MoF-DAC
Ministry of Finance
Data Analytics Community

Kemenkeu
Corporate University

MACHINE LEARNING

Disusun oleh :

Sindhu Wardhana

Ade Satya Wahan

Aris Budi Santoso

Leonard Yulianus

BASIC MACHINE LEARNING

Setelah mengikuti program pembelajaran, peserta diharapkan dapat:

Standar Kompetensi:

Menerapkan metode dan teknik machine learning tingkat dasar, evaluasi kualitas, dan validasi keakuratan model machine learning.

Kompetensi Dasar:

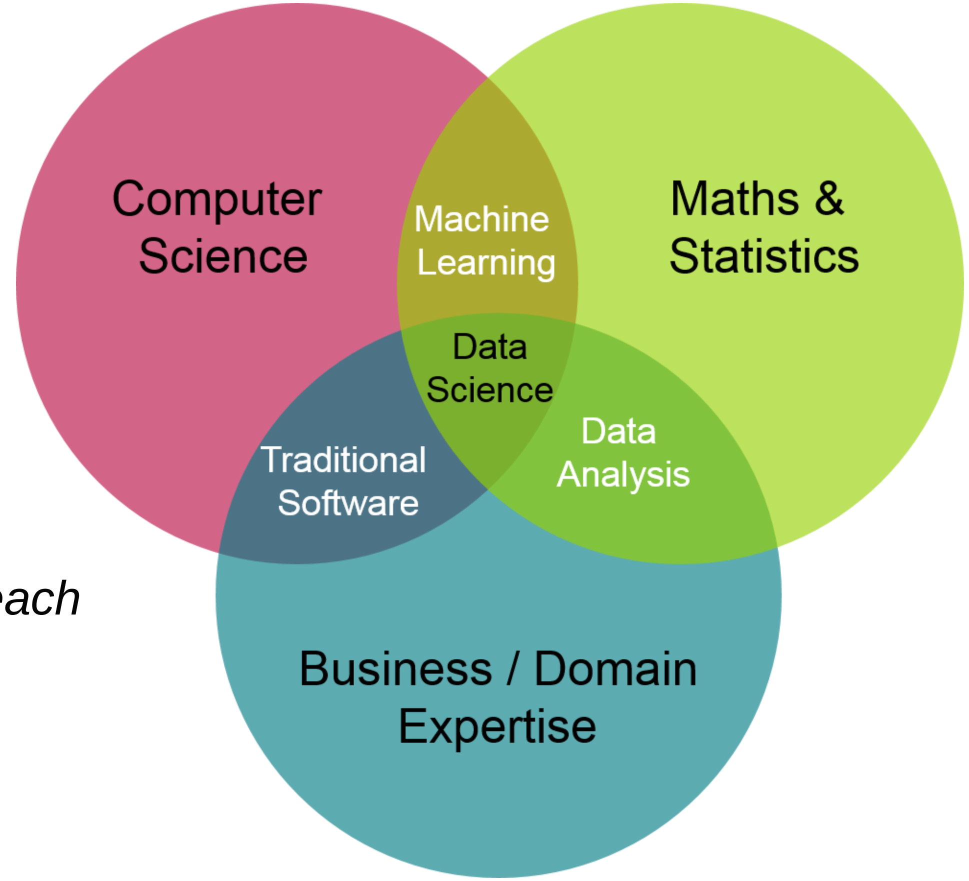
1. Menjelaskan konsep dasar machine learning;
2. Menerapkan pendekatan supervised learning algorithms;
3. Menerapkan unsupervised learning algorithms;
4. Menerapkan evaluasi/ pengukuran kinerja model yang telah disusun; dan
5. Menerapkan optimisasi kinerja model.

WHAT IS MACHINE LEARNING?

"Ability to learn without being explicitly programmed"
--- Arthur Samuel, 1959

"*Learn from **experience** (E) with respect to some **task** (T) and some **performance** measure (P)*"
--- Tom Mitchell, 1997

Machine learning is a field of computer science that aims to teach computers how to learn and act without being explicitly programmed
--- https://deepai.org/machine-learning-glossary-and-terms/machine-learning



KEY POINTS OF MACHINE LEARNING

TASK (T)

EXPERIENCE (E)

PERFORMANCE (P)

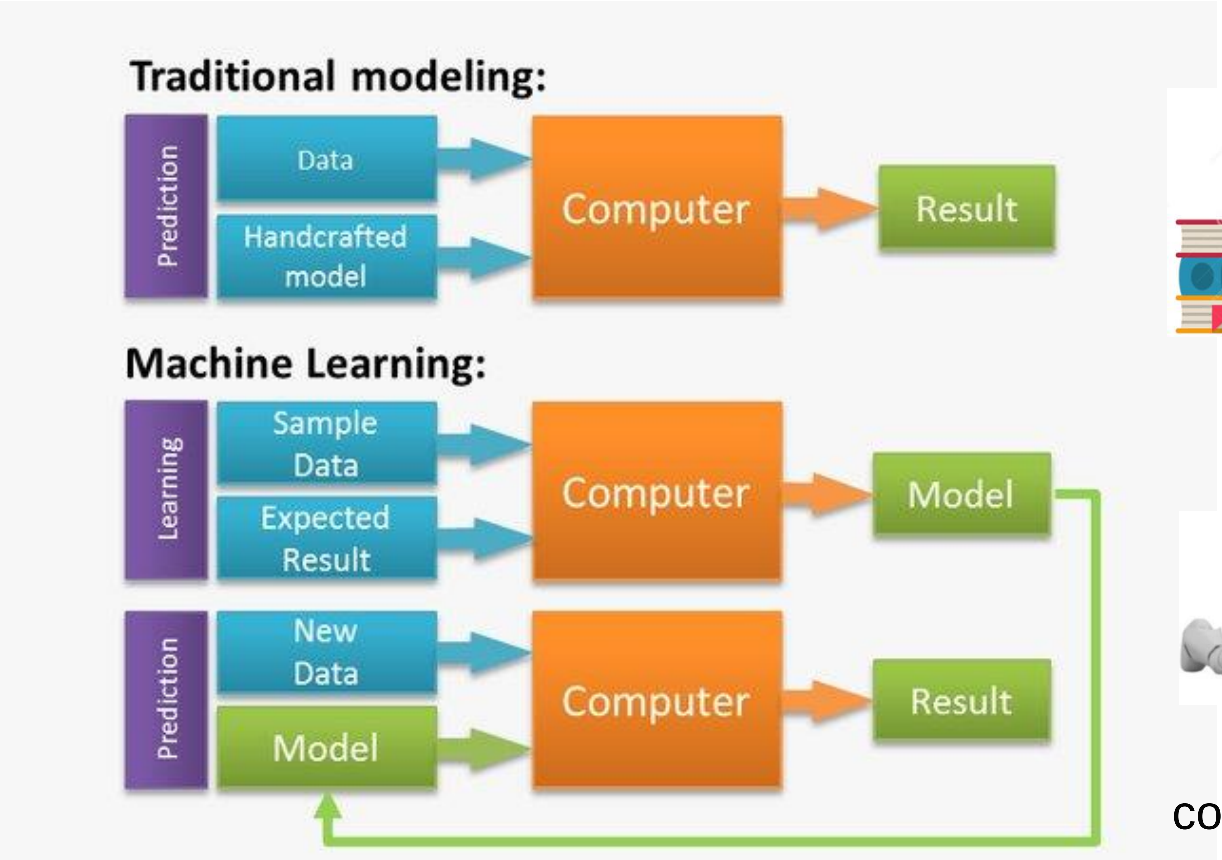
Machine learning untuk memprediksi cuaca

Prediksi cuaca

data riwayat indikator kecepatan angin, kelembaban udara, suhu, pembentukan awan, tingkat curah hujan pada lokasi tertentu

persentase kondisi cuaca yang diprediksi dengan tepat (akurasi)

TRADITIONAL PROGRAMMING VS MACHINE LEARNING



Mehra, Sidharth & Hasanuzzaman, Mohammed. (2020). Detection of Offensive Language in Social Media Posts

Orang menulis rule dalam bentuk kode aplikasi

Model (komputer) dilatih menggunakan data

contoh ril sederhana : [klik di sini]

pusdiklat keuangan umum.

BUT WHY MACHINE LEARNING?

