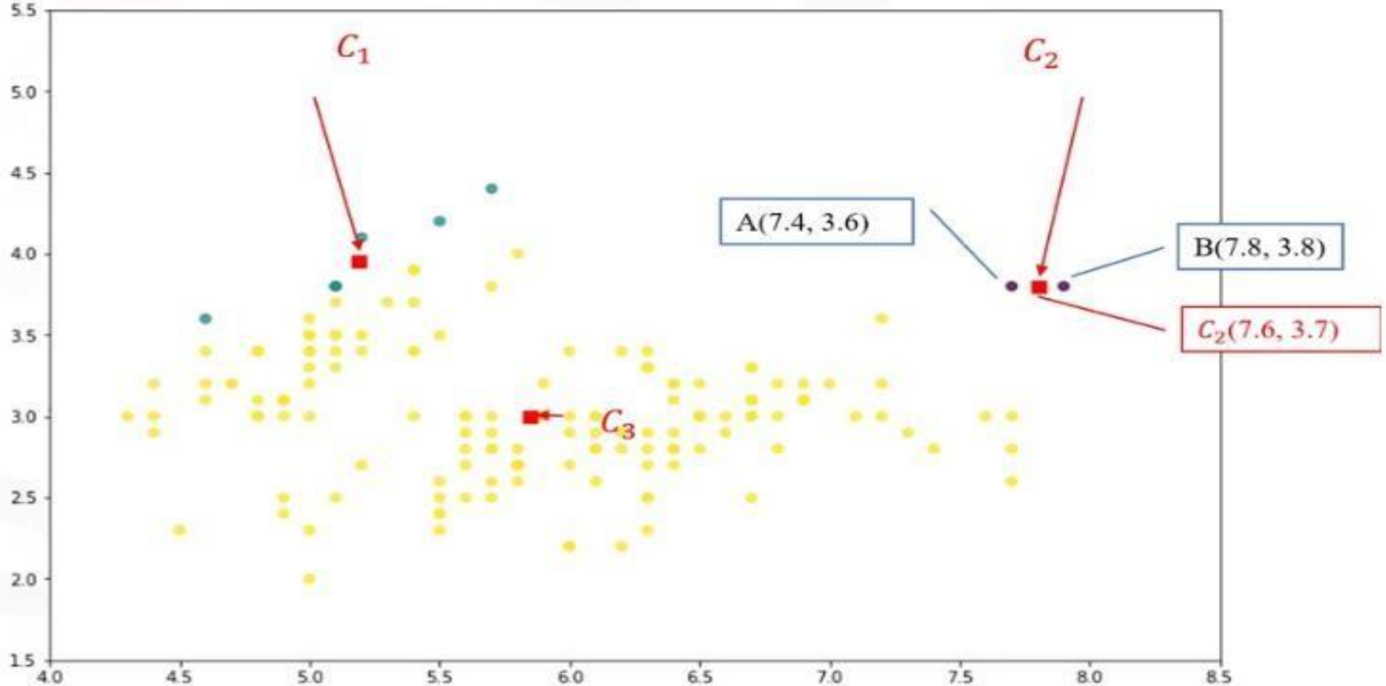




k-Means Clustering – computer new centroids

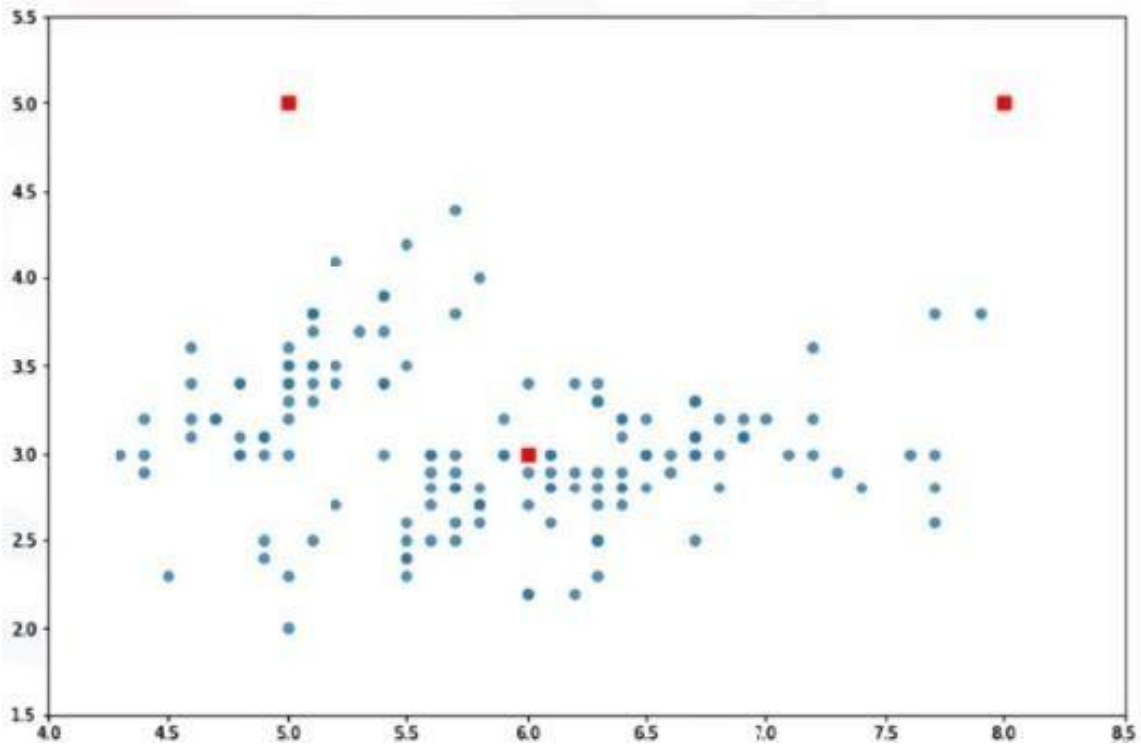
