



# DIGITAL TALENT SCHOLARSHIP 2019



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

# Machine Learning: Classification

M. Ramli & M. Soleh





# Apa itu Classification

- Dalam machine learning, **classification** termasuk pendekatan **supervised-learning**.
- Tugasnya melakukan kategorisasi data yang tidak diketahui ke dalam kelompok *class* yang diskrit
- Target pengelompokkan data tersebut berupa kategori, contoh :
  - Kelompok A atau B
  - Warna merah atau biru
  - Ya atau tidak

# Apa Tugasnya Classification?

- Diberikan sebuah himpunan observasi berupa data tabel, lengkap dengan label *class*-nya,

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

Data berupa  
Kategori

- Classification harus menentukan *class* dari observasi baru yang belum diberikan label class.

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
37	2	16	10	130	9.3	10.23	3.21	

# Contoh Kasus Classification

- Sebuah bank harus menentukan apakah salah satu nasabahnya yang ingin melakukan peminjaman uang dapat mengembalikan pinjamannya atau tidak.

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

- Yang bisa dilakukan bank ialah menganalisa data nasabah sebelumnya, nasabah dengan ciri-ciri seperti apa yang kemungkinan memiliki masalah dalam melakukan pengembalian pinjaman.
- Ciri-ciri yang dimaksud adalah informasi yang dimiliki nasabah, seperti:
  - Umur, edukasi, lama bekerja, pendapatan, debit yang dimiliki, dll.

# Contoh Kasus Classification

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
37	2	16	10	130	9.3	10.23	3.21	0

Modelling

Classifier model

Input

Kategorisasi



- Berdasarkan informasi-informasi tersebut, sebuah classifier model dapat dibuat yang akan digunakan untuk melakukan klasifikasi data baru!
- Jika nasabah baru muncul, classifier model dapat melakukan kategorisasi nasabah tersebut.



# Target / Class dari Classification

- Dari contoh yang sebelumnya, kita bisa melihat bahwa target / class dari nasabah tersebut berupa binary.
  - Memungkinkan melakukan pinjaman (1)
  - Tidak memungkinkan melakukan pinjaman (0)
- Pada dasarnya, classification tidak hanya dapat melakukan *binary classification* tetapi juga *multiclass classification*.
- Sebagai contoh:
  - Kelompok A, atau B, atau C.
  - Kucing, Harimau, atau Macan
  - Bunga anggrek, melati, atau bakung.



# Contoh Multiclass Classification

Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

- Terdapat data pasien yang mengalami sebuah penyakit yang sama.
- Namun karena kondisi tubuh pasien tersebut berbeda-beda, obat yang dianjurkan oleh dokterpun harus sesuai dengan kondisi tubuhnya.
- Ada tiga tipe obat yang tersedia: drugC, drugX, dan drugY.



# Contoh Multiclass Classification

Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

Age	Sex	BP	Cholesterol	Na	K	Drug
36	F	LOW	HIGH	0.697	0.069	

Modelling

Classifier model



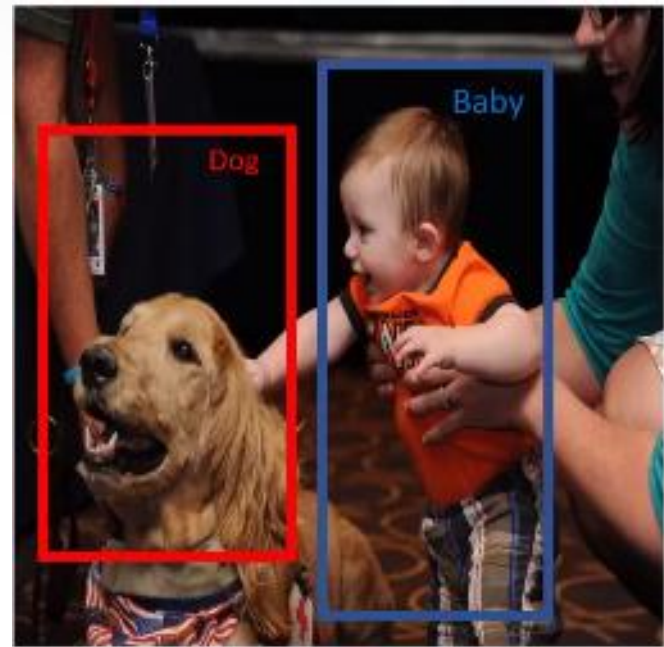
Input

Kategorisasi

- Kita bisa membuat classifier model berdasarkan data pasien yang tersedia.
- Jika pasien baru muncul, classifier model dapat melakukan kategorisasi, yang mana dari ketiga obat tersebut yang cocok untuk pasien tersebut.



# Contoh Penggunaan Classification





# Beberapa Algoritma Classification

- K-Nearest Neighbor
  - Decision Tree
- } 1 Hari (Hari ini)
- 
- Logistic Regression
  - Support Vector Machines
- } 1 Hari (Hari berikutnya)
- 
- Neural Networks
- 2 Hari (Setelah Ujian Tengah)

Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

# Pengukuran Evaluasi dalam Classification

M. Ramli & M. Soleh





Bagian Satu

# Memahami Pengukuran Evaluasi

# Training dan Testing Data

- Pengukuran Evaluasi (Evaluation Metrics) mendeskripsikan performa dari *model classifier* kita.
- Untuk membuat Evaluation Metrics, data training dibagi menjadi dua:

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn	
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1	} Training Data
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1	
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0	
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0	} Testing Data
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0	
5	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1	

- Training data = Membuat *model classifier*.
- Testing data = Memeriksa akurasi dari classifier



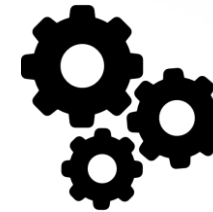
# Training dan Testing Data

## Training Data

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0

Modelling

Classifier model



Seberapa akurat model kita?

Class yang diprediksi

churn
1
0
1

$\hat{y}$

Class yang sebenarnya

churn
0
0
1

$y$

## Testing Data

3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0
5	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1

Prediksi



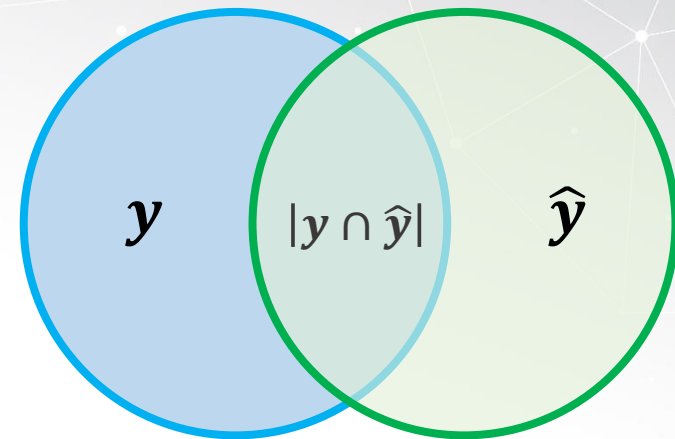
Bagian Dua

# Beberapa Evaluation Metrics yang Umum

# Jaccard Index

- $y$  = Label/Class yang sebenarnya
- $\hat{y}$  = Label/Class yang diprediksi

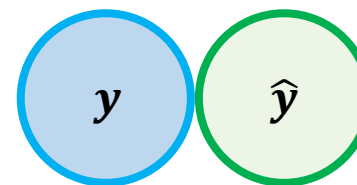
$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$