No Human Experience Yet

Can't explain the experience

Many solutions adaptation

Situation changes

Large amount of Data

Human are too expensive

You wouldn't want be this guy

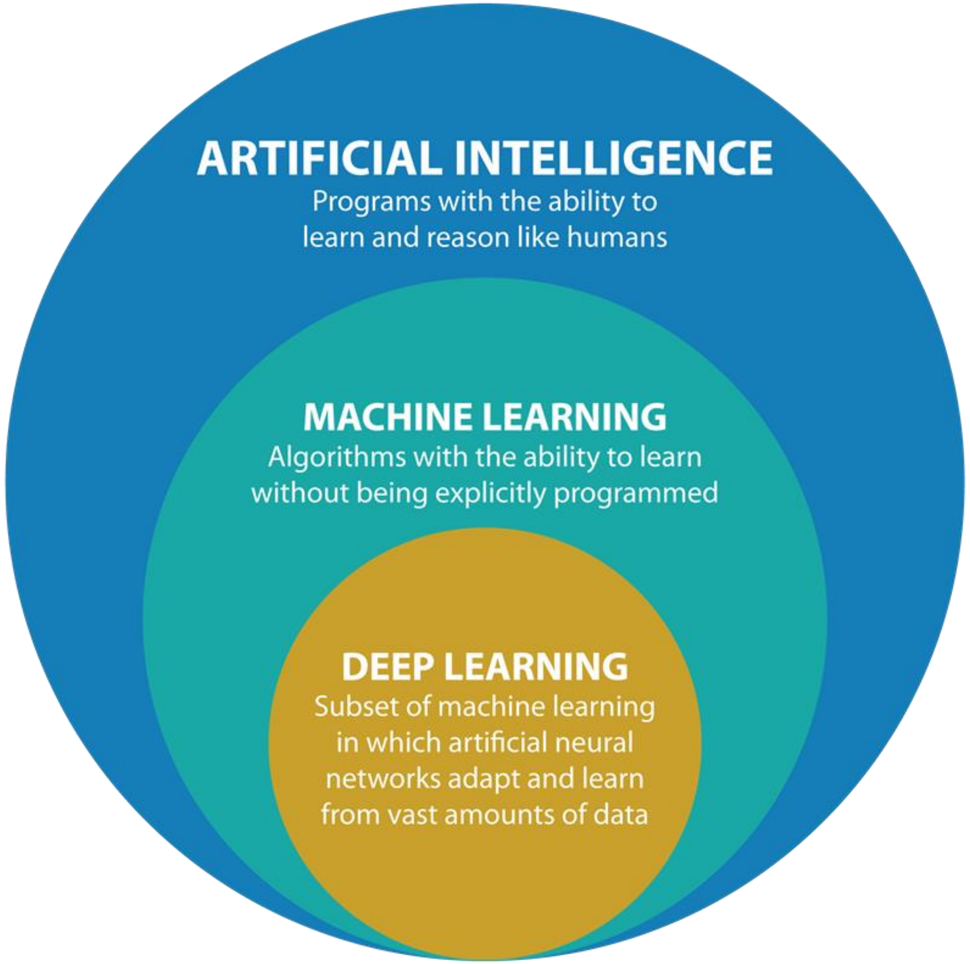
Checking all data by eyes and hands

EXERCISE – T / E / P / NONE

- P — Jumlah makanan yang dengan benar diklasifikasikan sebagai seafood
- N — Mengubah daftar menu menjadi matrix/angka
- T — Mengklasifikasikan label makanan sebagai seafood atau bukan seafood
- N — Download daftar makanan dari internet
- E — Dataset berisi makanan yang telah dilabeli seafood dan bukan seafood

Aplikasi Machine Learning di restoran seafood

JARGONS ??


ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

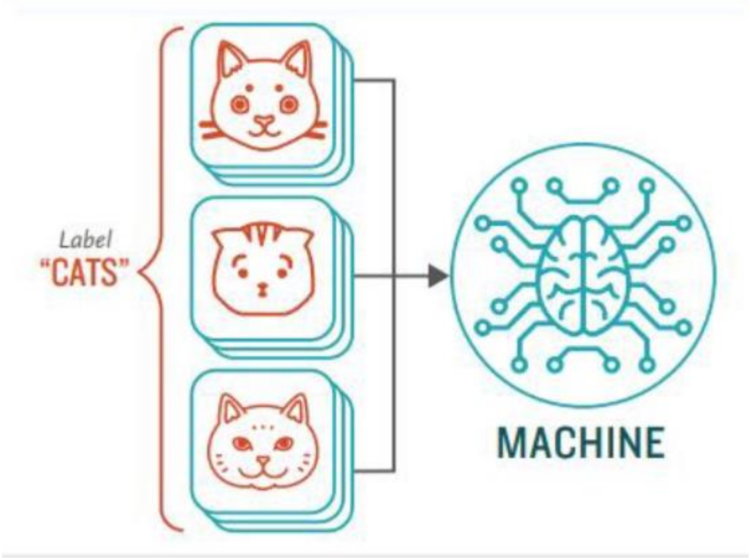
MACHINE LEARNING TYPES

Supervised	• Menggunakan dataset **memiliki label** (E) untuk memprediksi varible target (T)
Unsupervised	• Menggunakan dataset **tanpa label** (E) untuk melihat/mempelajari pola (T)
Semi-supervised	• Menggunakan data **dg label** dan **tanpa label** (E) untuk memprediksi / mempelajari pola (T)
Reinforced Learning	• Menggunakan data hasil simulasi secara iterative (E) untuk mencapai tujuan (T) (memperbesar **reward** / mengurangi error)