4. Hitung centroid baru dari setiap cluster (C1, C2, dan C3)





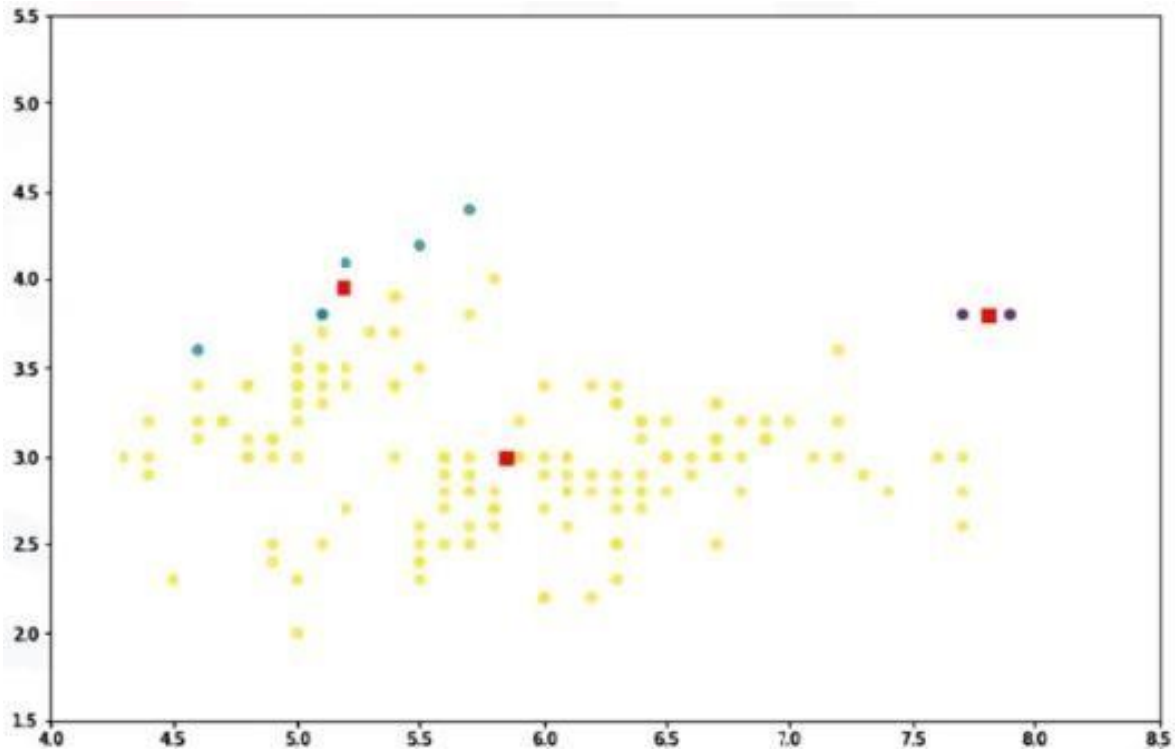
k-Means Clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)



