Machine Learning Basic Algorithms

Unsupervised Algorithm: K-means ...

Supervised Algorithm: Linear Regression/Regression

Supervised Algorithm: Decision tree ...

Supervised Algorithm: KNN ...

Meta Algorithm: Ensemble ...

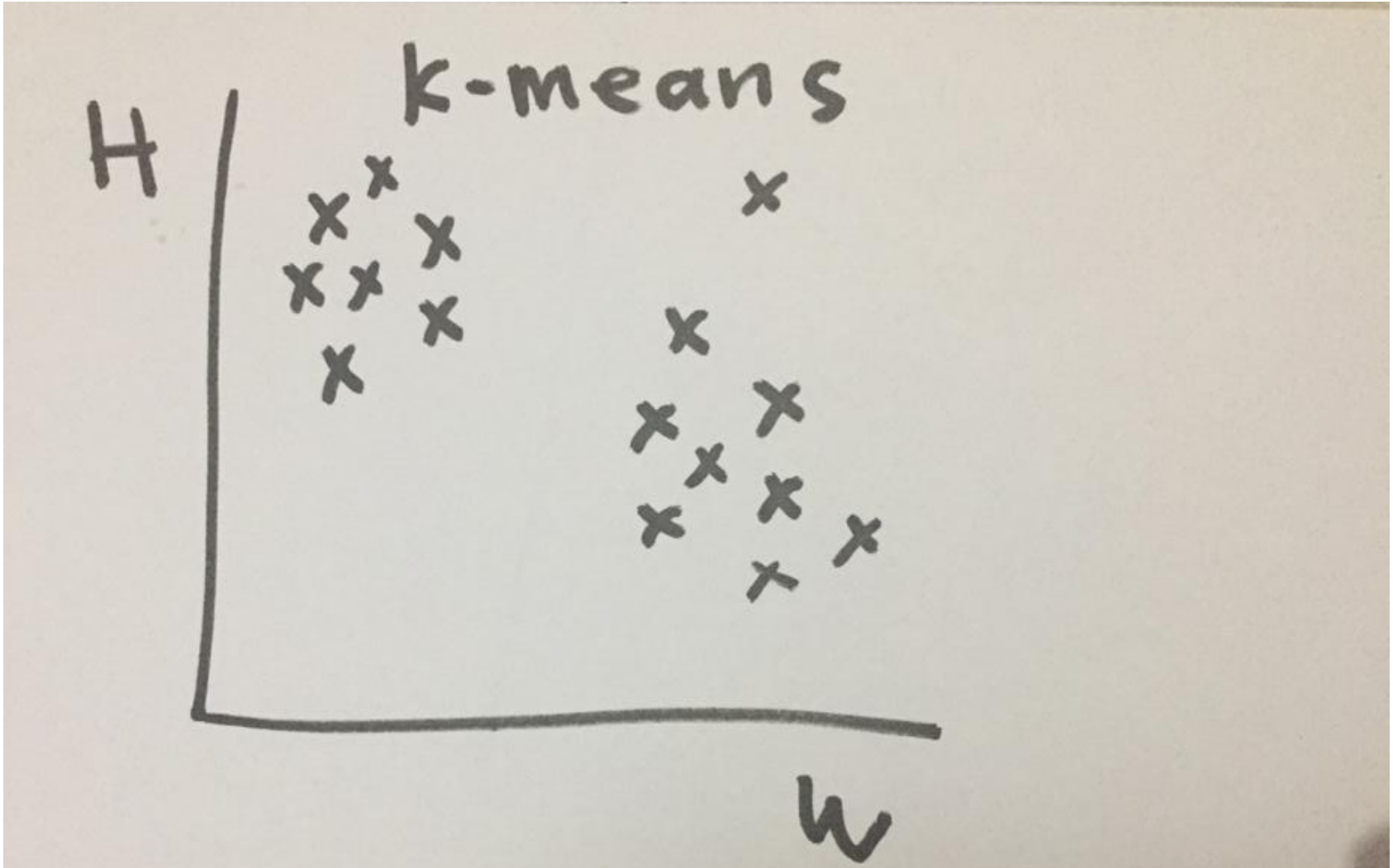

K-MEANS?!
Y NOT A-MEANS
memegenerator.net

K-means

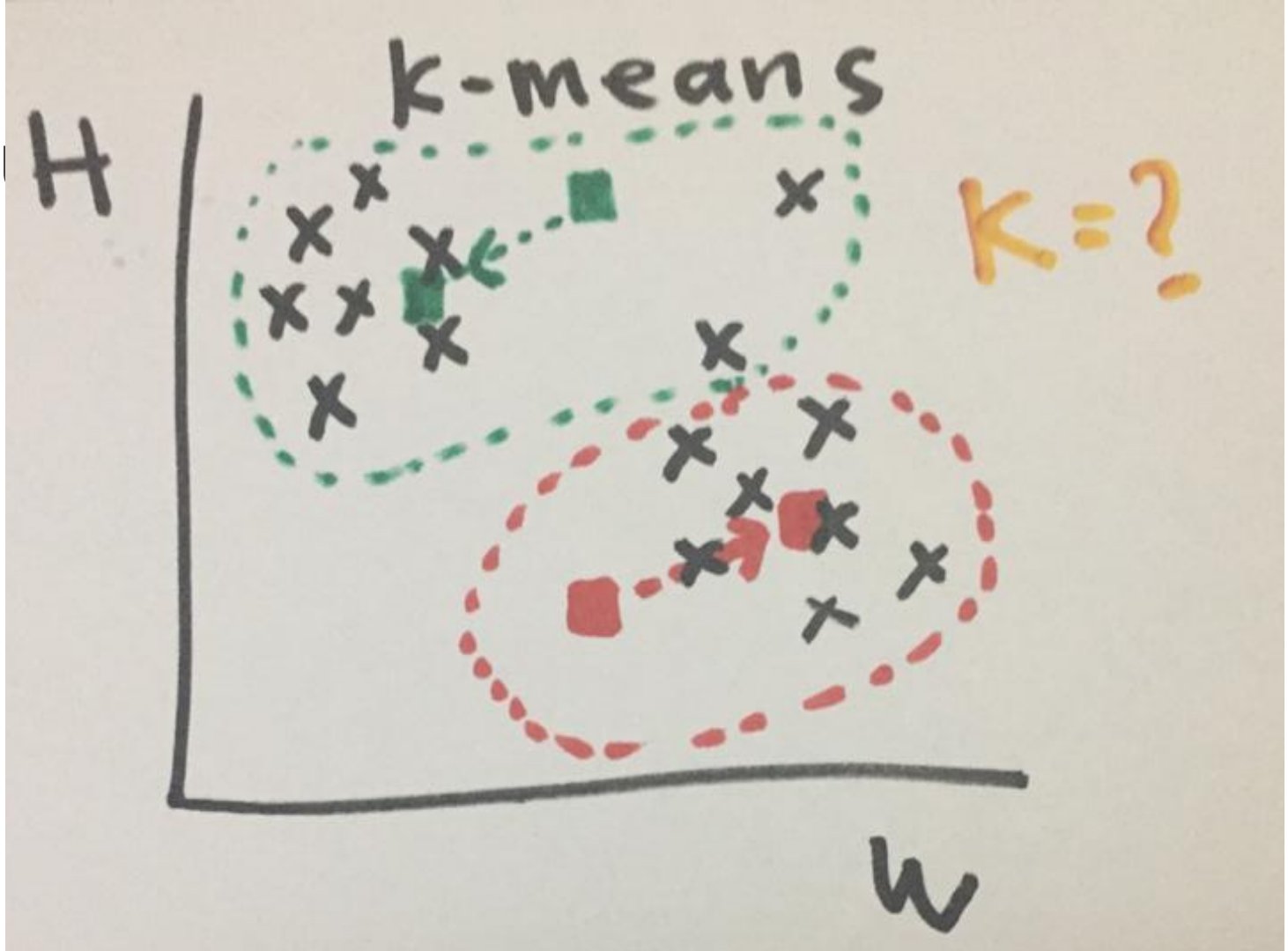
Adalah clustering teknik dimana observasi dikelompokkan ke dalam cluster yang berbeda berdasarkan nilai rata-rata (means) dari tiap cluster.