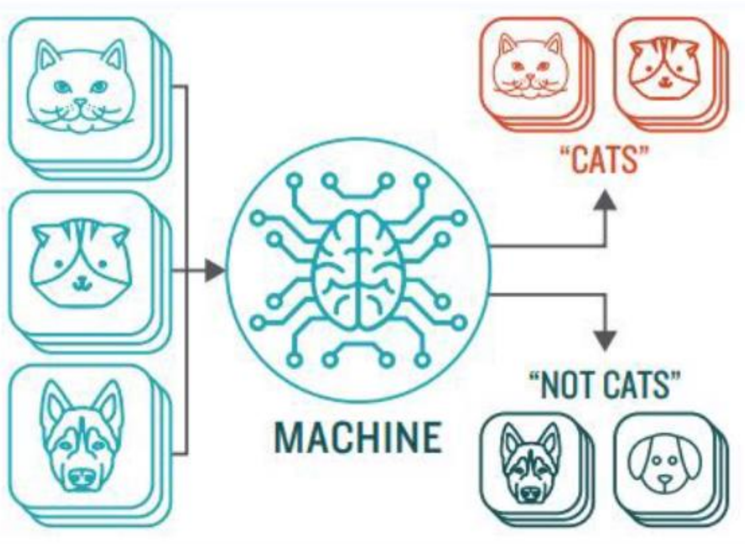
SUPERVISED LEARNING

Supervised Learning

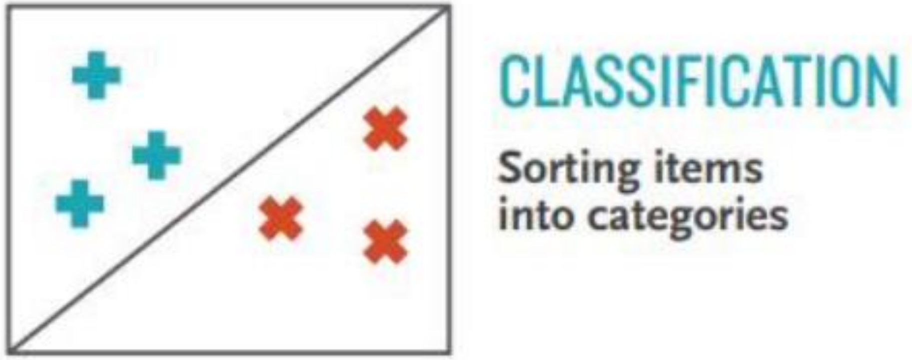
STEP 1: Training



STEP 2: Predicting

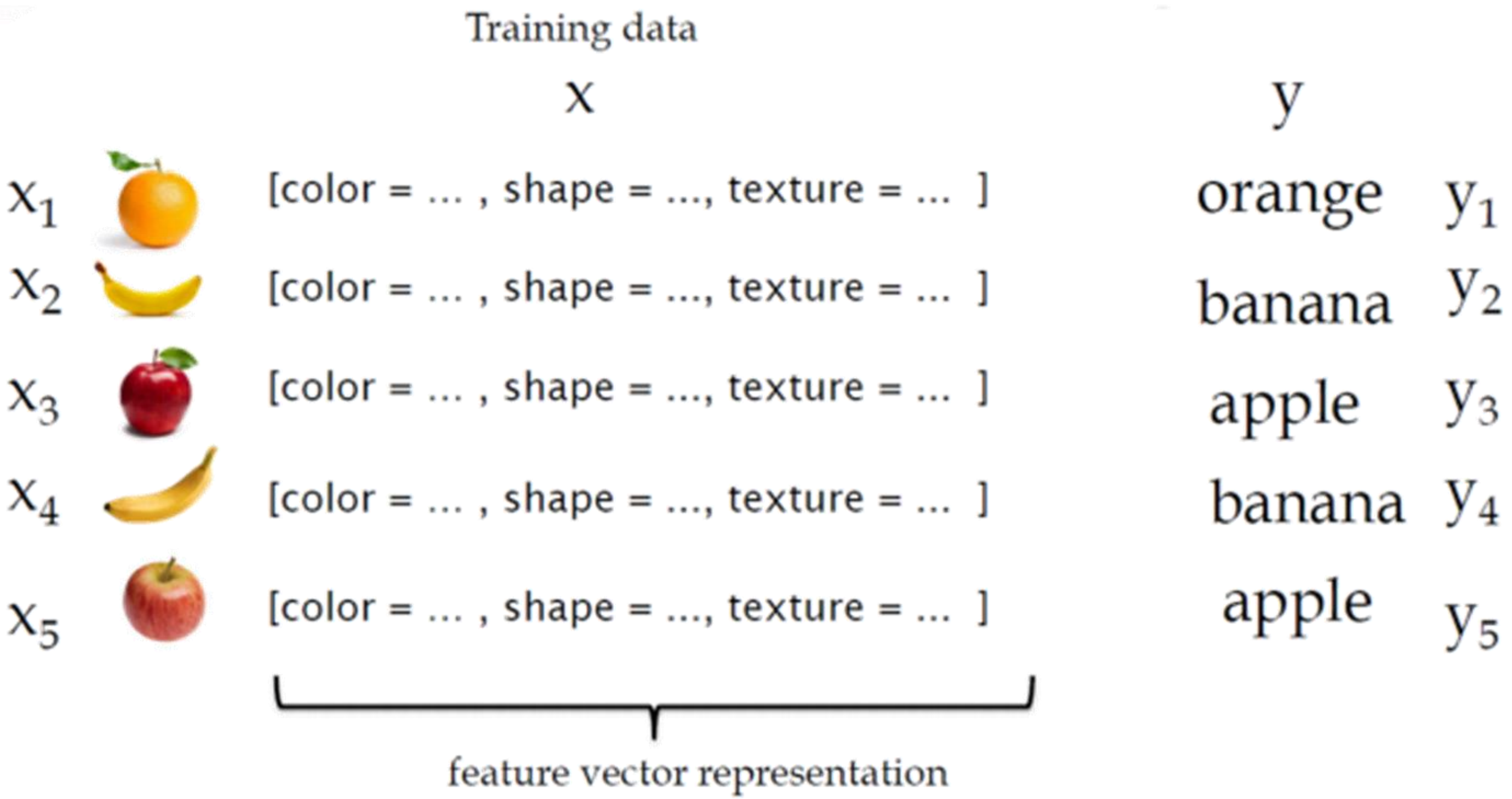


Different Types Based on Target Variable

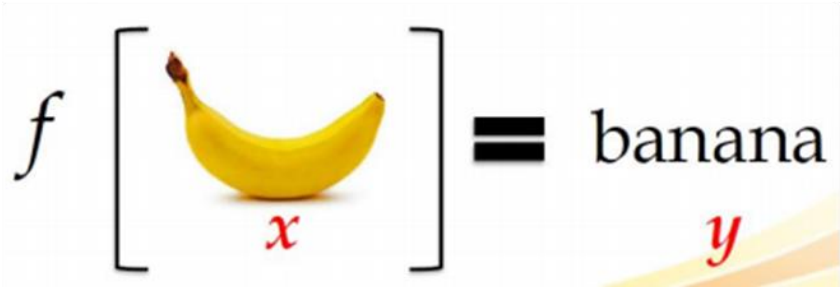


Let's go to math…

Simple Math Notation on Training Step



finding best f(x)
to predict new data



Some Supervised Learning Models

Linear Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Training = Find the optimal **β**

Related models

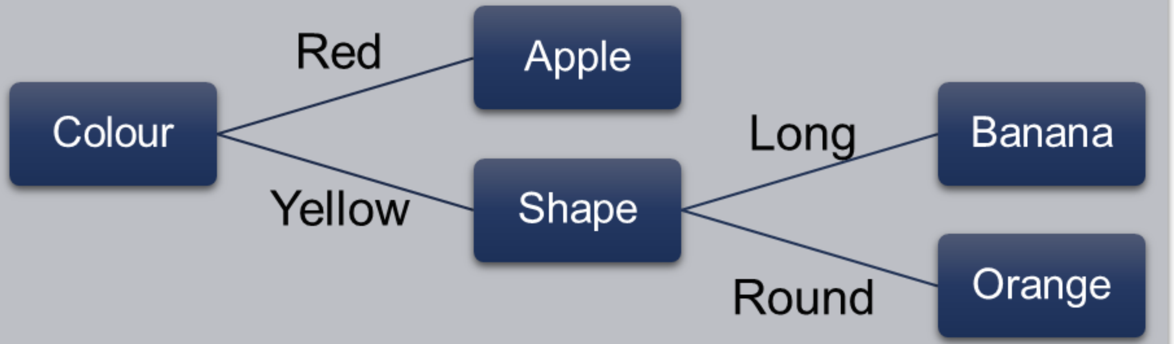
Logistic
Add sigmoid function

Lasso / Ridge
Add regularization term

Polynomial
Add polynomial transformation

Deep Neural Network
Stacking multiple linear model with non linear activation function

Tree Based



Training = Find the optimal **split**

Related models

Decision Tree
Create one tree

Random Forest
Create multiple tree

Ada / Gradient Boost
Create multiple tree sequentially based on info of previous tree

Let’s Coba

Buka notebook di google colab

EVALUATION

Model Evaluations

But, how good is it’s performance?

Ok, so I have a machine learning model now

Is it the best model?

Step to effectively evaluate

- Choose the correct performance metric
- Pick a preferred evaluation approach/method
- Analyse the result

Classification

- Confussion Matrix
  - Accuracy
  - Precision
  - Recall
  - F1-Score
- Area Under the Curve (AUC)

Regression

- Root Mean Squared Error (RMSE) / MSE
- Mean Squared Error (MAE)
- Other:
  - MAPE
  - Adj R^2 / R^2

Classification Model Evaluations

Classification Cases
Confusion Matrix

Binary example
(one class set as positive / target)

		Actual Values (Correct answers)	
		Positive	Negative
Predicted Values (from Model)	Positive	True Positive (TP)	False Positive (FP) Type I error
	Negative	False Negative (FN) Type II error	True Negative (TN)

Accuracy:
percentage of test data that are correctly classified
Accuracy = (TP + TN)/All