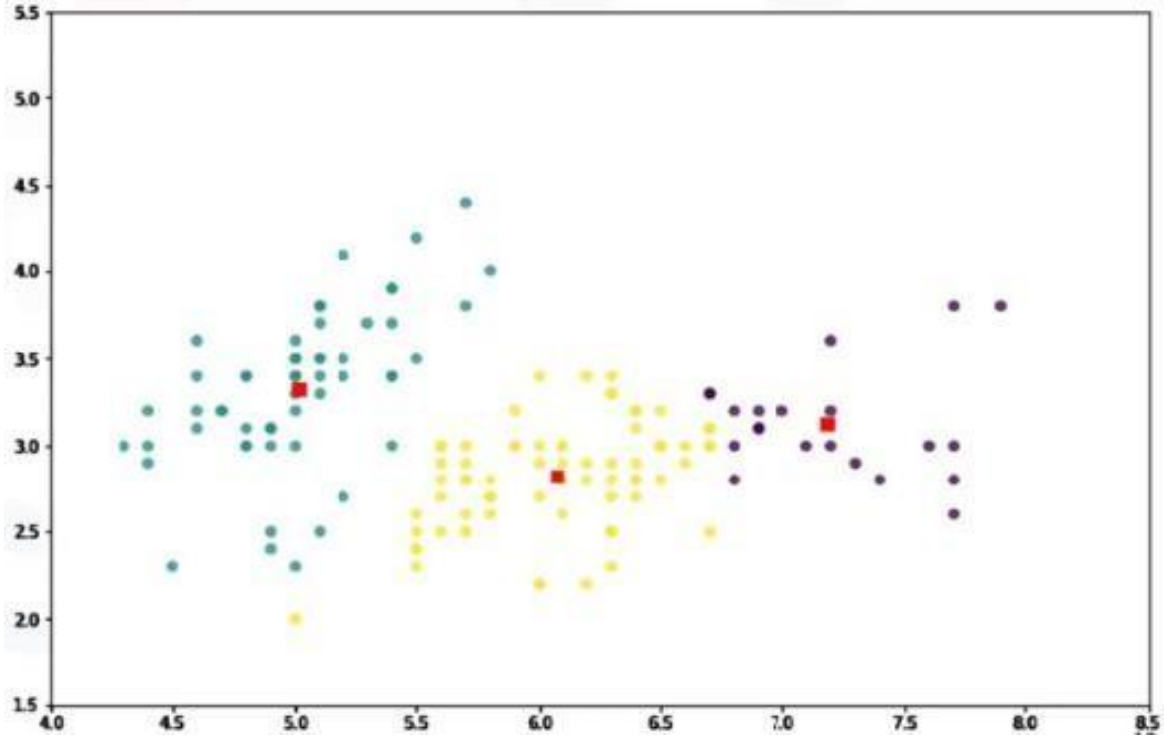
k-Means Clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)