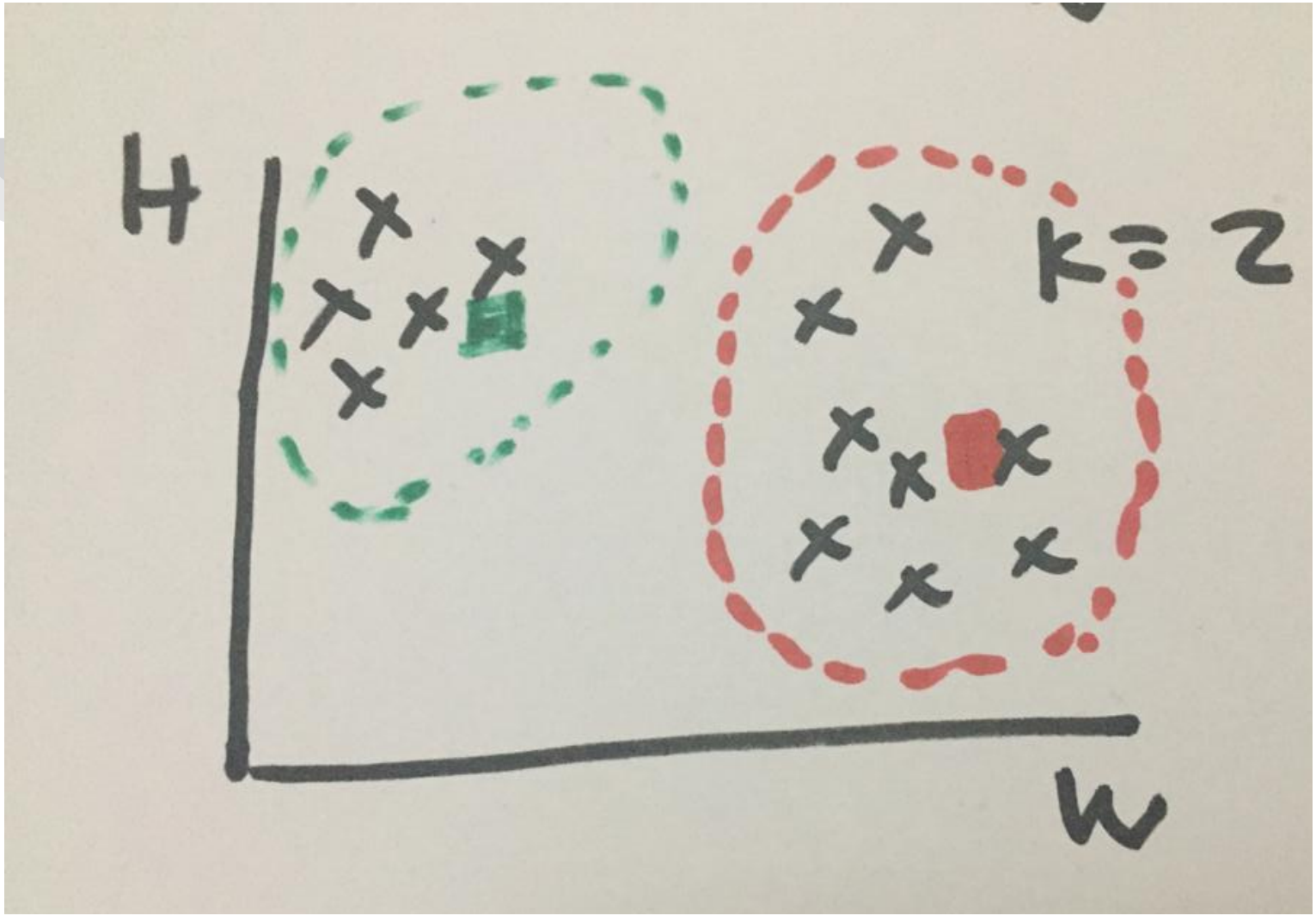
Problem: Seorang pegawai KPPN diminta atasannya untuk membuat kelompok profil dari satker -satker mitra untuk kebijakan baru

Hasil yang diharapkan: kelompok satker yang memiliki kesamaan


k-means
H
W


k-means
H
K=?
W


k-means
H
K=?
W


H
k = 2
w


H
k = 2
W

Unsupervised K-Means

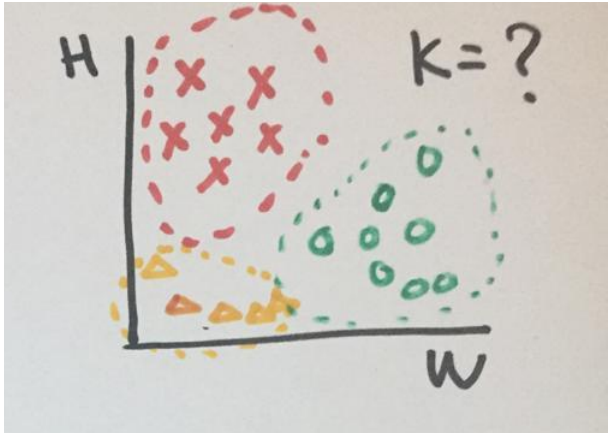
When to stop?