Error rate: 1 – accuracy, or
Error rate = (FP + FN)/All

Classification Model Evaluations

Accuracy will have an issue when is used on imbalance target variable

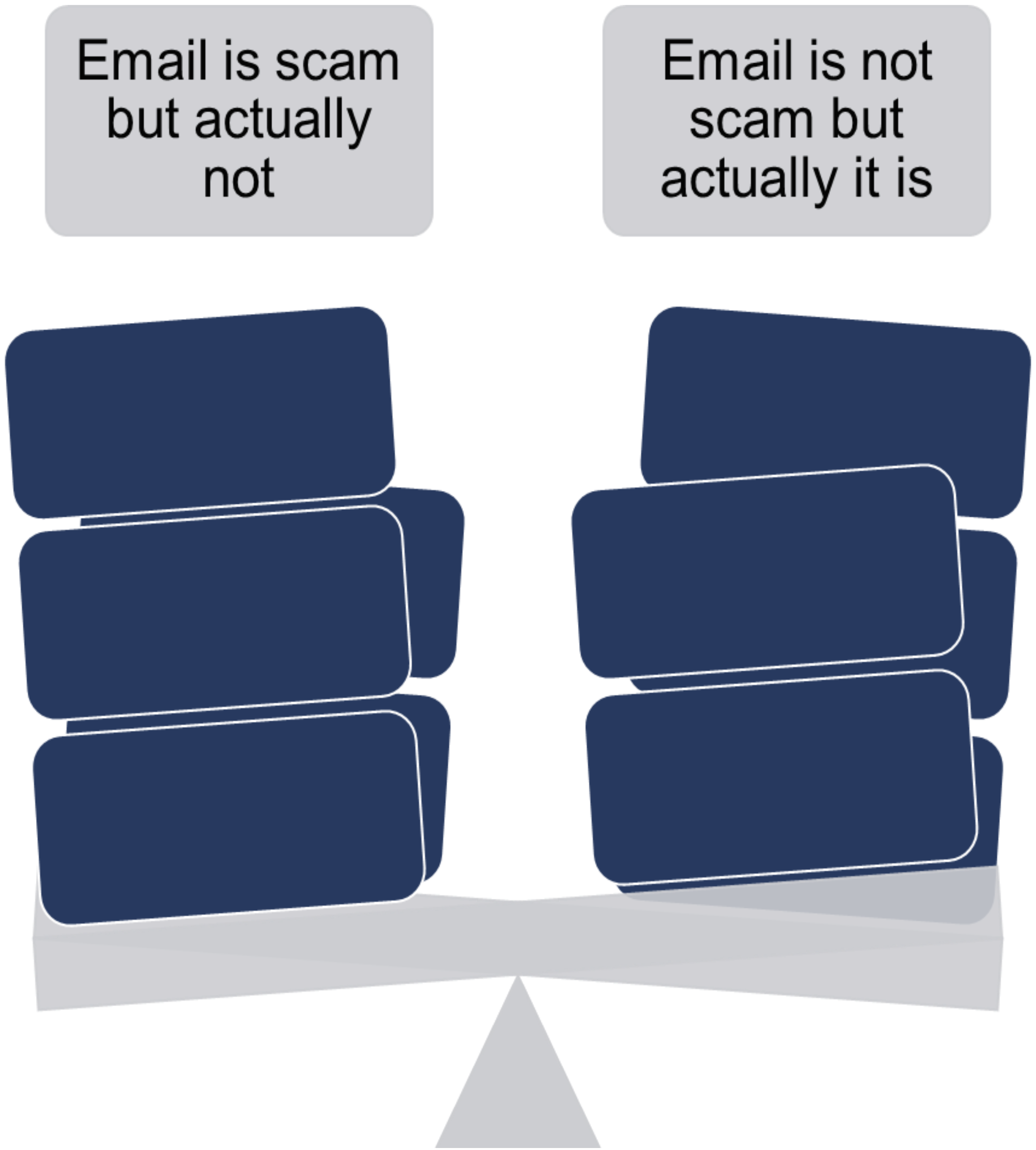
Imbalance = one class may be rare, e.g., fraud, or Scam

So, we need to consider the prediction false cost and use other metric

Let’s discuss:

in a case of predicting scam,

Which false is more costly?



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

- Precision = when the costs of false positives are high
- Recall = when the cost of false negatives is high

F1 / F-score is an overall measure of a model's accuracy that combines precision and recall

Mean Squared Error (MSE)

- Error (true – prediction), squared, get average, rooted if RMSE

MAE

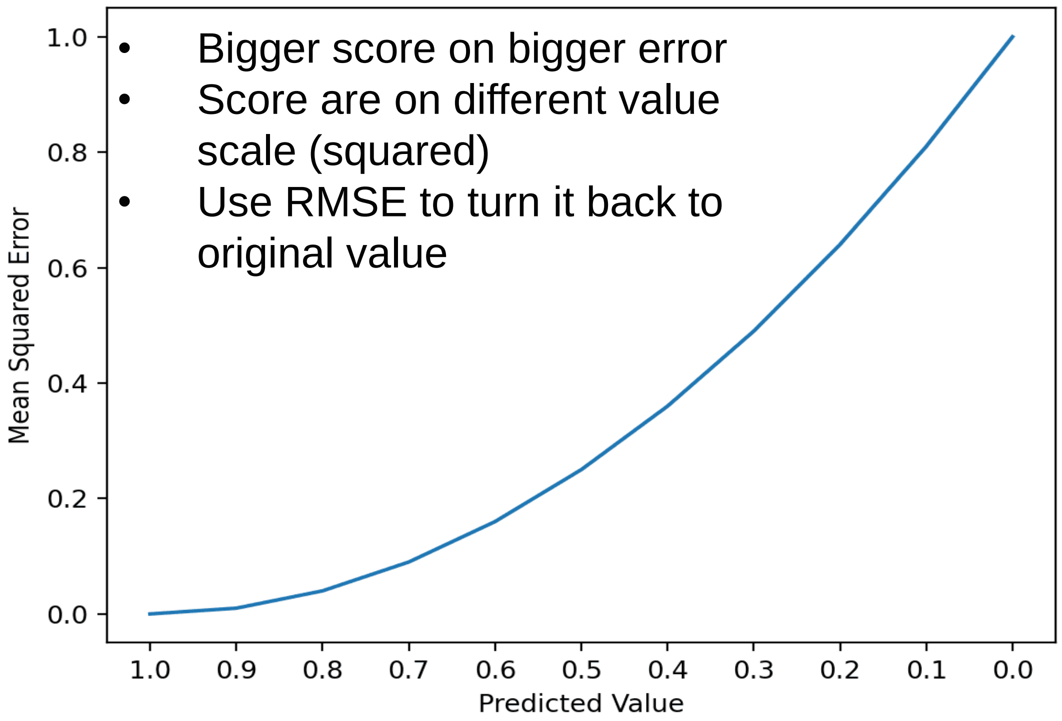
- Error (true – prediction), turn to positive value (absolute), get average

Caveat

- Value can be from 0 to ∞
- Minimized is better
- Minimal means predictions are near true values

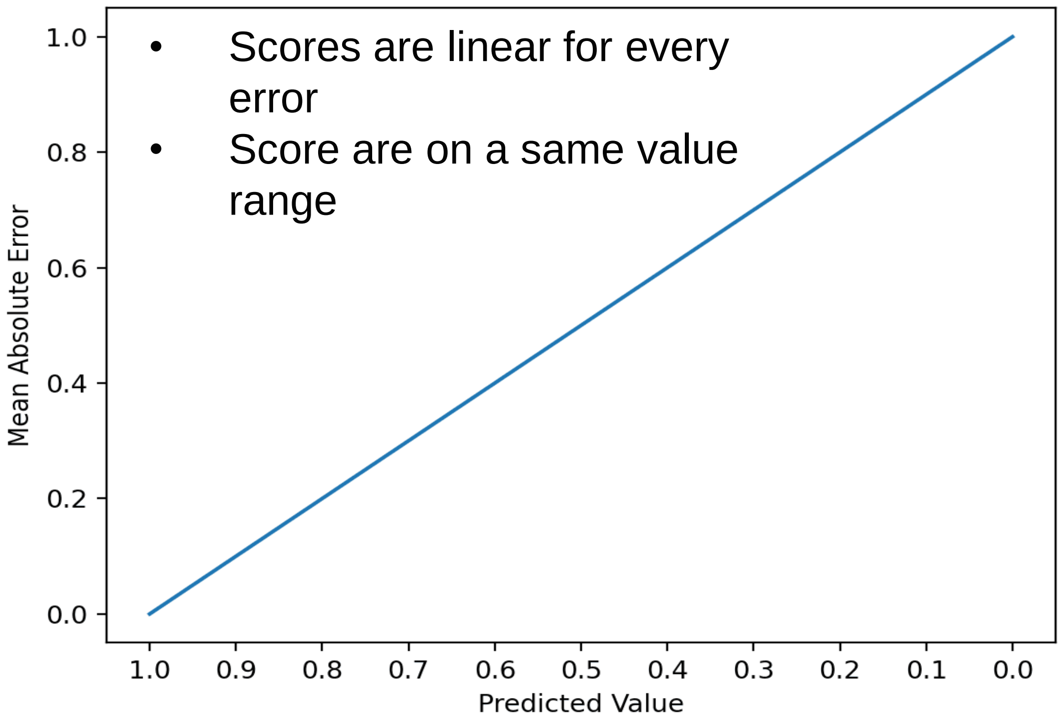
Regression Model Evaluations

Mean Squared Error (MSE)