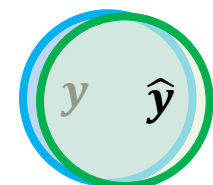
$y = [0,0,0,0,0,1,1,1,1,1]$

$\hat{y} = [1,1,0,0,0,1,1,1,1,1]$

$$J(y, \hat{y}) = \frac{8}{10 + 10 - 8} = 0.66$$



$$J(y, \hat{y}) \approx 0$$

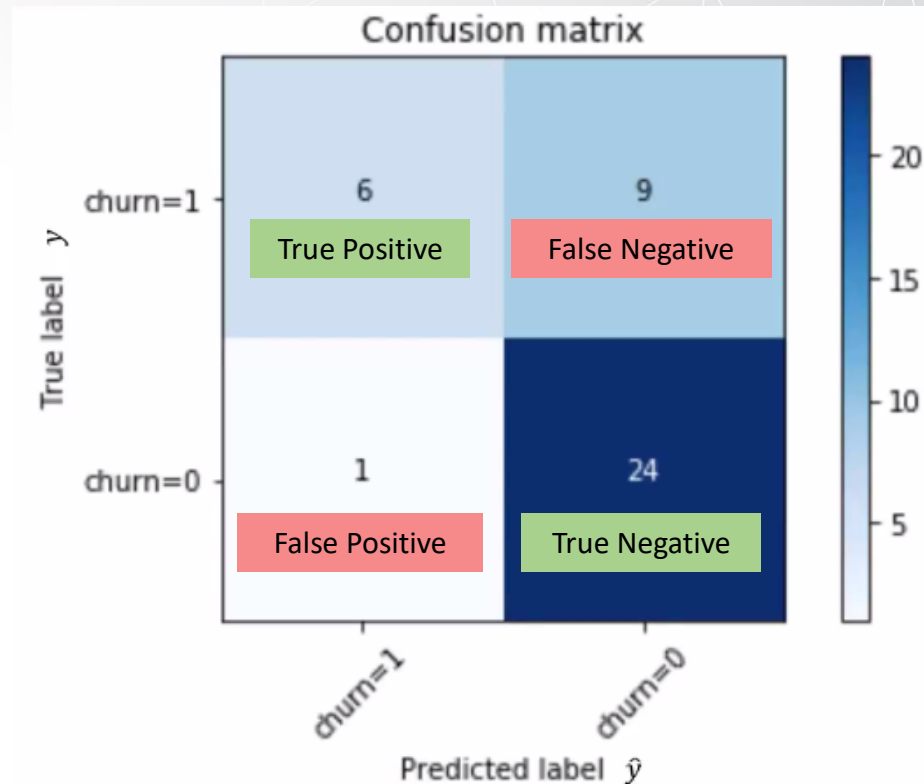


$$J(y, \hat{y}) \approx 1$$

Higher Accuracy

# F1 Score

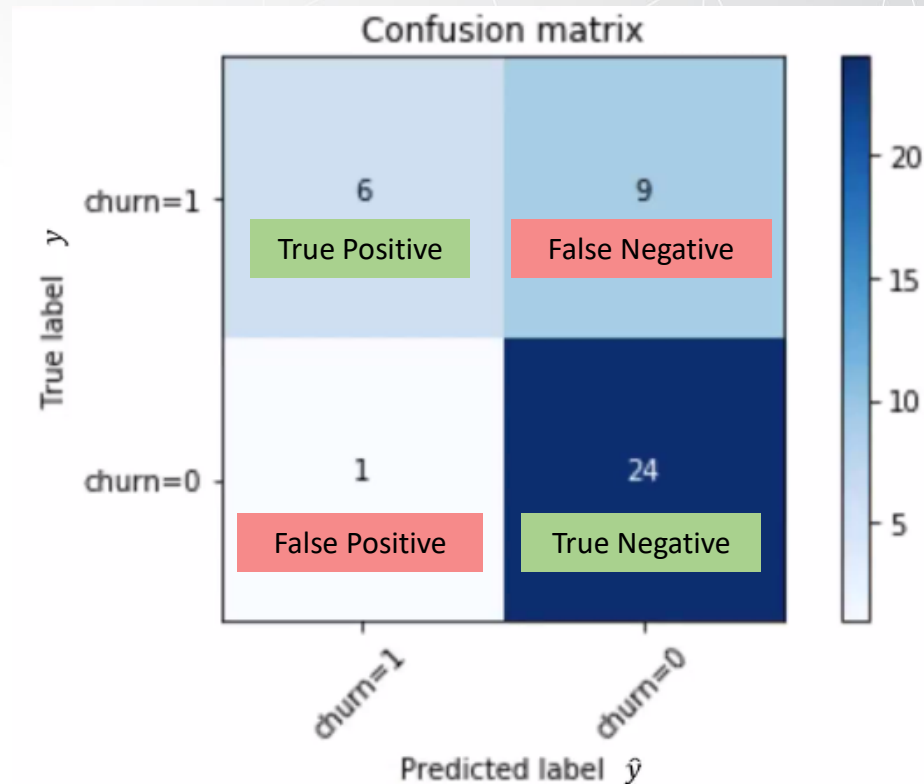
- Cara membaca Confusion Matrix
- True Positive:
  - Diprediksi *True*
  - Class sebenarnya *True*
- False Negative:
  - Diprediksi *False*
  - Class sebenarnya *True*
- False Positive:
  - Diprediksi *True*
  - Class sebenarnya *False*
- True Negative
  - Diprediksi *False*
  - Class sebenarnya *False*



# F1 Score

## Confusion Matrix Evaluation Metric:

- Precision =  $\frac{TP}{(TP+FP)}$
- Recall =  $\frac{TP}{(TP+FN)}$



	precision	recall
Churn = 0	0.73	0.96
Churn = 1	0.86	0.40

# F1 Score

## Confusion Matrix Evaluation Metric:

- Precision =  $\frac{TP}{(TP+FP)}$
- Recall =  $\frac{TP}{(TP+FN)}$
- F1-Score =  $2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$

Harmonic Average dari Prec. & Rec.

**F1-score:**

0.00

...

0.20

....

0.55

....

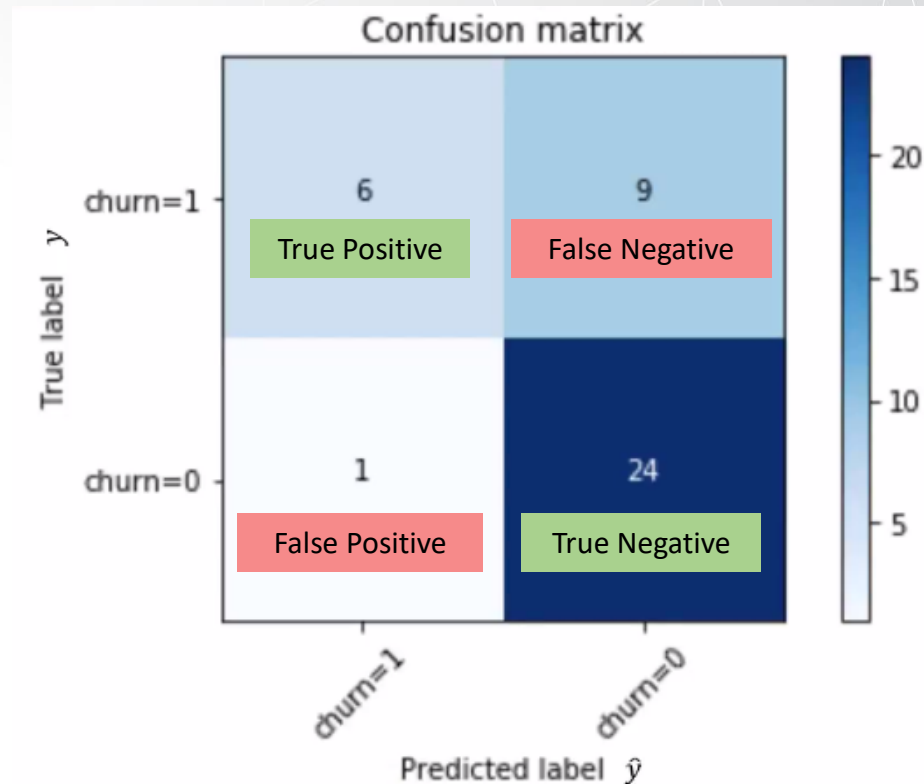
0.83

...

1.00

Higher Accuracy

	precision	recall	f1-score
Churn = 0	0.73	0.96	0.83
Churn = 1	0.86	0.40	0.55





Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

# Classification : K-Nearest Neighbor

M. Ramli & M. Soleh





Bagian Satu

# Intuisi dibalik K-Nearest Neighbor

# Contoh Kasus KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

- Bayangkan sebuah perusahaan provider telekomunikasi.
- Perusahaan tersebut mengumpulkan data pelanggannya lengkap dengan penggunaan layanan yang dipilihnya.

# Contoh Kasus KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

- Berdasarkan data tersebut, perusahaan seharusnya dapat menemukan pola:
- Pelanggan seperti apa yang senang memilih basic, e-service, plus, dan total.



# Contoh Kasus KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

- Ini merupakan permasalahan klasifikasi :
  - Diberikan sebuah dataset, lengkap dengan label yang telah ditentukan (layanan),
  - Kita harus membuat model yang dapat digunakan untuk melakukan klasifikasi,
  - Sebagai contoh: akan memilih layanan mana pelanggan baru ini?