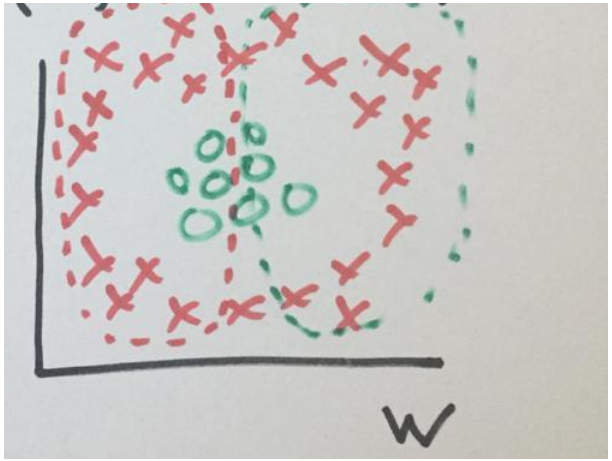
- Sampai centroid (kotak biru dan kotak hijau) tidak bergerak
- Sampai jumlah iterasi yang ditetapkan sebelumnya tercapai

Unsupervised K-Means

Plus : intuitif dan sederhana

Negatif: menentukan jumlah k; terlalu sederhana





Linear Regression

Supervised-Linear Regression

Regresi Linear adalah tehnik yang memprediksi variabel dependent (y) yang bersifat continous berdasarkan variabel independen (x1,x2...xn) berdasarkan fungsi linear.

Problem: Seorang pegawai KPKNL diminta atasannya untuk memperkirakan nilai sewa kantin berdasarkan luas ruangan

Hasil yang diharapkan: nilai sewa kantin dalam rupiah

Supervised-Linear Regression

Task: memperkirakan nilai yhat

$$yhat = b + WX;$$

Dimana yhat: perkiraan nilai sewa kantin

X: vector luas ruangan kantin

W: Parameter weight/vector berupa koefisien

b: bias term.

Supervised-Linear Regression

Performance (P) : kita dapat hitung error untuk tiap observasi i sebagai:

$$e_i = abs(\hat{y}_i - y_i)$$

More popular: **Mean Squared Error (MSE)**

$$MSE = 1/2m \sum_i (\hat{y}_i - y_i)^2$$

Regularization *(remember bias vs variance?)*

Misal sebuah model yang kompleks (model dengan banyak variable) digunakan. **Low bias high Variance.**

Regularization : Reduce model complexity (reduce variance)

- Mengurangi Variance saat kita menggunakan banyak variable (complex model)
- Ridge : mengurangi koefisien/parameter shrinkage hingga variable yang tidak signifikan akan mendekati 0. Jumlah variabel yang digunakan oleh model tetap sama
- Lasso: mengurangi koefisien/parameter shrinkage sampai 0. Jumlah variable yang digunakan dapat berkurang
- Efek: **Higher Bias**

OLS *(remember bias vs variance?)*

- **Ridge** pada umumnya lebih cepat
- **Ridge** digunakan jika sebagian besar variable pengaruh yang sama terhadap response.
- **Lasso** umumnya digunakan saat hanya ada beberapa variabel yang signifikan/berpengaruh terhadap response

Supervised- Logistic regression

Problem : Pegawai DJP diminta untuk memprediksi apakah wajib pajak akan menyampaikan SPT pada akhir tahun

Kemungkinan yang akan terjadi: (1) wajib pajak menyampaikan (2) wajib pajak tidak menyampaikan

Probabilias wajib pajak menyampaikan SPT: 0 =< P =< 1

Supervised- Logistic regression

Pada dasarnya Logistic regression sama seperti linear regression

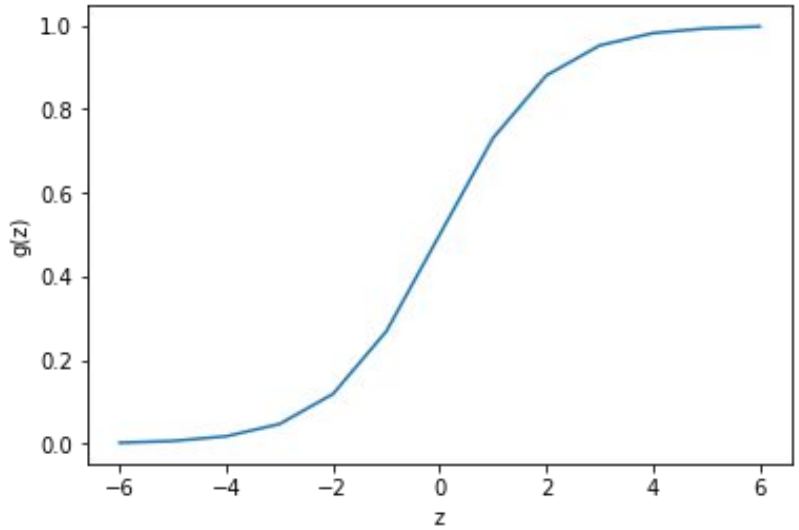
Logistic regression memprediksi probabilitas sebuah observasi berada dalam kategori positif

Target variabel pada Logistic regression adalah kategori (A atau bukan A)

Karena output merupakan probabilitas, gunakan sigmoid function

Supervised- Logistic regression

Sigmoid function: g(z) = 1/(1+e^-z)