- Bigger score on bigger error
- Score are on different value scale (squared)
- Use RMSE to turn it back to original value

Mean Absolute Error (MAE)



- Scores are linear for every error
- Score are on a same value range

Both don't show indication on how good is the model
But they are useful to compare model
The best practice is to make a benchmark score

Other metric:
- R^2 / Adj R^2
- MAPE

Let’s Coba

Buka notebook di google colab

Model Evaluations

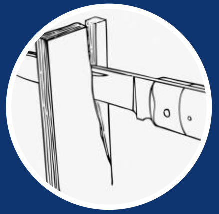
Ok, I’ve picked the best machine learning model

But, will it perform as good as it’s training performance

Evaluation Method

We need Evaluation Method

- Measure the model performance when used on unseen data
- Differentiate data for train and evaluate models

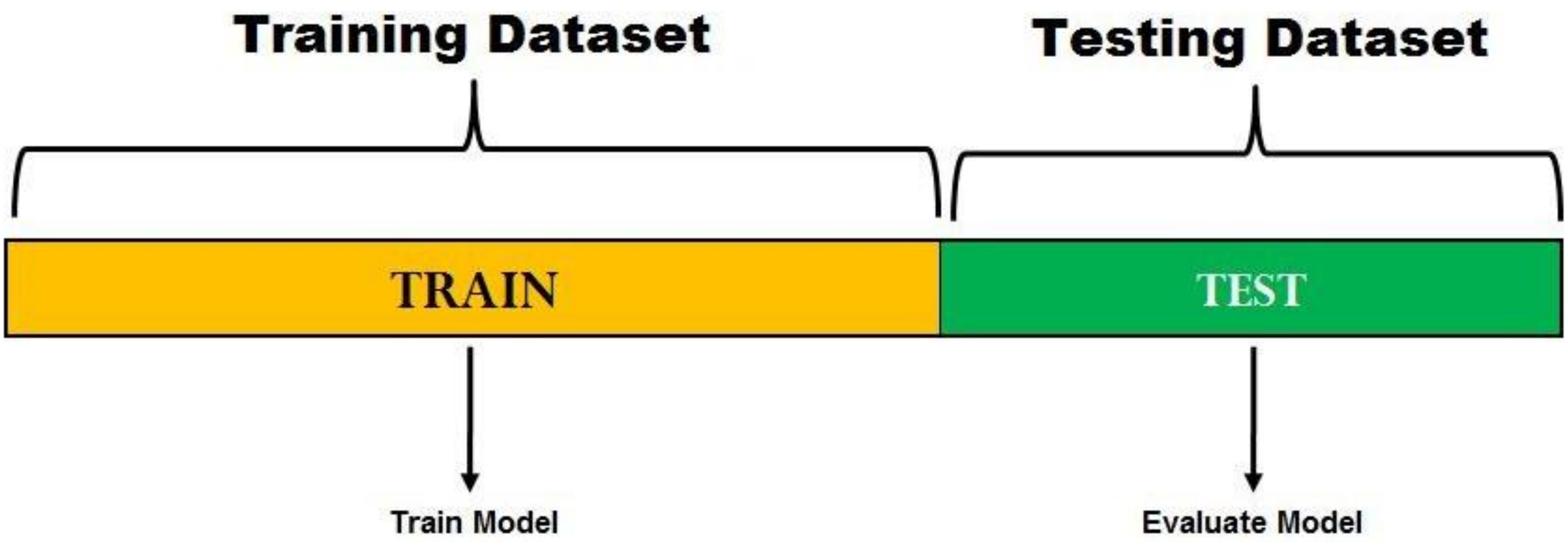


Hold Out

Bootstrap CV

K - Fold CV

Hold Out Method (Splitting)

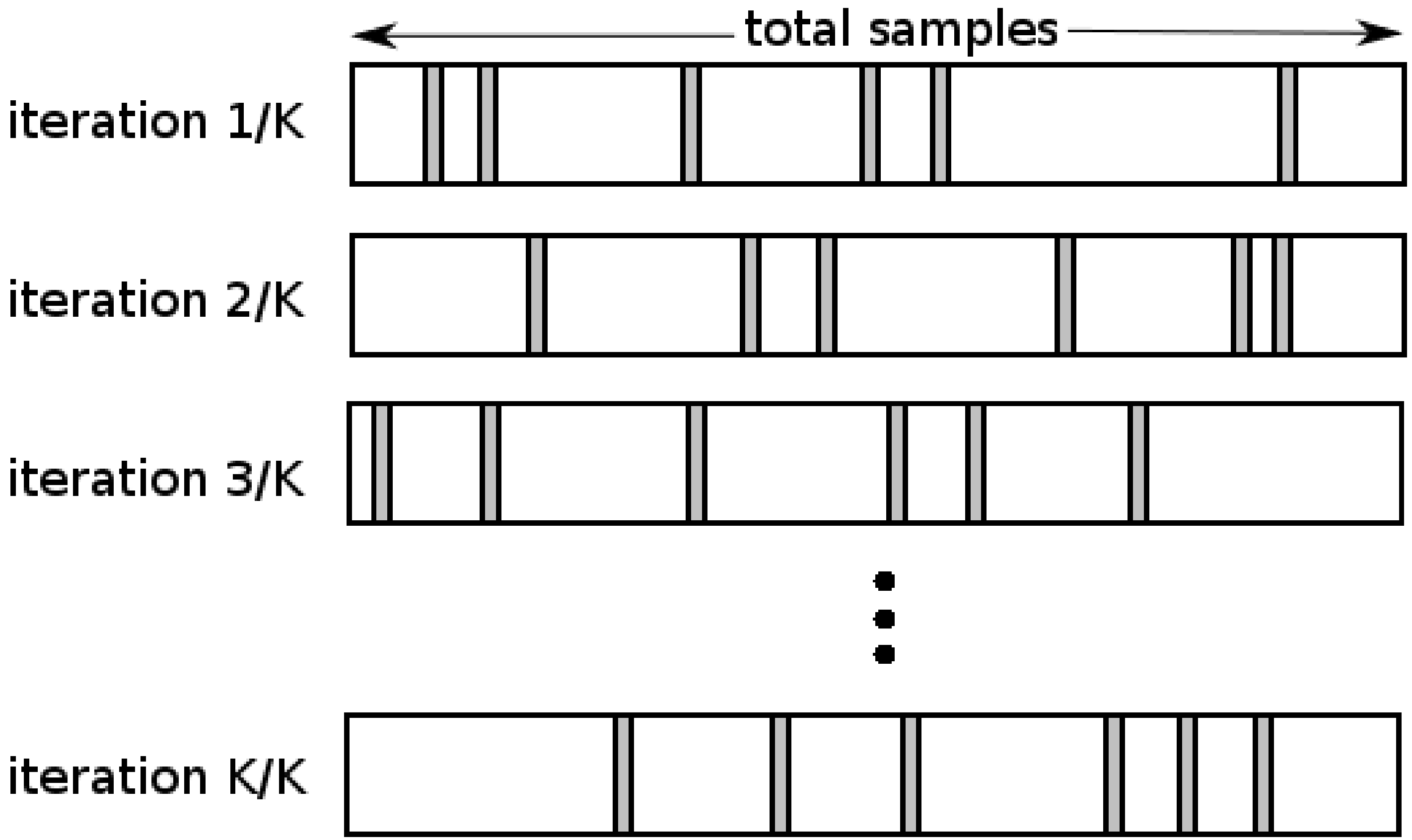


- No golden rules for splitting ratio (75:25, 80:20, 90:10)
- Important to make sure test data represents unseen new data
- Good approach if we have limited data
- Only gives one performance score

RapidMiner:
Splitting Validation Widget

Bootstrapping Cross Validation

Use sampling on creating Training and Testing data (random / stratified)