# Contoh Kasus KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

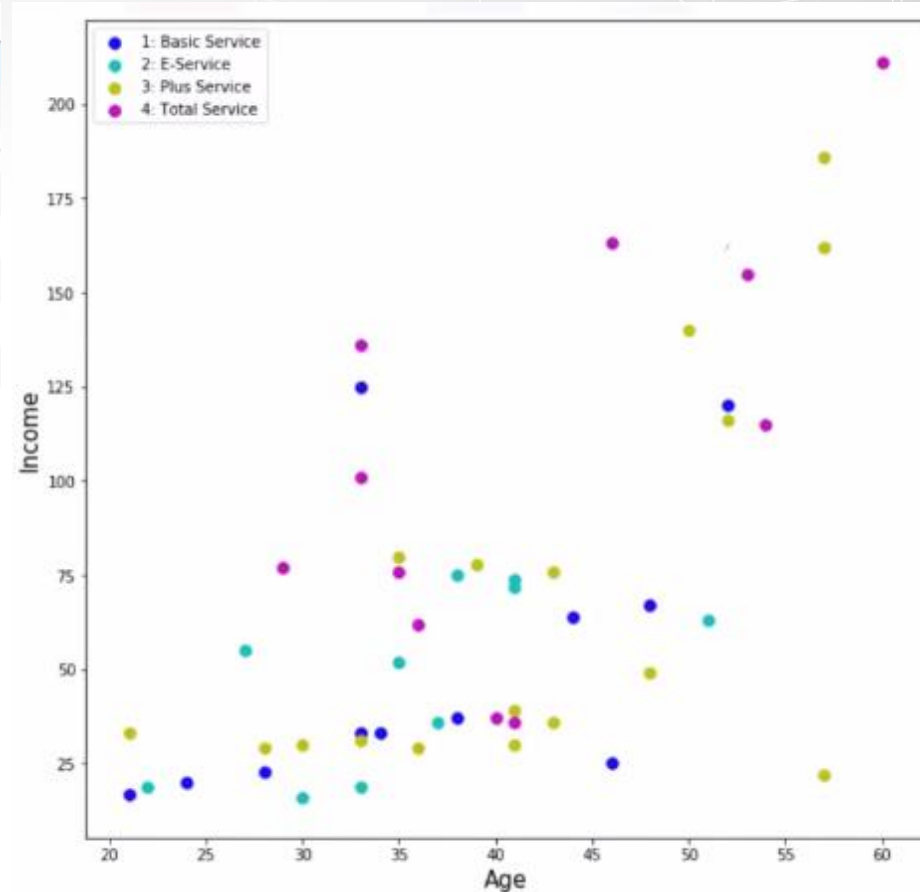
- Tujuan kita adalah membuat model classifier yang akan melakukan klasifikasi pelanggan baru tersebut.
- Kita akan gunakan salah satu algoritma machine learning: K-Nearest Neighbor



# Menentukan Class dengan 1<sup>st</sup> KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

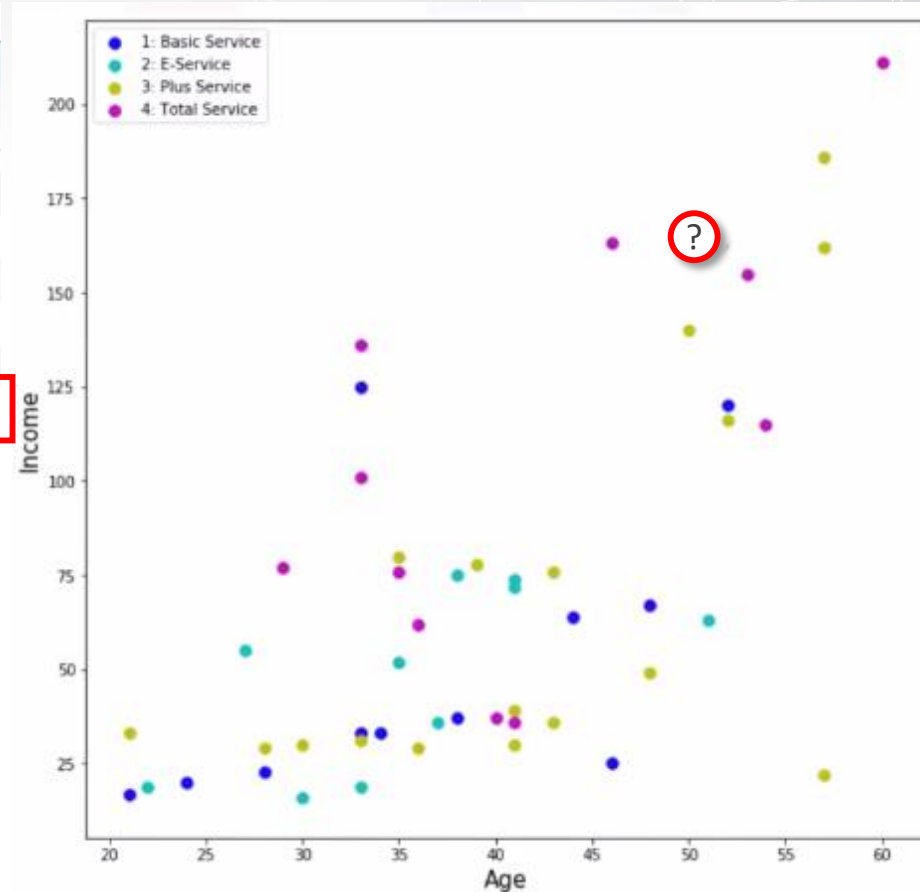
- Untuk penyederhanaan, kita hanya akan menggunakan dua variable untuk melakukan klasifikasi.
- *Income* dan *Age*
- Kemudian kita plot kedua variable kedalam grafik.



# Menentukan Class dengan 1<sup>st</sup> KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

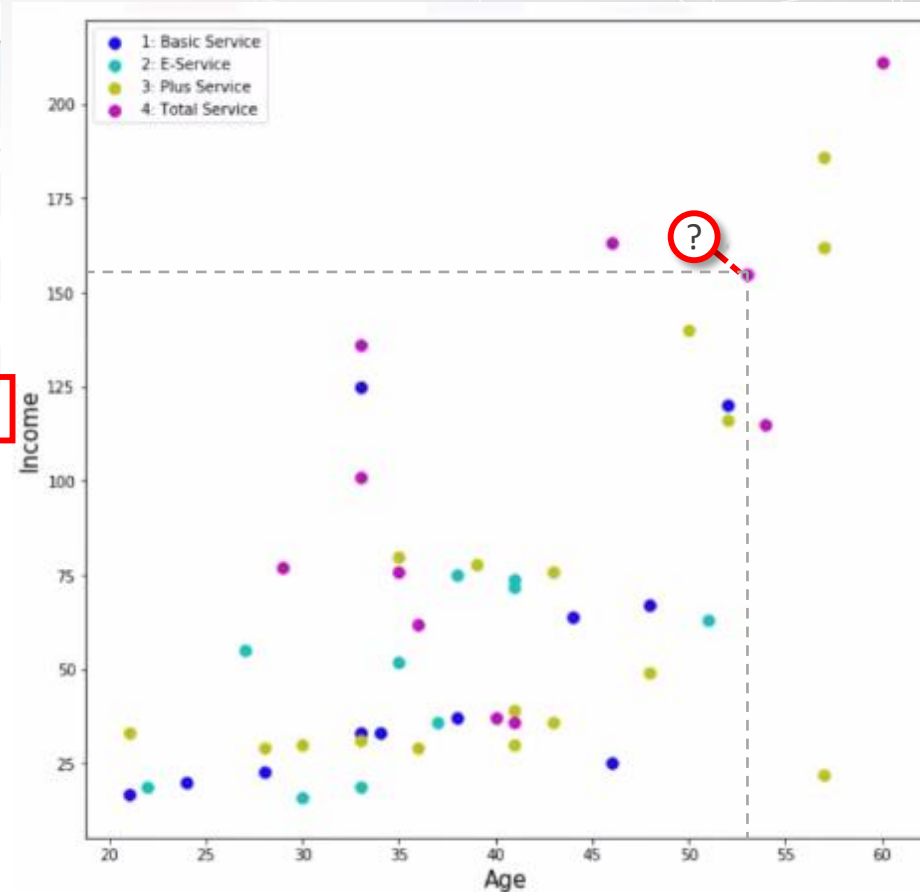
- Sekarang, andai pelanggan baru muncul, dengan:
  - Age = 50
  - Income = 166
- Termasuk kategori manakah pelanggan baru ini?