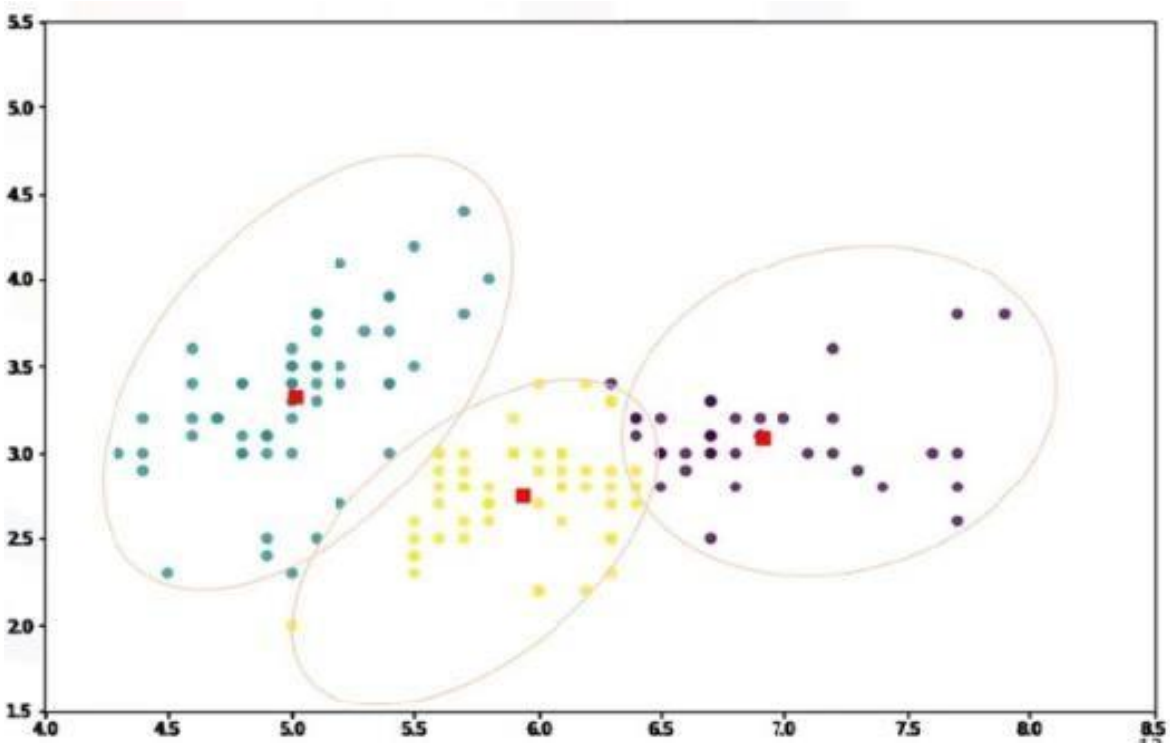
k-Means Clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)



k-Means Clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)

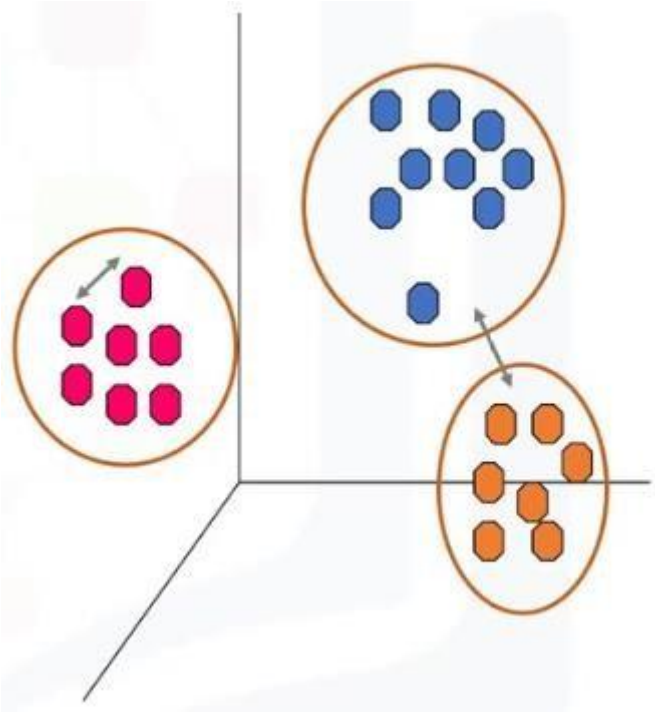


Algoritma k-Means clustering

1. Tempatkan secara acak k centroid, satu untuk setiap cluster
2. Hitung jarak setiap titik dari setiap centroid
3. Tetapkan setiap titik data (objek) ke pusat centroid terdekatnya (membuat cluster)
4. Hitung ulang posisi centroid k
5. Ulangi langkah 2-4, hingga centroid tidak lagi bergerak