Repeating K times and final score is the average of all performance score

Out-of-Bag problem

RapidMiner:
Bootstrap Validation Widget

K-Fold Cross Validation

Eksperimen	Dataset	Akurasi
1		93%
2		91%
3		90%
4		93%
5		93%
6		91%
7		94%
8		93%
9		91%
10		90%
	Akurasi Rata-Rata	92%

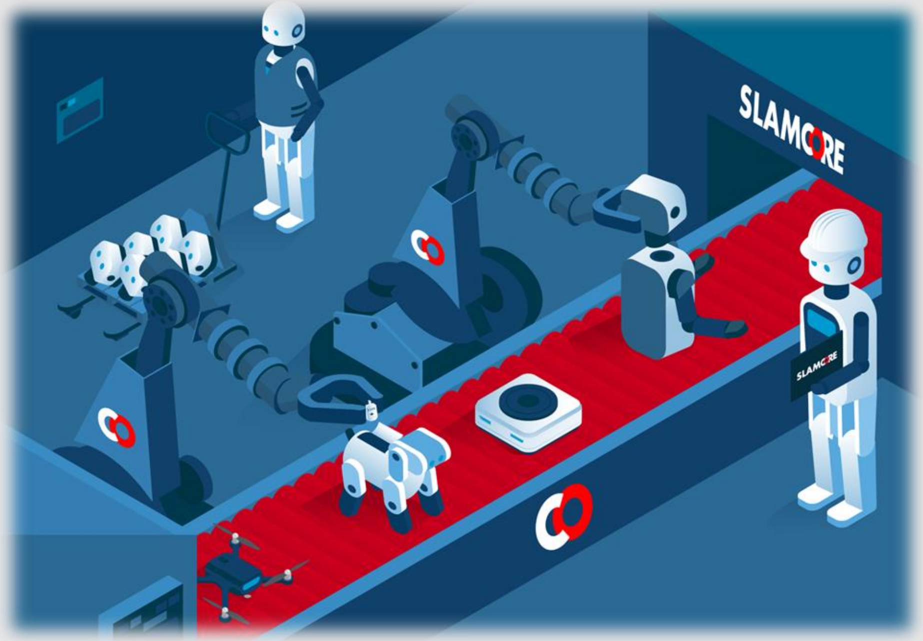
Need more time to finish training process