Supervised- Logistic regression

Jika sigmoid function dimasukkan ke dalam linear regression:

$$g(x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

Dimana g(X) adalah probabilitas yhat merupakan kategori positive

Supervised- Logistic regression

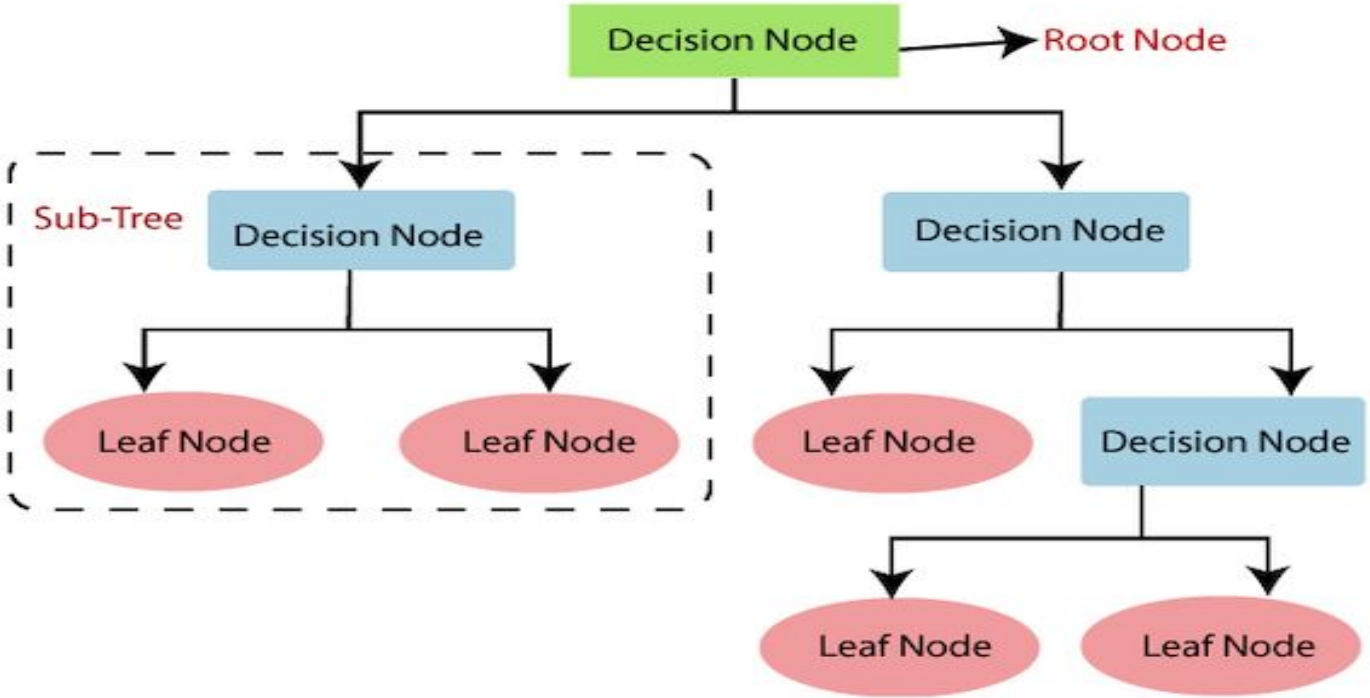
Output logistic regression: angka probabilitas