# Menentukan Class dengan 1<sup>st</sup> KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

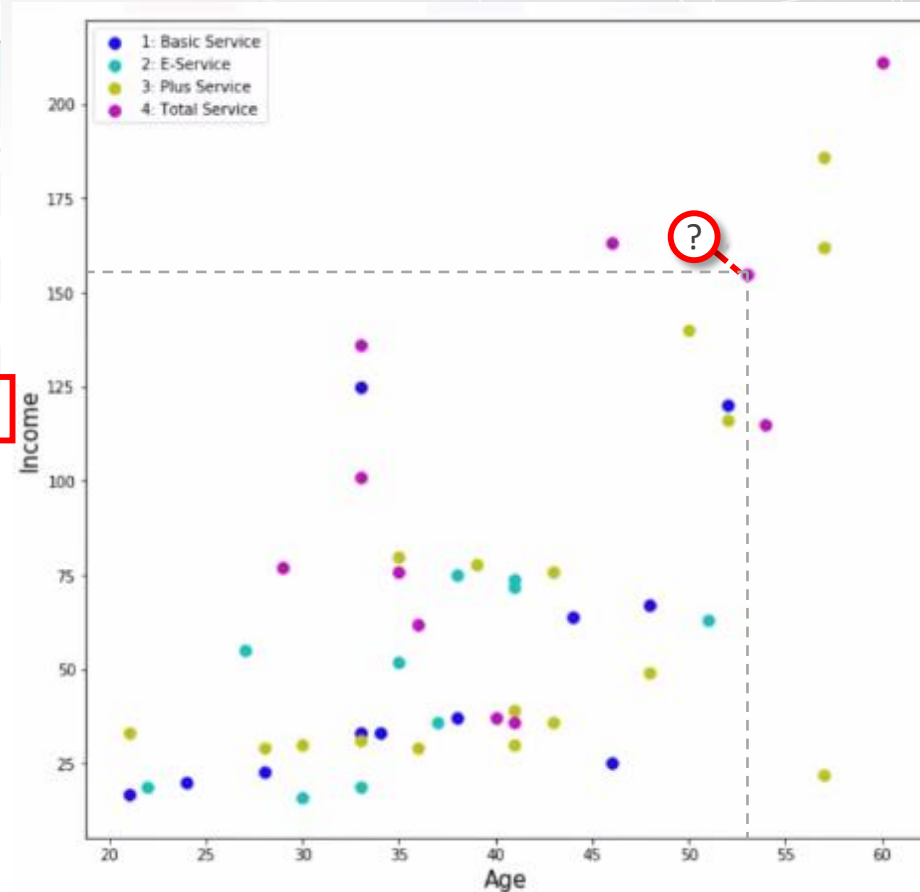
- Salah satu logika paling sederhana:
  - Data terdekat dengan kita memiliki ciri-ciri (umur dan pendapatan) yang serupa dengan kita.
- Mari kita lihat data yang terdekat dengan pelanggan baru.



# Menentukan Class dengan 1<sup>st</sup> KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

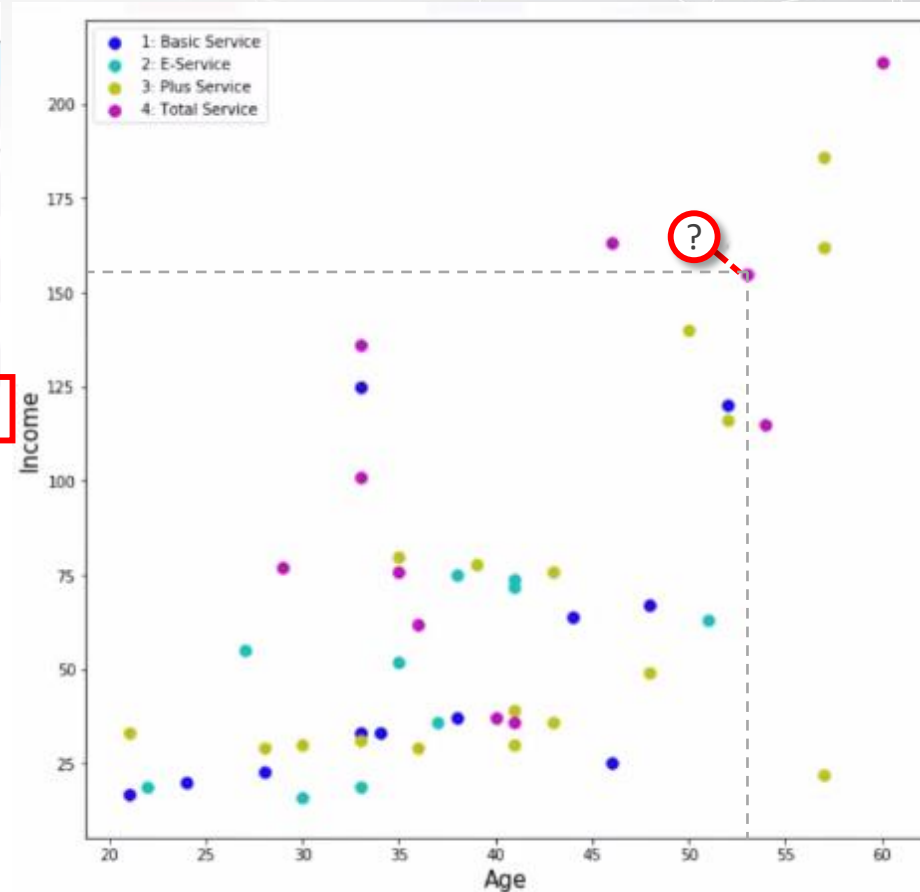
- Data terdekat (tetangga terdekat) memiliki umur  $\approx 56$  dan pendapatan  $\approx 158$ , berlangganan layanan Total Service.



# Menentukan Class dengan 1<sup>st</sup> KNN

region	age	marital	address	income	ed	employ	retire	gender	reside	custcat	
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

- Secara logika, pelanggan baru (umur  $\approx 50$ , pendapatan  $\approx 166$ ), mungkin juga akan berlangganan layanan Total Service.
- Maka kita klasifikasikan pelanggan baru dengan Total Service.



# Menentukan Class dengan 1<sup>st</sup> KNN

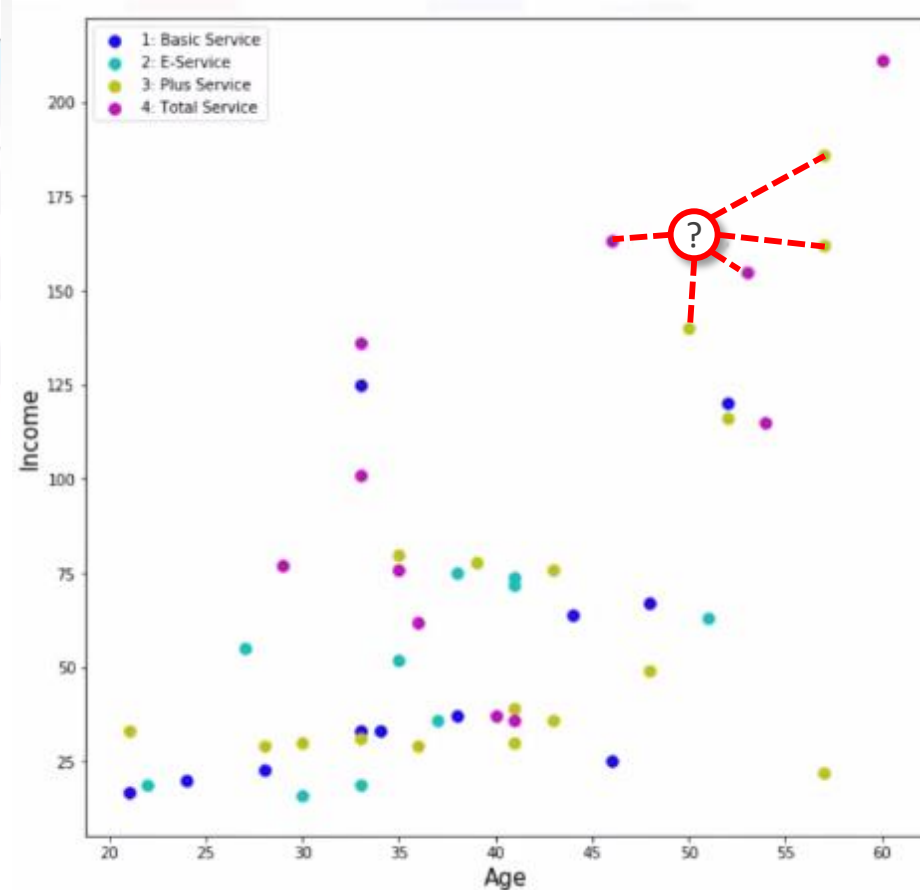
- Proses ini disebut dengan 1<sup>st</sup> KNN.
- Melihat kecenderungan class data baru dari tetangga terdekat kita.
- Namun bagaimana jika tetangga terdekat kita ternyata salah memilih layanan?
  - Misalnya, secara tidak sengaja memilih layanan tersebut, dan sebetulnya tidak menginginkan layanan tersebut.
- Melakukan klasifikasi hanya dengan melihat satu tetangga terdekat saja sangat rentan terjadi kesalahan klasifikasi.