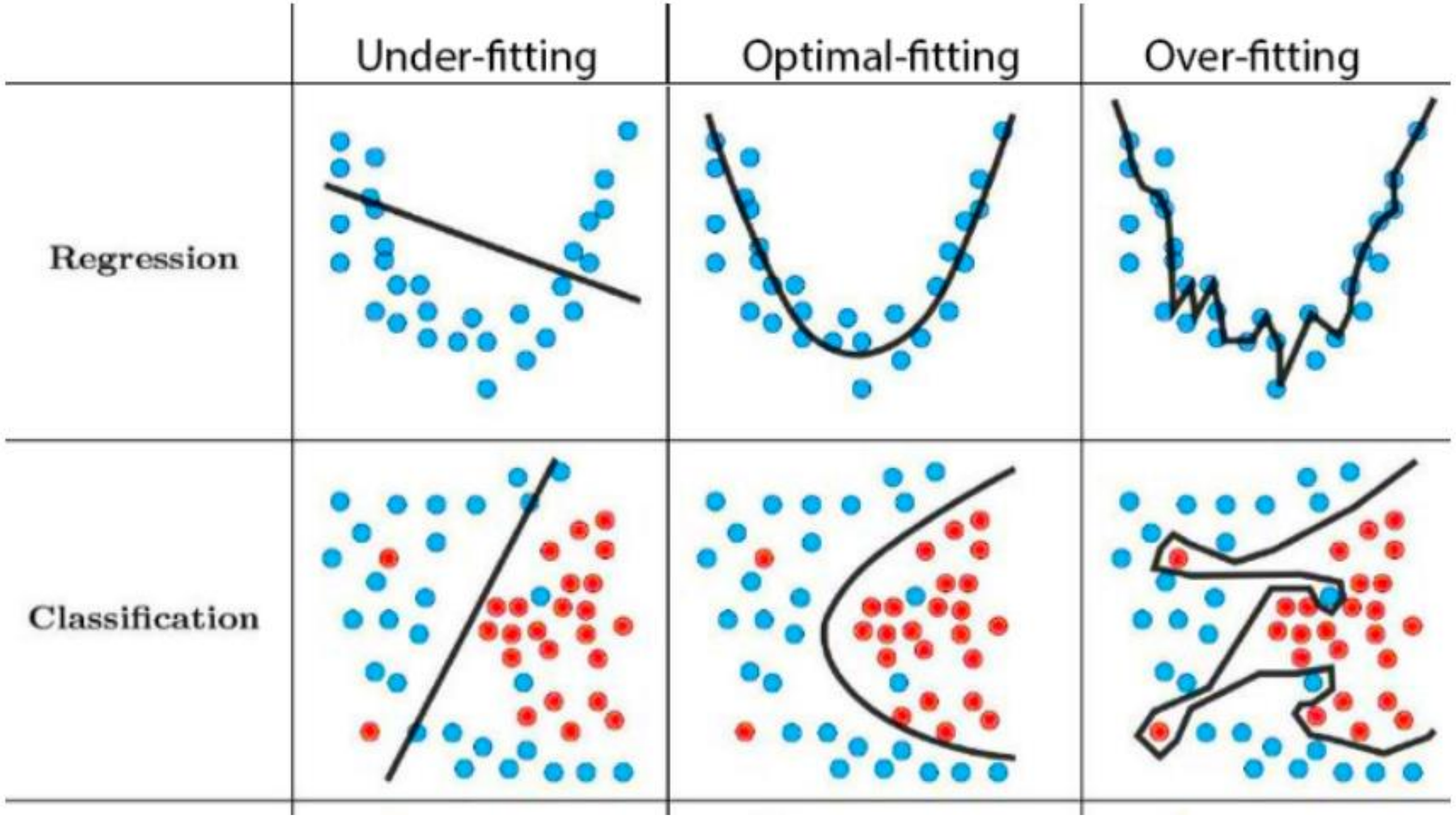
RapidMiner:
Cross Validation Widget

Let’s Coba

Buka notebook di google colab

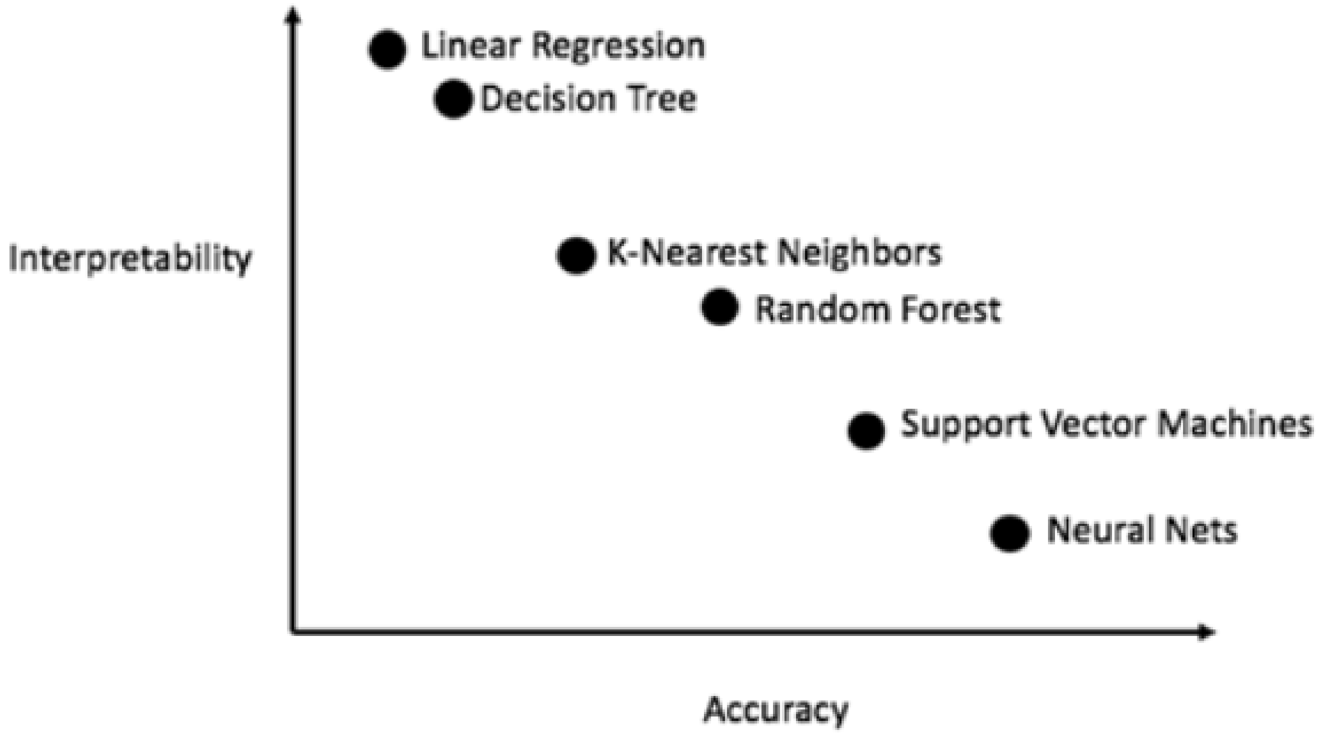
PERFORMANCE TUNING





Maximizing model's performance but with an acceptable generalization level

Model Complexity

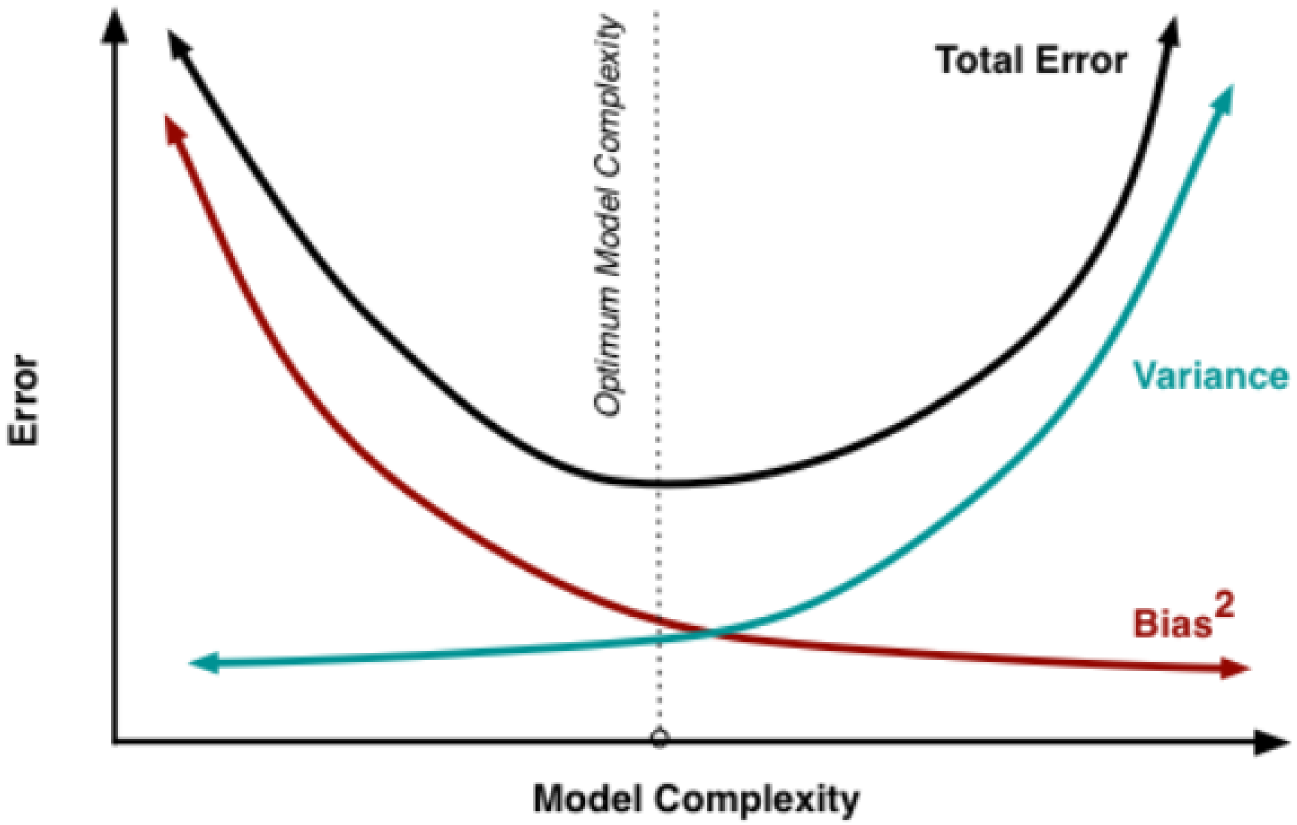


Accuracy vs Interpretability

Model complexity can be defined by:

1. How many parameters learned by the model
2. How difficult for human to explain the model
3. How well the model learned training data

Bias Variance Trade-off



Bias is the difference between the average predicted results of our model and the actual value.

Variance is the variability of our model's prediction of the data points that show the distribution of the data.

How to Tune

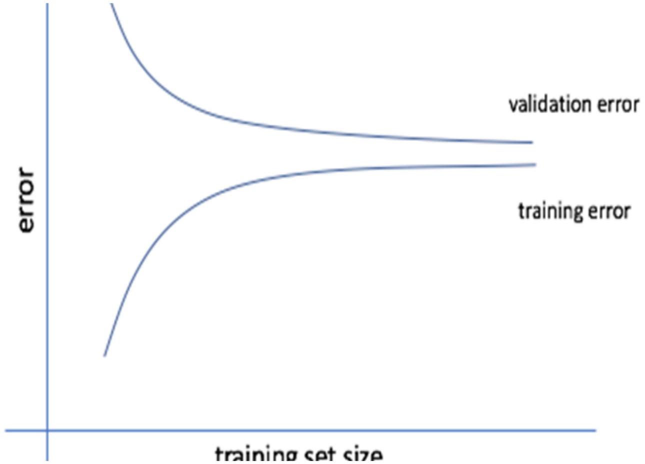
Change ML Algo

Use more/less complex model

Tuning Hyperparameter

GridSearchCV

RandomizedSearchCV



Feature Engineering

Add new rows (introduce more data to lower variance)

Add new / reduce columns (change complexity)