Performance : Akurasi klasifikasi (yhat) dengan threshold tertentu

Training: mengoptimalkan akurasi

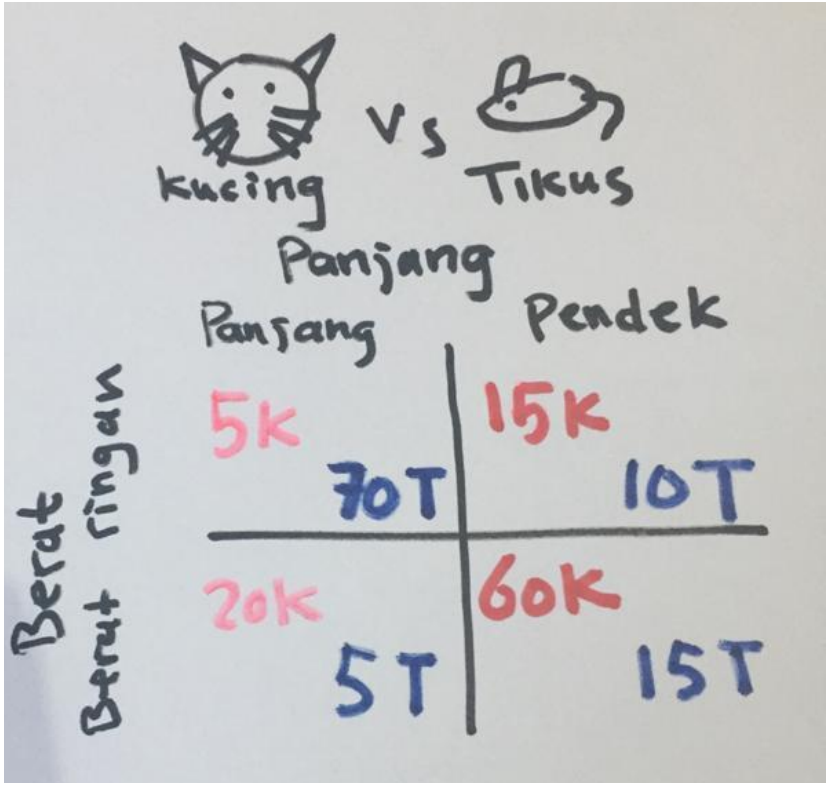

Decision Tree

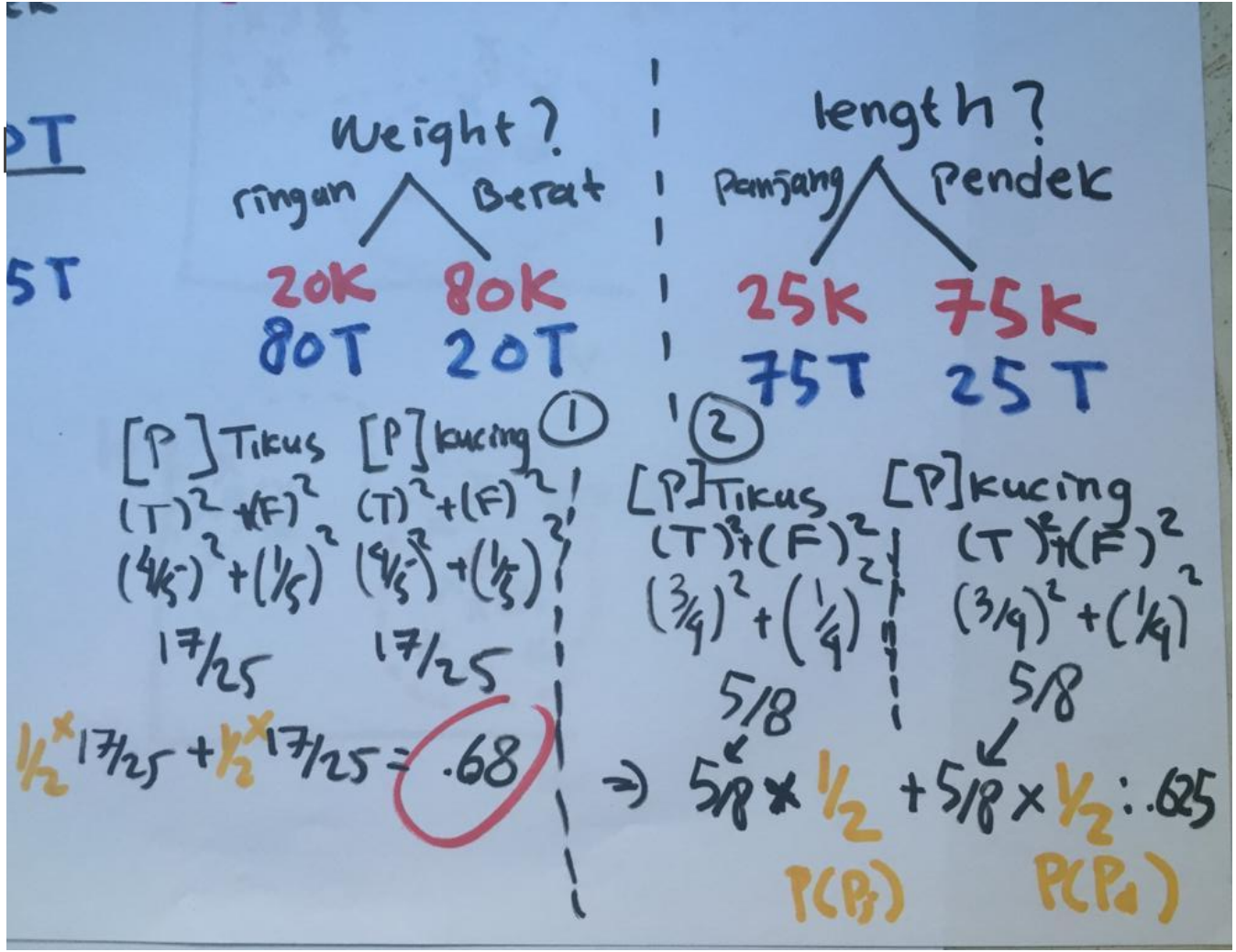
Decision Tree

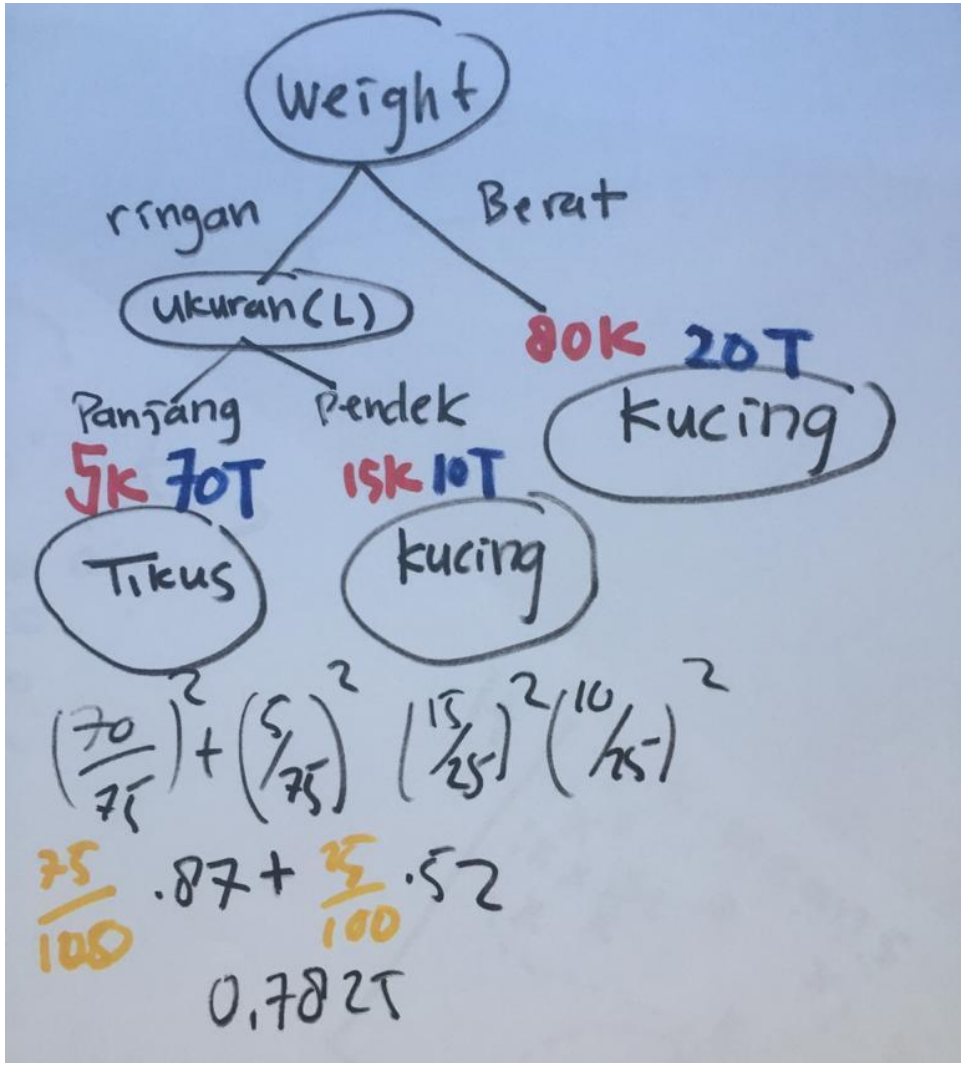


Source: glints.com

Tree (Decision Tree)