# Menentukan Class dengan 5 KNNs

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

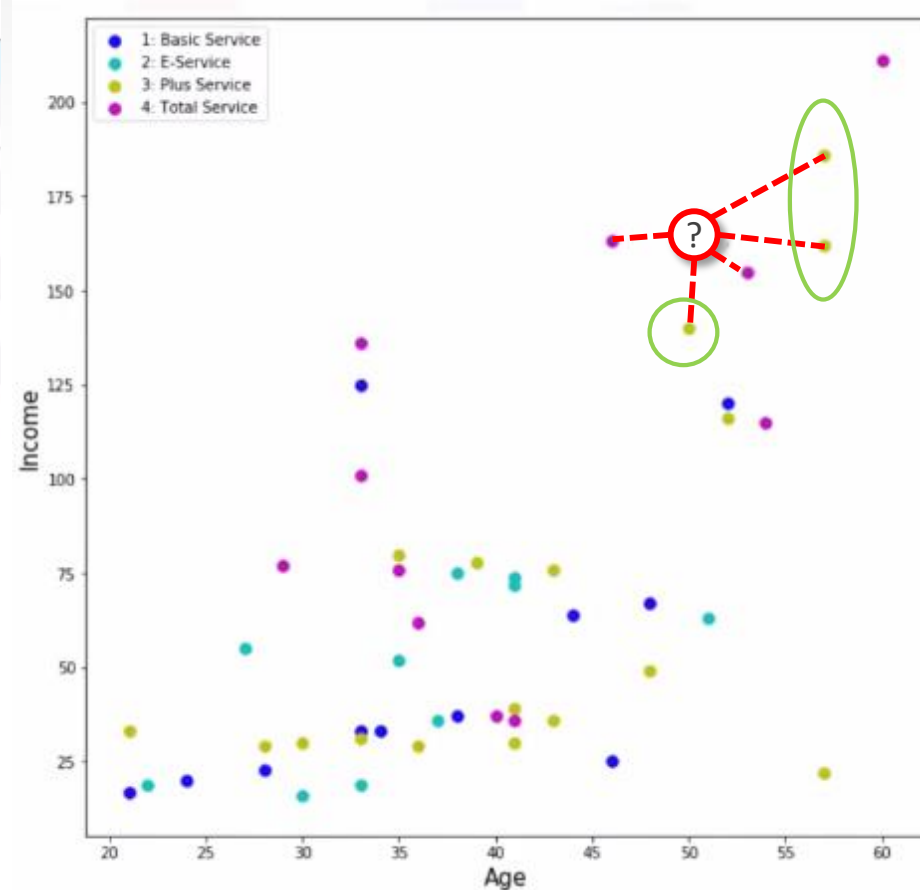
- Sekarang, daripada hanya satu, bagaimana jika kita melihat 5 tetangga terdekat kita?
- Dan penentuan class, berdasarkan mayoritas class disekelilingnya?



# Menentukan Class dengan 5 KNNs

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

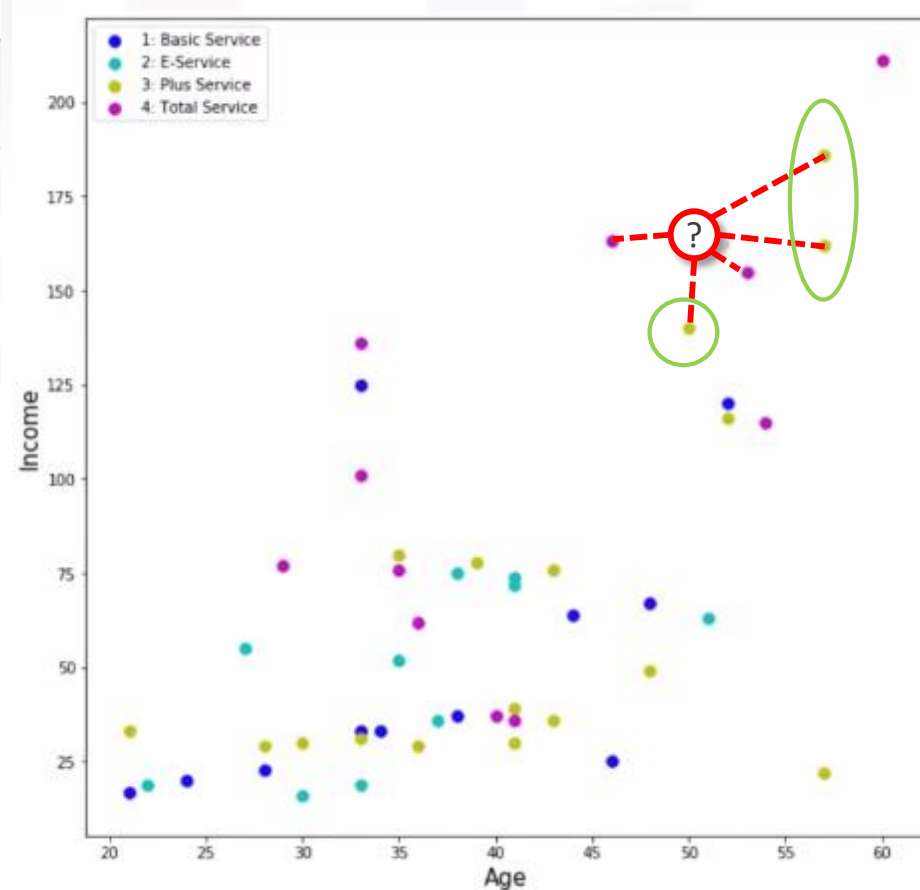
- 3 tetangga terdekat = *Plus Service*
- 2 tetangga terdekat = *Total Service*
- Dari 5 ketetanggaan, pelanggan baru ternyata memiliki kecenderungan memilih *Plus Service*



# Menentukan Class dengan 5 KNNs

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Ini merupakan intuisi dasar dari  
**K-Nearest Neighbor**



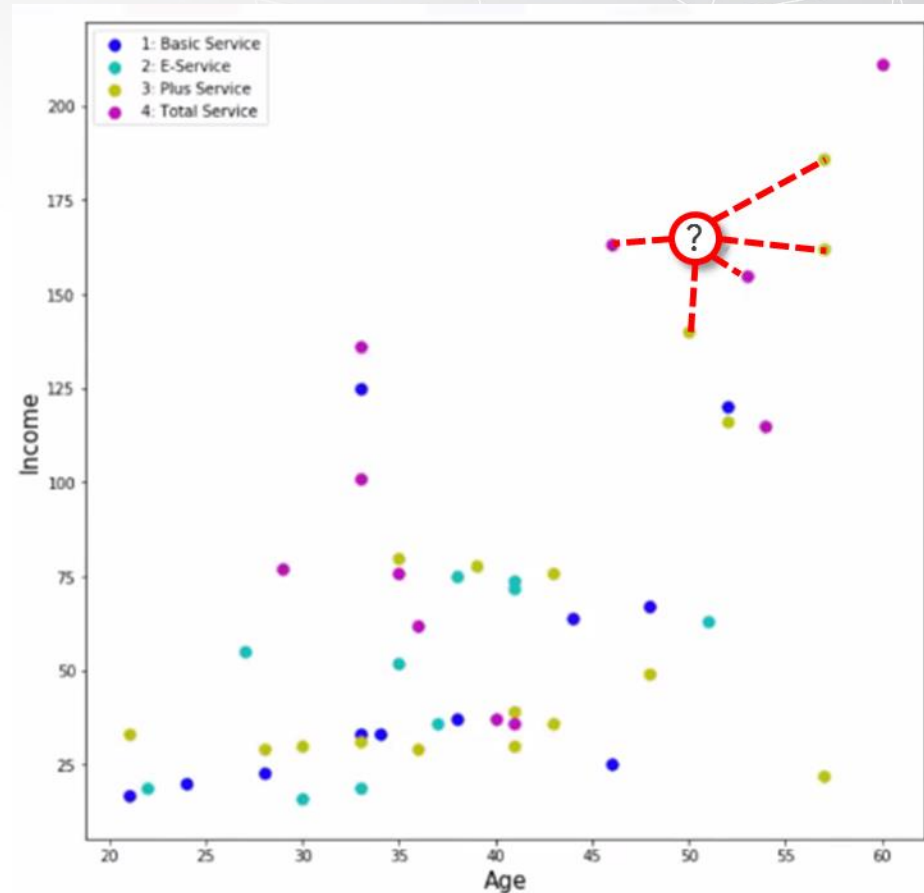


Bagian Dua

# Algoritma K-Nearest Neighbor

# Apa itu K-Nearest Neighbor

- Sebuah **metode** untuk melakukan **klasifikasi** berdasarkan **similaritas terhadap data lainnya**.
- Beberapa data terdekat disebut dengan “**Tetangga**” atau “**Neighbors**”
- Tetangga yang terdekat dianggap memiliki fitur yang serupa dengan data yang dimaksud.