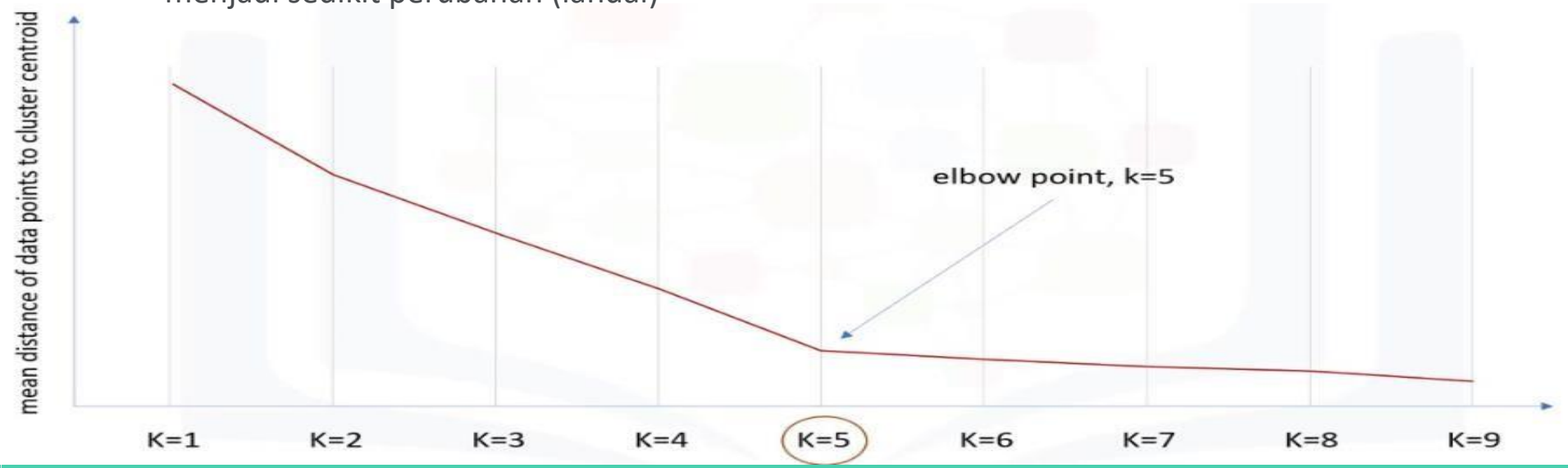
Akurasi k-Means dengan Distance

1. Pendekatan Eksternal
   - Bandingkan cluster dengan nilai sebenarnya
2. Pendekatan Internal
   - Rata-rata jarak antara titik data dalam sebuah cluster



Memilih k

- Dengan menghitung jarak rata-rata antara titik data ke centroid, maka dapat ditentukan jumlah centroid k yang optimal
- Jumlah k yang optimal terletak pada elbow point
- Hal tersebut dikarenakan setelah adanya penurunan rata-rata jarak yang signifikan berubah menjadi sedikit perubahan (landai)



Rangkuman

- Machine Learning (ML) merupakan ilmu yang mampu memberikan komputer kemampuan untuk belajar tanpa diprogram secara eksplisit.
- Model ML yaitu supervised and unsupervised learning.
- Supervised learning adalah mengajarkan model dan melatihnya dengan beberapa data dari dataset yang berlabel.
- Tipe supervised learning adalah klasifikasi dan regresi.
- Unsupervised learning adalah algoritma yang melatih dataset, dan menarik kesimpulan pada data tidak berlabel.
- Tipe unsupervised learning adalah clustering.