Weight ?
ringan
Berat
20K
80K
80T
20T
length ?
Panjang
Pendek
25k
75k
75T
25 T
[P] Tikus
[P] kucing
(T)² + (F)²
(4/5)² + (1/5)²
17/25
17/25
1/2 x 17/25 + 1/2 x 17/25 = .68
1
2
[P] Tikus
[P] kucing
(3/4)² + (1/4)²
5/8
5/8
⇒ 5/8 x 1/2 + 5/8 x 1/2 : .625
P(P)
P(P)


Weight
ringan
Berat
ukuran(L)
80K 20T
kucing
Panjang
Pendek
5K 70T
15K 10T
Tikus
kucing
$\left(\frac{70}{75}\right)^2 + \left(\frac{5}{75}\right)^2 \left(\frac{15}{25}\right)^2 \left(\frac{10}{25}\right)^2$
$\frac{75}{100} .87 + \frac{25}{100} .52$
0,782T

Decision Tree

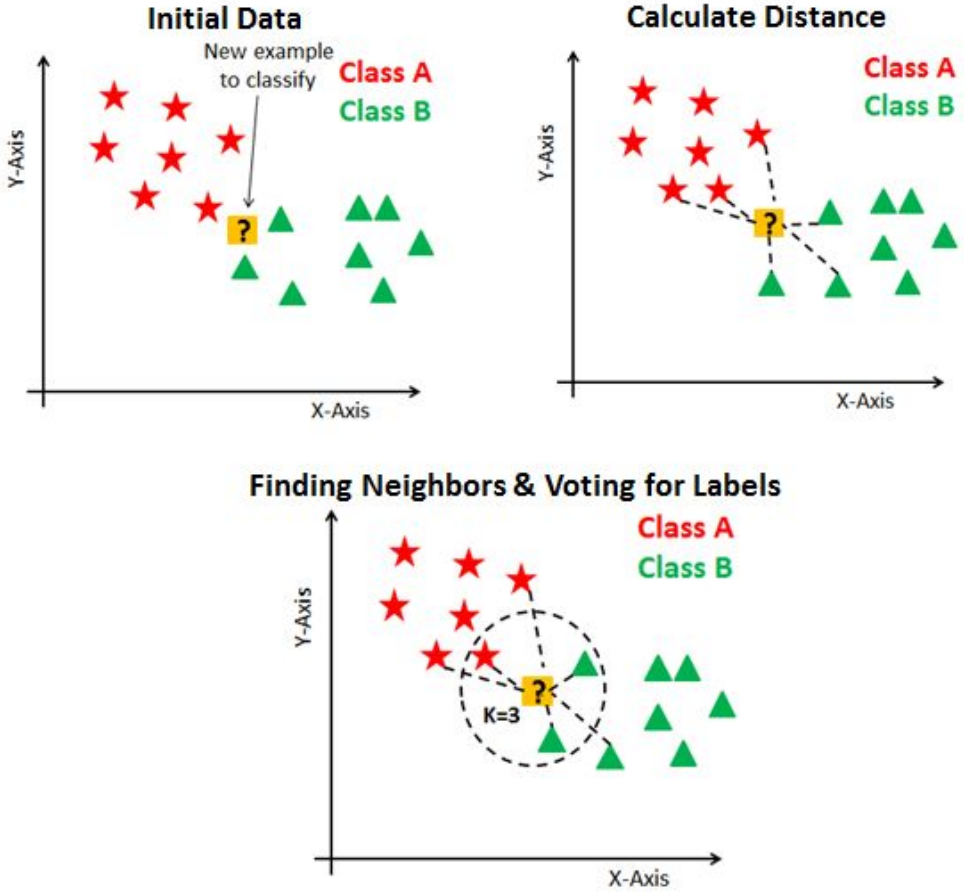
- Seperti namanya, Decision tree berbentuk cabang pohon
- Dapat digunakan baik untuk regression maupun classification problem
- Mampu memvisualisasi model dan keputusan secara jelas/eksplisit sehingga interetabilitasnya sangat tinggi
- Robust, **tidak rentan** terhadap outlier


Tetangga
Sebelah

KNN- K nearest Neighbour

1. Menentukan sejumlah (k) observasi yang terdekat berdasarkan ukuran jarak yang ditetapkan sebelumnya (contoh nya euclidean distance)
2. Berdasarkan kategori “tetangga”, model memprediksi kategori suatu observasi dengan voting.
3. Jika hasil voting imbang maka kategori, dipilih secara random