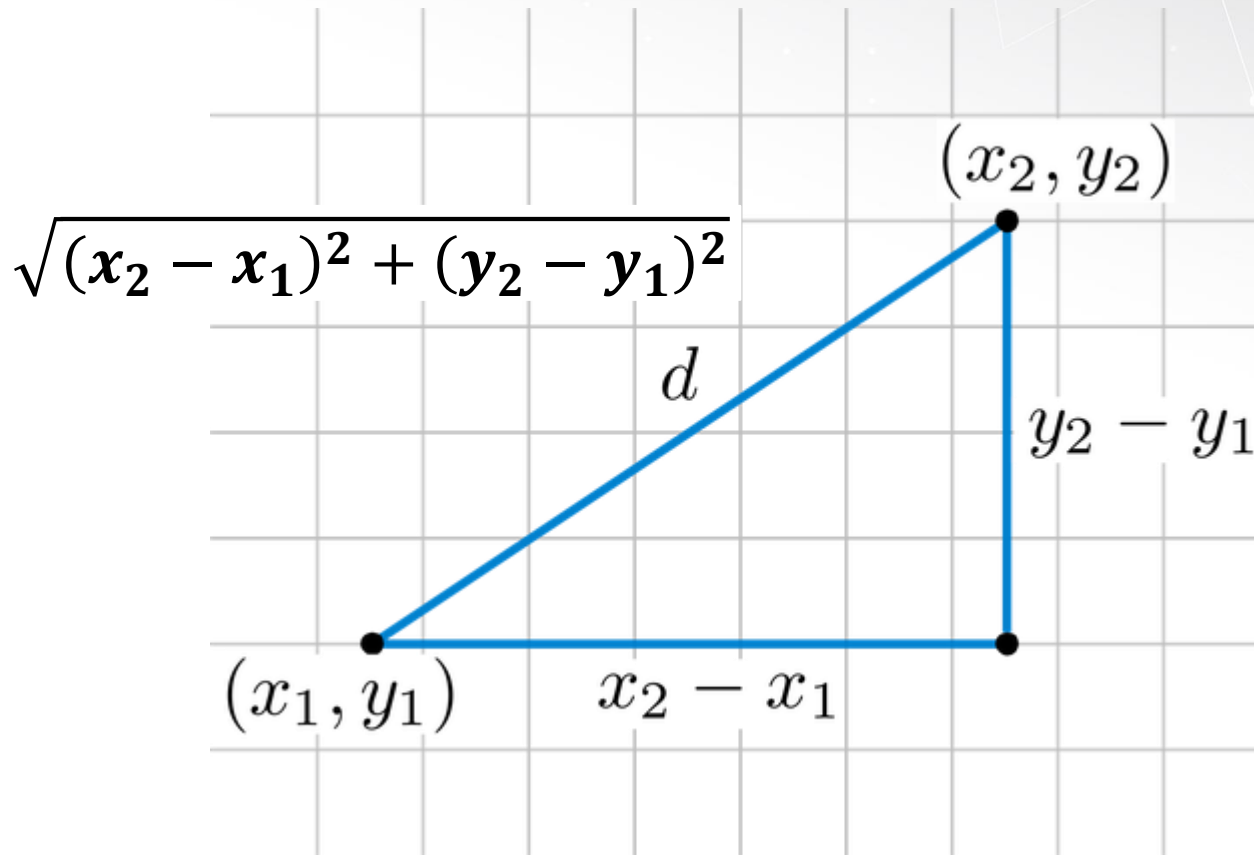




# Algoritma K-Nearest Neighbor

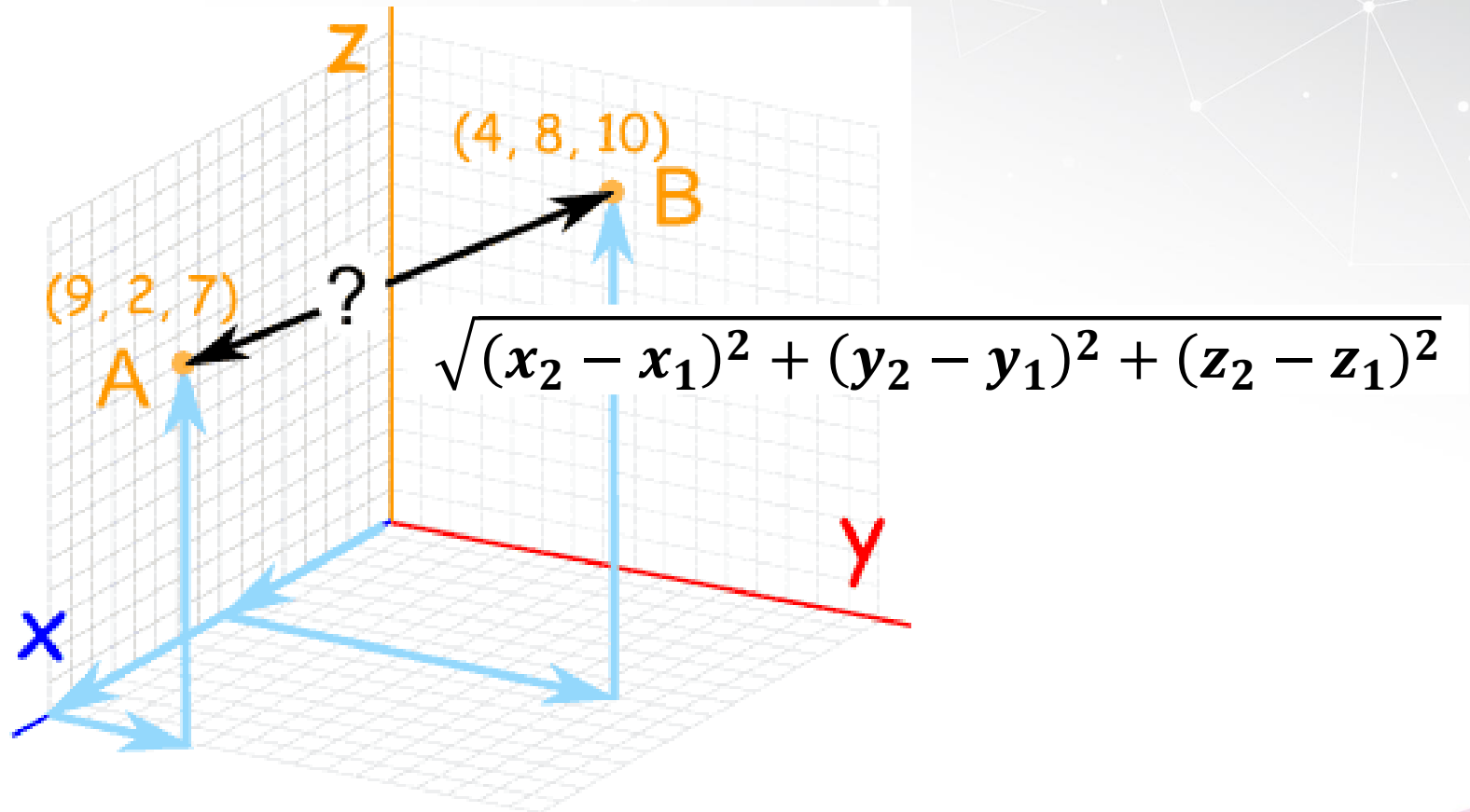
1. Diberikan sebuah data tidak terklasifikasi  $p$ , dan kumpulan data training  $P$  yang telah dilengkapi label classnya .
2. Pilih nilai dari jumlah ketetanggan  $K$ .
3. Hitung jarak antara  $p$  ke seluruh data yang ada dalam  $P$ .
4. Ambil  $K$  observasi yang merupakan data terdekat dengan  $p$ .
5. Klasifikasikan data tersebut dengan mayoritas class dari  $K$ -Tetangga terdekatnya.

# Mencari Jarak Antara 2 Titik di 2D



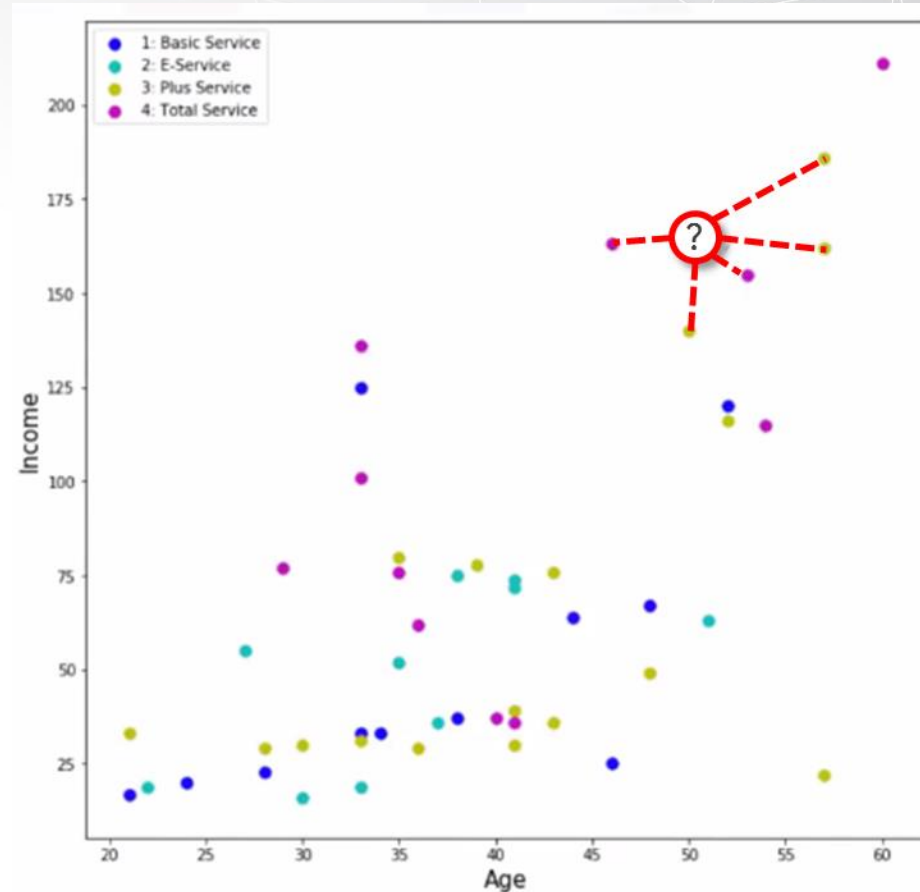


# Mencari Jarak Antara 2 Titik di 2D



# Menentukan Jumlah $K$

- Bagaimana menentukan jumlah  $K$  yang tepat?
  - Terlalu sedikit = rentan outlier
  - Terlalu banyak = tidak relevan