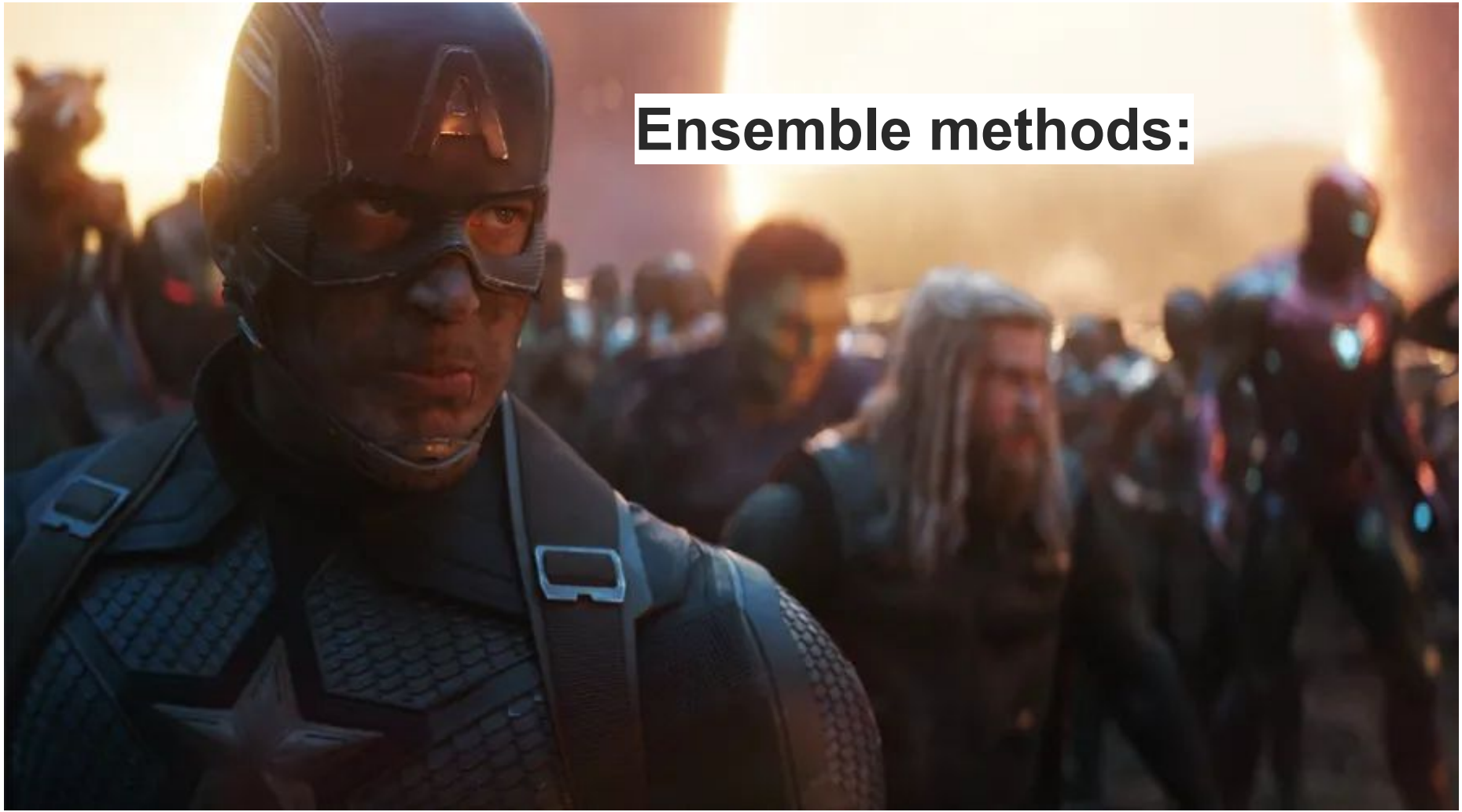
KNN- K nearest Neighbour



KNN

\+ Robust, tidak terlalu terpengaruh oleh outliers

\- Rentan terhadap satu variabel yang skalanya lebih besar daripada variabel lain

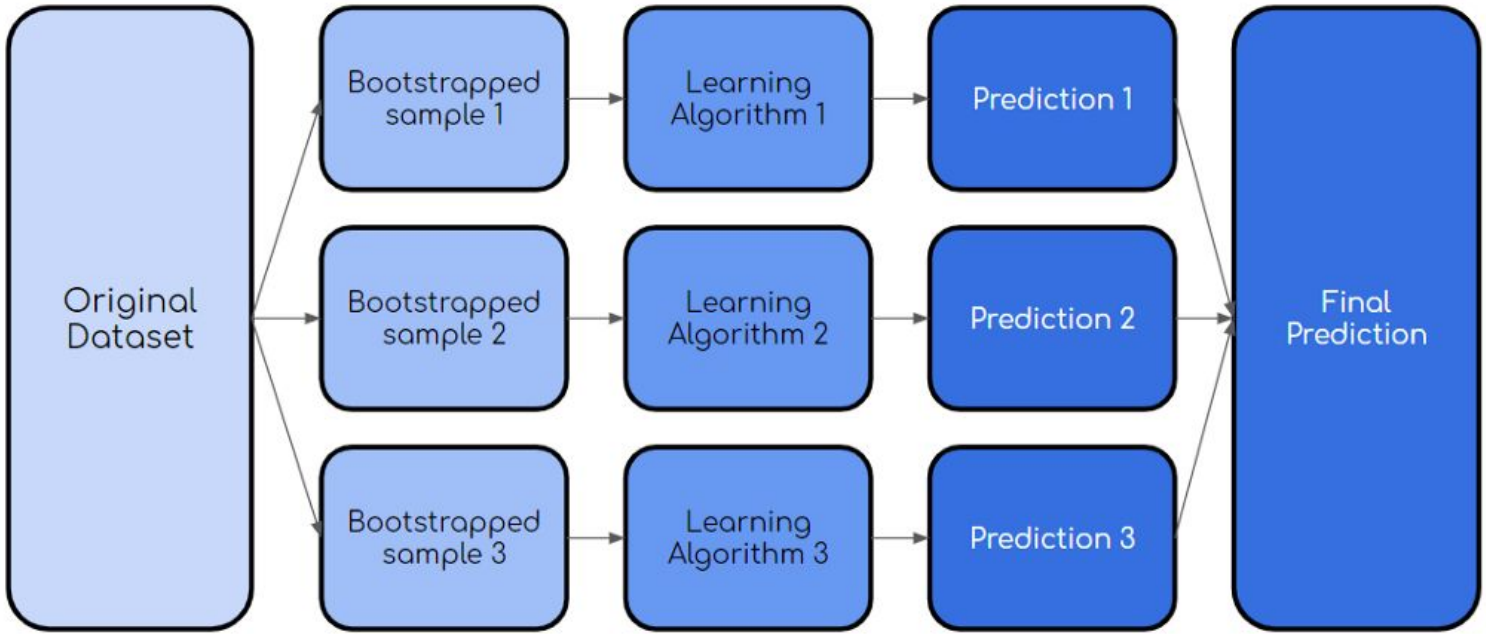

Ensemble methods:

Ensemble methods: *bagging, boosting* and *stacking*

***----- weak learners**: Model-model yang memiliki bias atau variance yang tinggi namun **dapat** digunakan bersama untuk membentuk model yang lebih baik-----*

- **Bagging** to decrease the model's **variance**;
- **Boosting** to decrease the model's **bias**;
- **Stacking** to increase the predictive force of the classifier.

Bagging-mengumpulkan banyak weak learners (sejenis) kemudian di aggregate

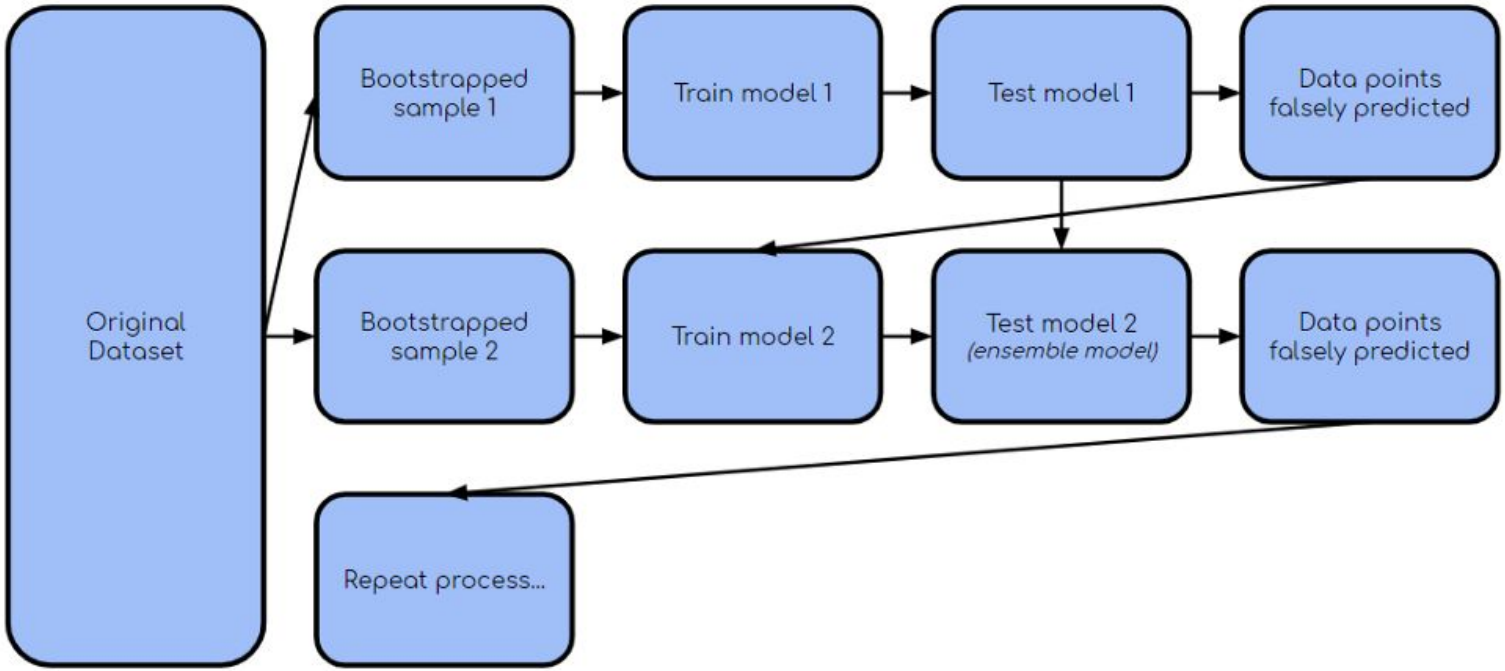


Source picture: Terence Shin (towarddatascience.com)

Example

- Bagged Decision Tree
- Random Forest

Boosting - mengembangkan weak learner secara sequential dan adaptive dengan weak learner lain (sejenis).

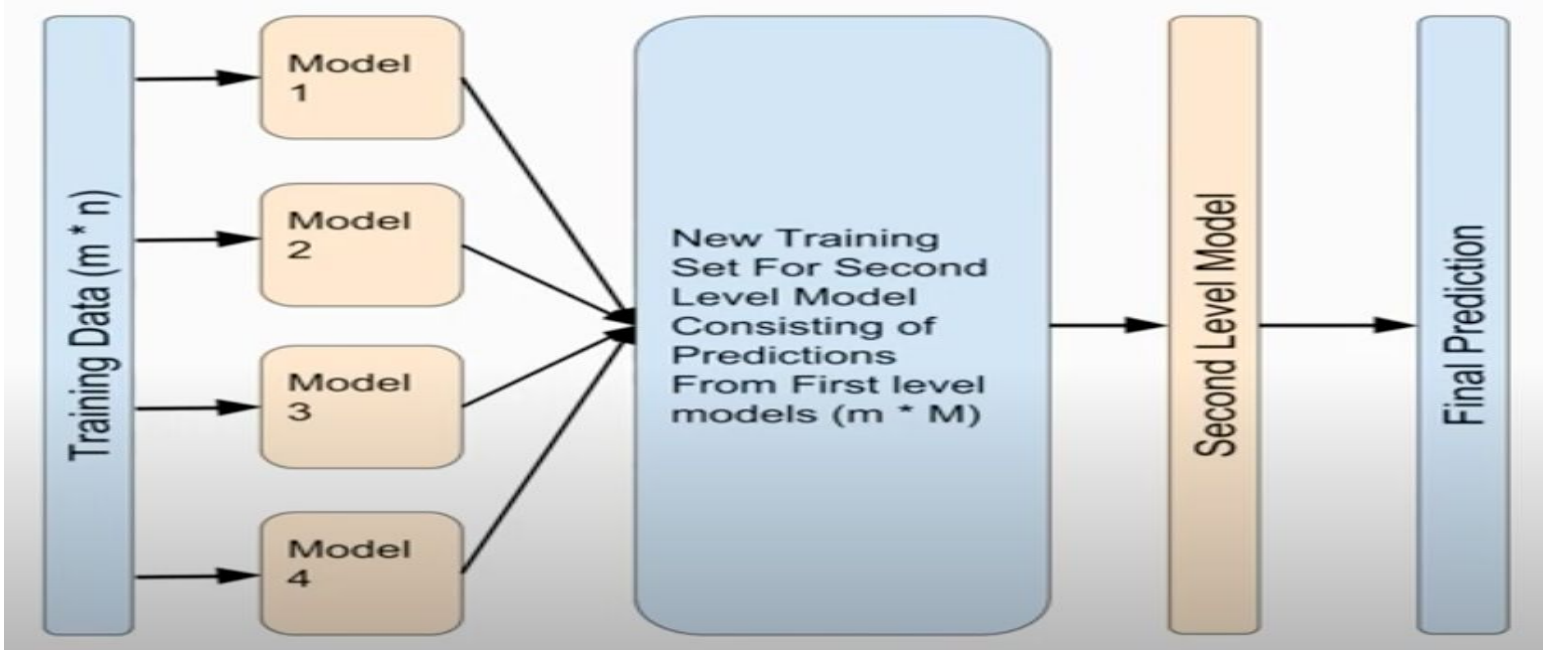


Source picture: Terence Shin (towarddatascience.com)

Example

- Adaptive Boosting
- Gradient Boosting

Stacking - menggabungkan weak learners dengan weak learner yang berbeda algoritma (heterogeneous)



Source picture: Terence Shin (towarddatascience.com)

Example

- Stacked Model