# Training dan Testing Data

- Bagi training data kita menjadi 2 :

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4

Training Data

Testing Data

- Training data = Membuat *model classifier*.
- Testing data = Memeriksa akurasi dari classifier

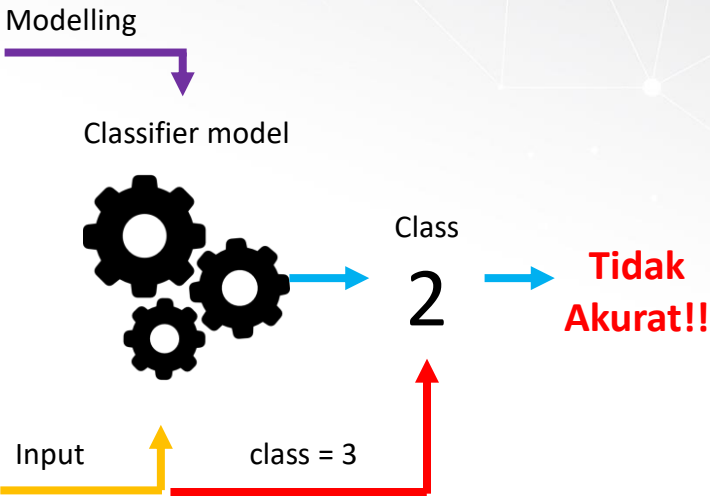
# Training dan Testing Data

## Training Data

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3

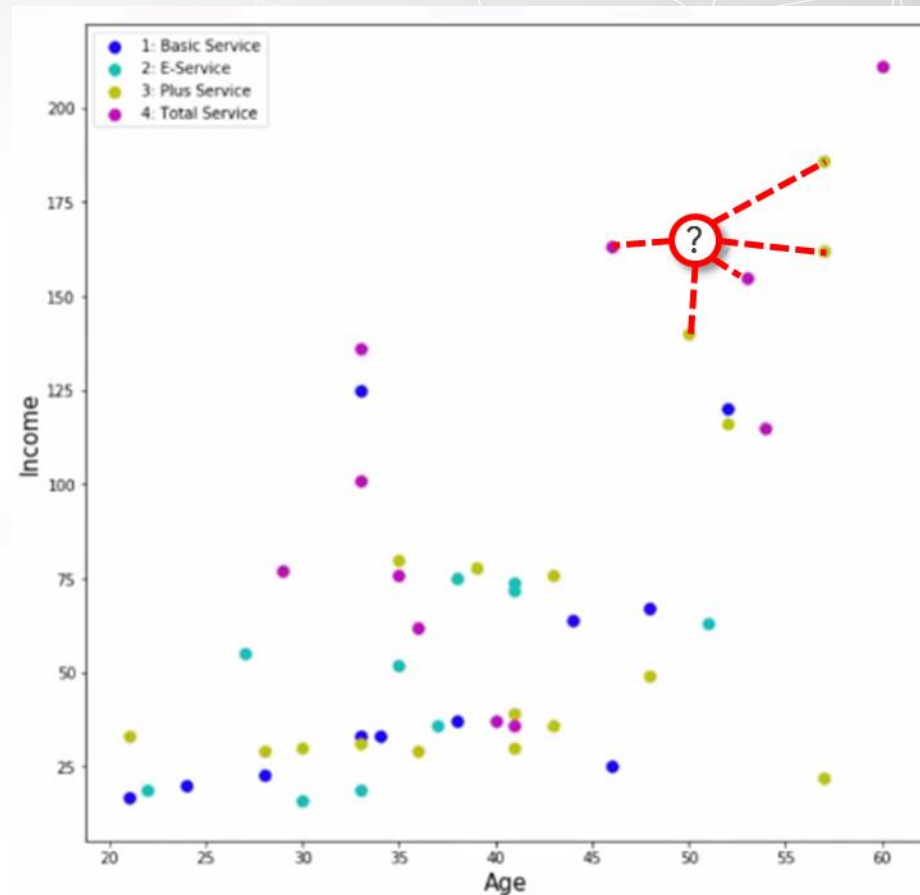
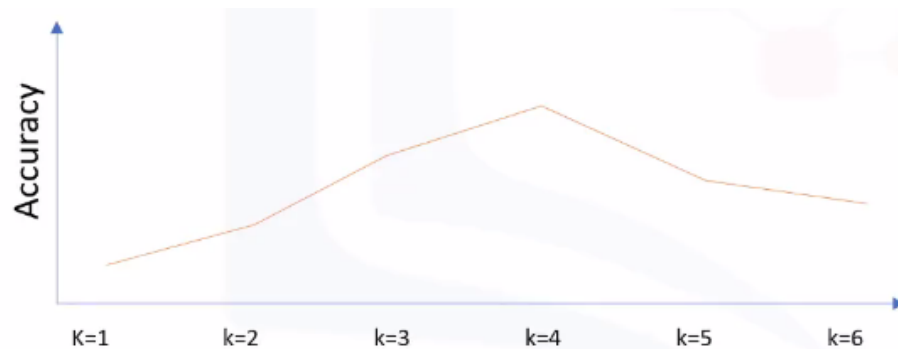
## Testing Data

5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4



# Menentukan Jumlah $K$

- Kita gunakan testing data untuk memeriksa berapakah jumlah  $K$  yang tepat!
- Dari beberapa kali observasi, kita bisa menemukan jumlah  $K$  yang sesuai dengan data kita.



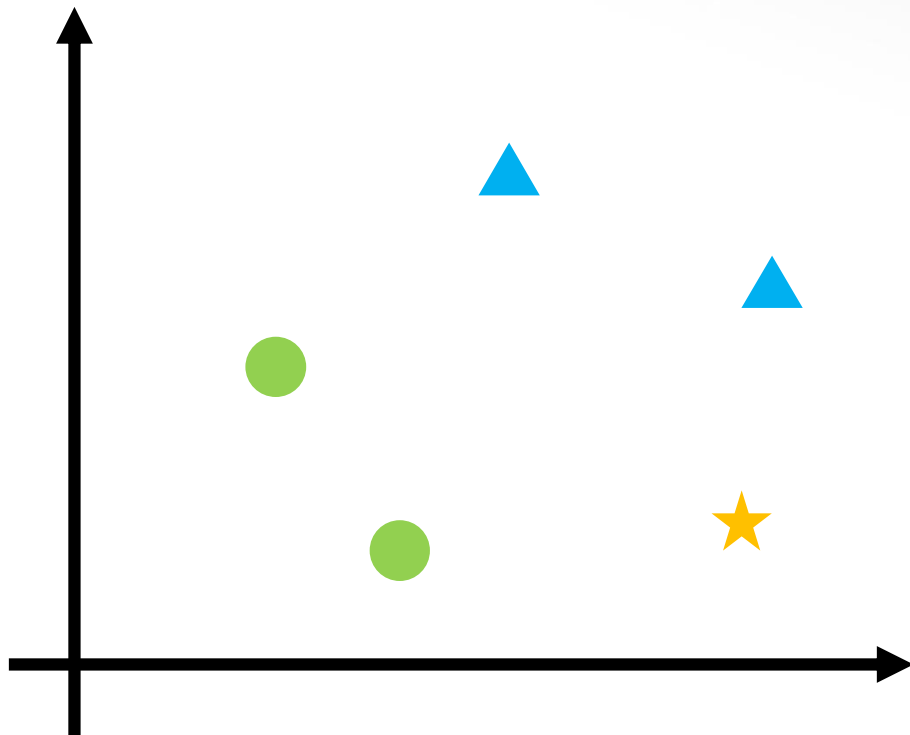


Bagian Tiga

# Membuat Boundary Decision dengan KNN



# Membuat Boundary Decision

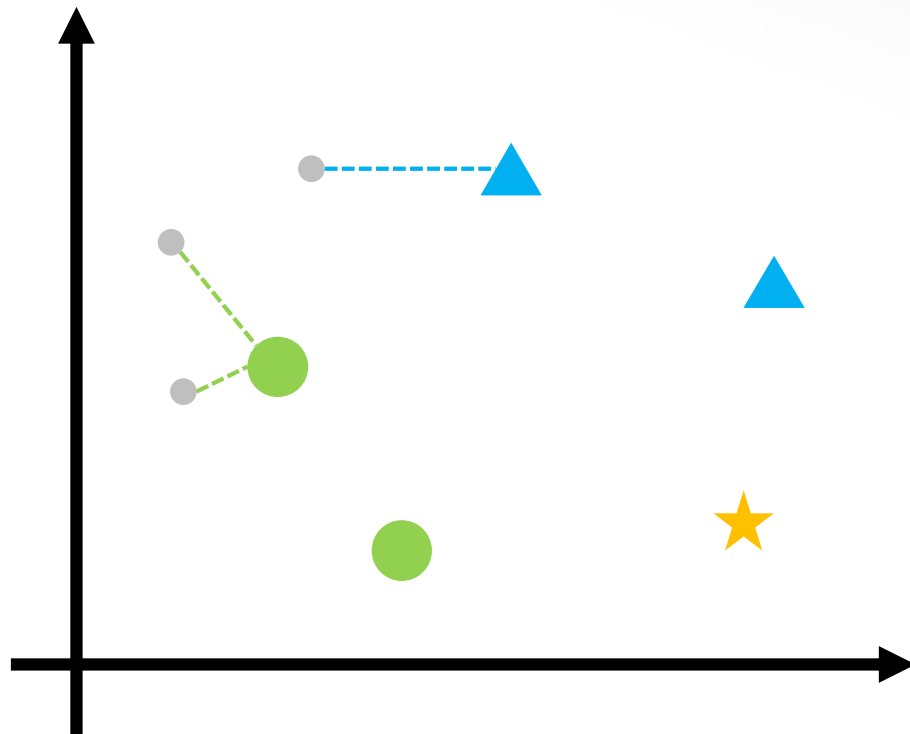


- Diberikan data training seperti berikut
- Gunakan  $K = 1$ .
- Untuk setiap titik yang ada dalam ruang, tentukan data terdekatnya.
- Setiap titik tersebut akan membentuk sepotong-sepotong bagian dalam ruang
  - Voronoi Constellation
  - Setiap titik yang jatuh dalam salah satu bagian Voronoi Constellation pasti memiliki jarak terdekat ke data yang bersangkutan.
- Boundary terbentuk.





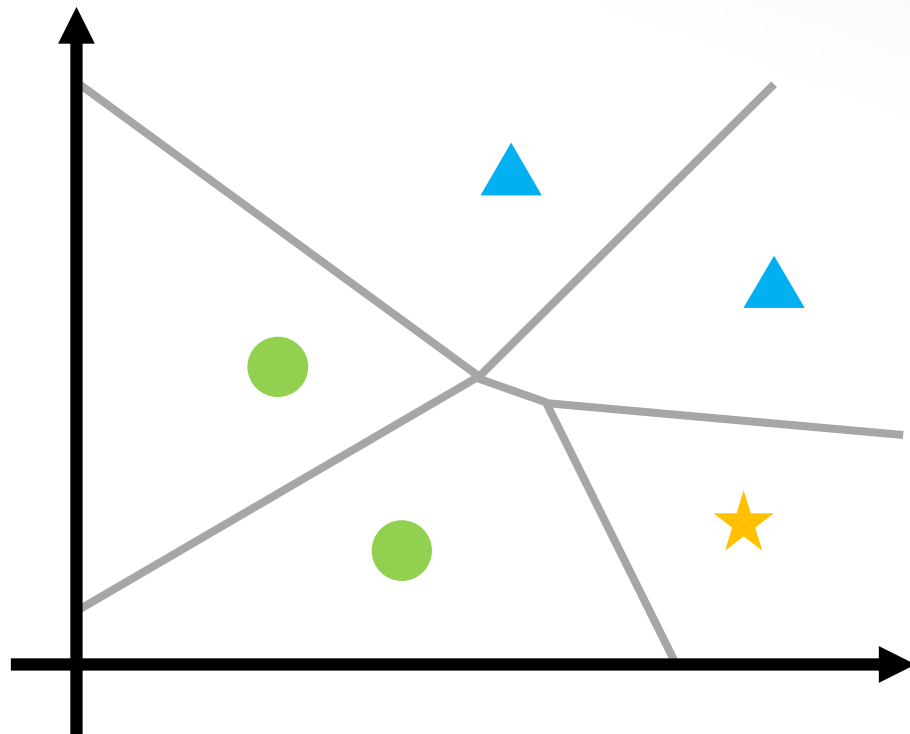
# Membuat Boundary Decision



- Diberikan data training seperti berikut
- Gunakan  $K = 1$ .
- Untuk setiap titik yang ada dalam ruang, tentukan data terdekatnya.
- Setiap titik tersebut akan membentuk sepotong-sepotong bagian dalam ruang
  - Voronoi Constellation
  - Setiap titik yang jatuh dalam salah satu bagian Voronoi Constellation pasti memiliki jarak terdekat ke data yang bersangkutan.
- Boundary terbentuk.



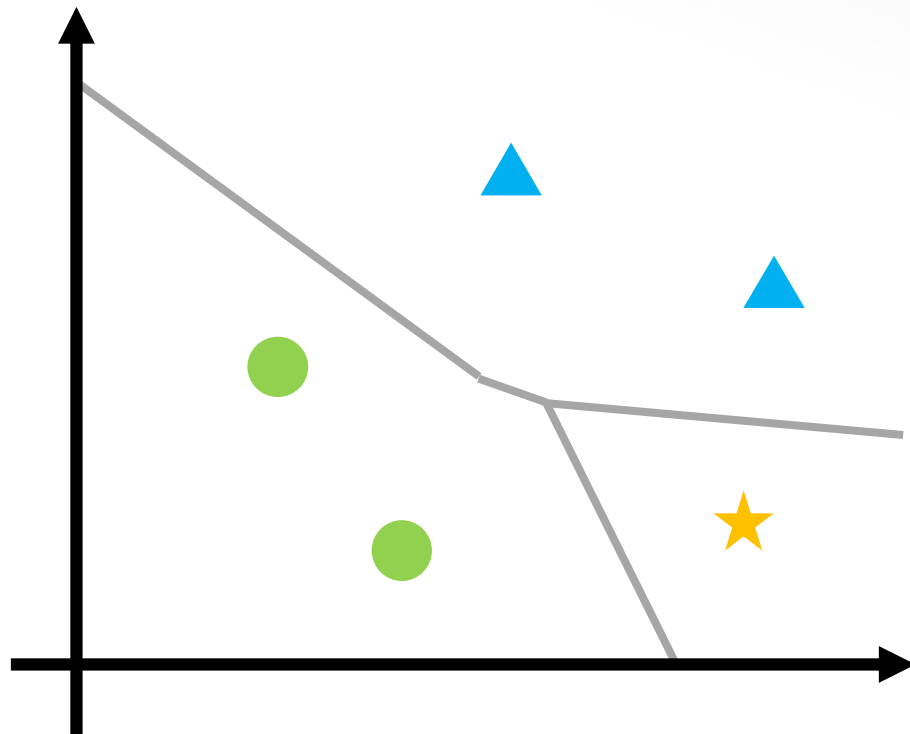
# Membuat Boundary Decision



- Diberikan data training seperti berikut
- Gunakan  $K = 1$ .
- Untuk setiap titik yang ada dalam ruang, tentukan data terdekatnya.
- Setiap titik tersebut akan membentuk sepotong-sepotong bagian dalam ruang
  - Voronoi Constellation
  - Setiap titik yang jatuh dalam salah satu bagian Voronoi Constellation pasti memiliki jarak terdekat ke data yang bersangkutan.
- Boundary terbentuk.



# Membuat Boundary Decision



- Diberikan data training seperti berikut
- Gunakan  $K = 1$ .
- Untuk setiap titik yang ada dalam ruang, tentukan data terdekatnya.
- Setiap titik tersebut akan membentuk sepotong-sepotong bagian dalam ruang
  - Voronoi Constellation
  - Setiap titik yang jatuh dalam salah satu bagian Voronoi Constellation pasti memiliki jarak terdekat ke data yang bersangkutan.
- Boundary terbentuk.

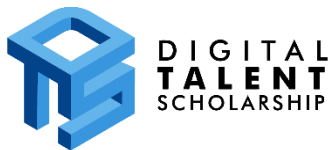






Bagian 4

# Praktikum Lab

ML0101EN-Clas-K-Nearest-neighbors-CustCat-py-v1.ipynb

IKUTI KAMI



-  digitalent.kominfo
-  digitalent.kominfo
-  DTS\_kominfo
-  Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi  
Badan Penelitian dan Pengembangan SDM  
Kementerian Komunikasi dan Informatika  
Jl. Medan Merdeka Barat No. 9  
(Gd. Belakang Lt. 4 - 5)  
Jakarta Pusat, 10110