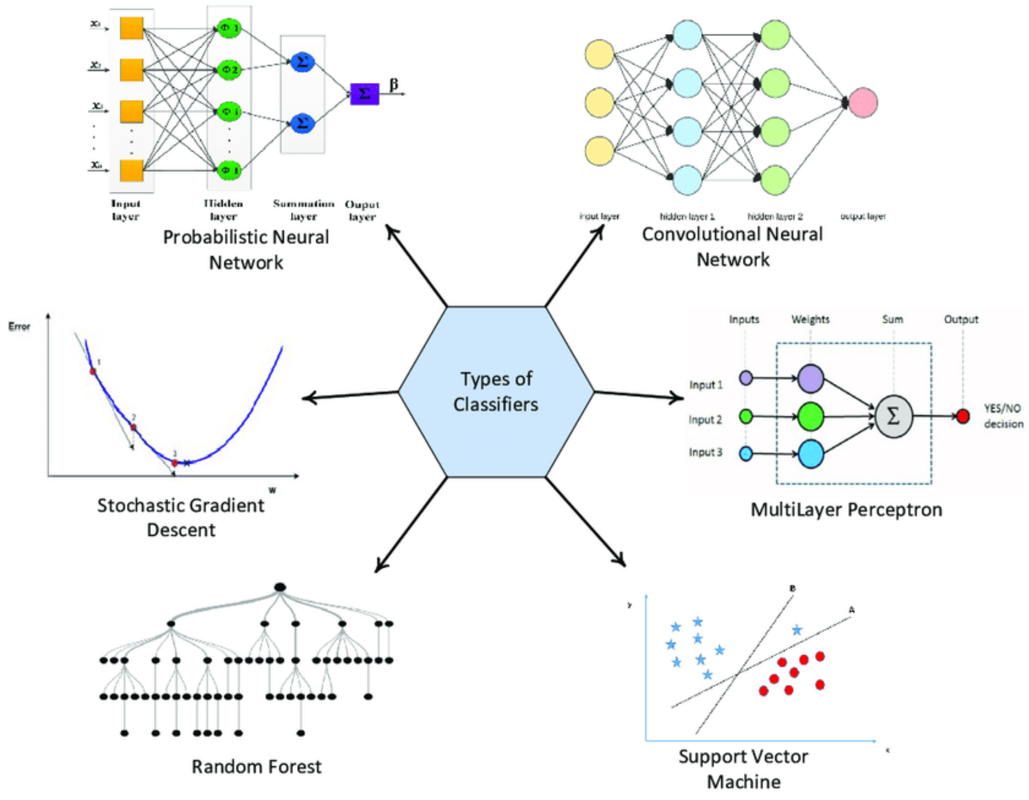
Ensemble Approach

Combine uncorrelated models to make an unified model

- ML Algorithm Alternatives

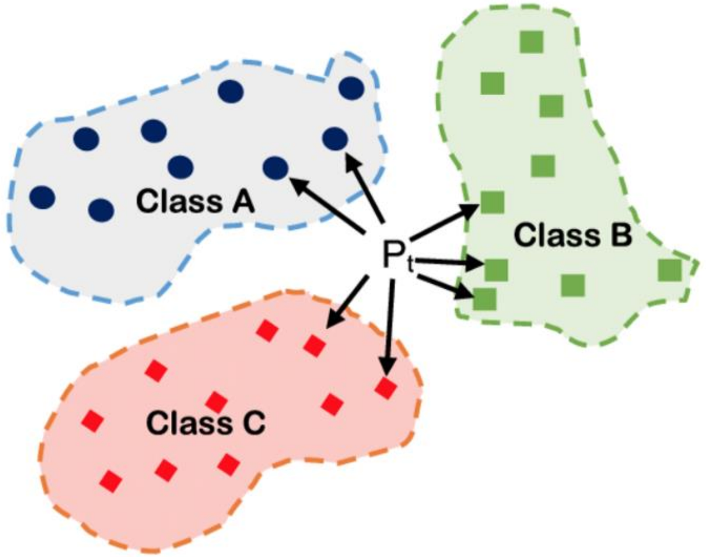
SUPERVISED LEARNING ALGORITHMS

1. K Nearest Neighbor
2. Naïve Bayes
3. Support Vector Machine
4. Logistic Regression
5. Decision Tree
6. Bagging : Random Forest
7. Boosting : AdaBoost, XGBoost, LGBM
8. Stacking: Voting, Stacking
9. Linear Model Family
10. Artificial Neural Network



K NEAREST NEIGHBOR (K-NN)

- Klasifikasi berdasarkan jarak antara titik (Eucleadean distance)
- Menggunakan sejumlah titik terdekat (k) sebagai penentu



Classification : class mode of neighbours
Regression : mean of neighbour’s values

NAÏVE BAYES ALGORITHMS

- Classification berdasarkan conditional probability, Bayes Theorem
- Asumsi bahwa setiap *predictor* tidak saling terkait

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- c = Kelas/ kategori yang menjadi target prediksi
- X = Data yang akan diprediksi kelasnya
- x1, x2, x3, …. xn = Feature dari data X yang diprediksi kelasnya

Only Classification

NAÏVE BAYESIAN ALGORITHMS

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

Training Data

We want to predict unlabeled instance #8

$$P(PG=Y|i_8) \propto P(O=sunny|PG=Y)P(T=mild|PG=Y)P(H=high|PG=Y)P(PG=Y)$$

$$\propto \frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{4}{7}=\frac{1}{28}$$

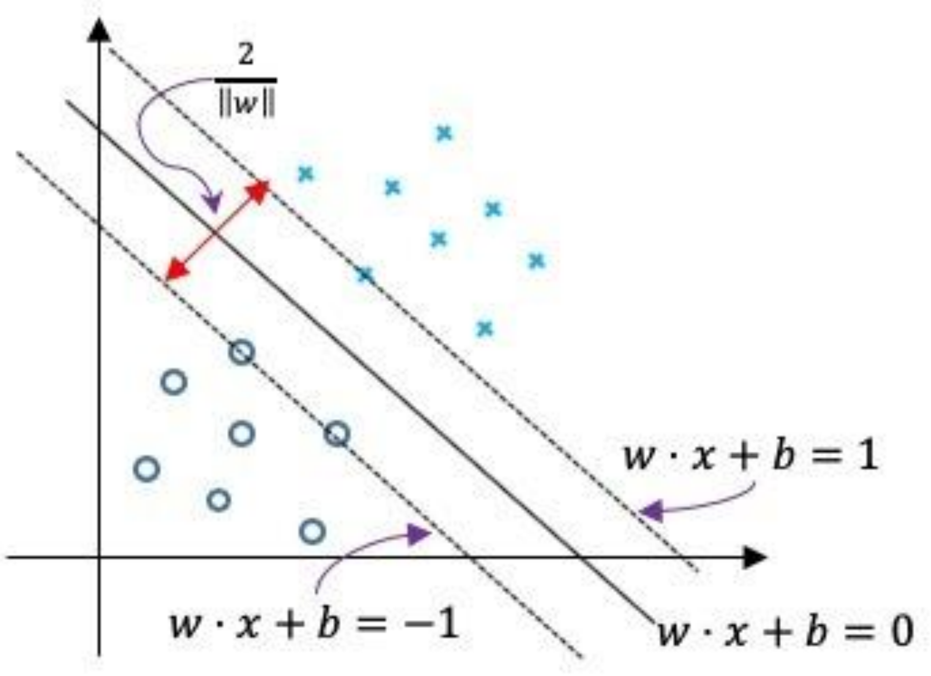
$$P(PG=N|i_8) \propto P(O=sunny|PG=N)P(T=mild|PG=N)P(H=high|PG=N)P(PG=N)$$

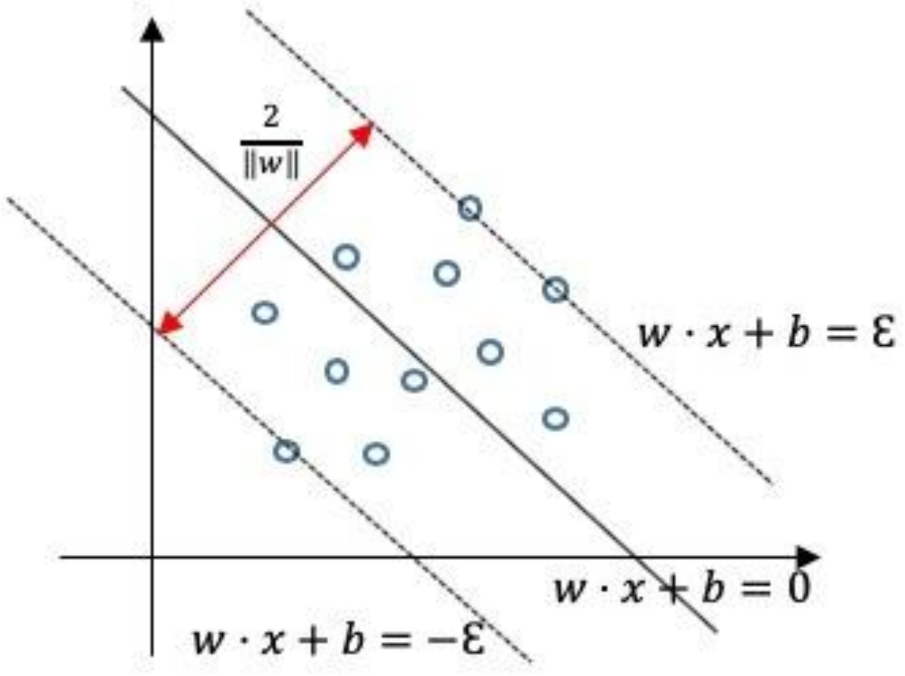
$$\propto \frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\times\frac{3}{7}=\frac{4}{63}$$

P(PG=Y|i8) > P(PN=Y|i8)
Sehinga kemungkinan Play Golf dengan kondisi i8 adalah NO

SUPPORT VECTOR MACHINE

- Menemukan hyperplane yang optimal untuk membagi data ke dalam 2 atau lebih kelas
- Hyperplane dapat linear maupun non linear
- Kernel Trick

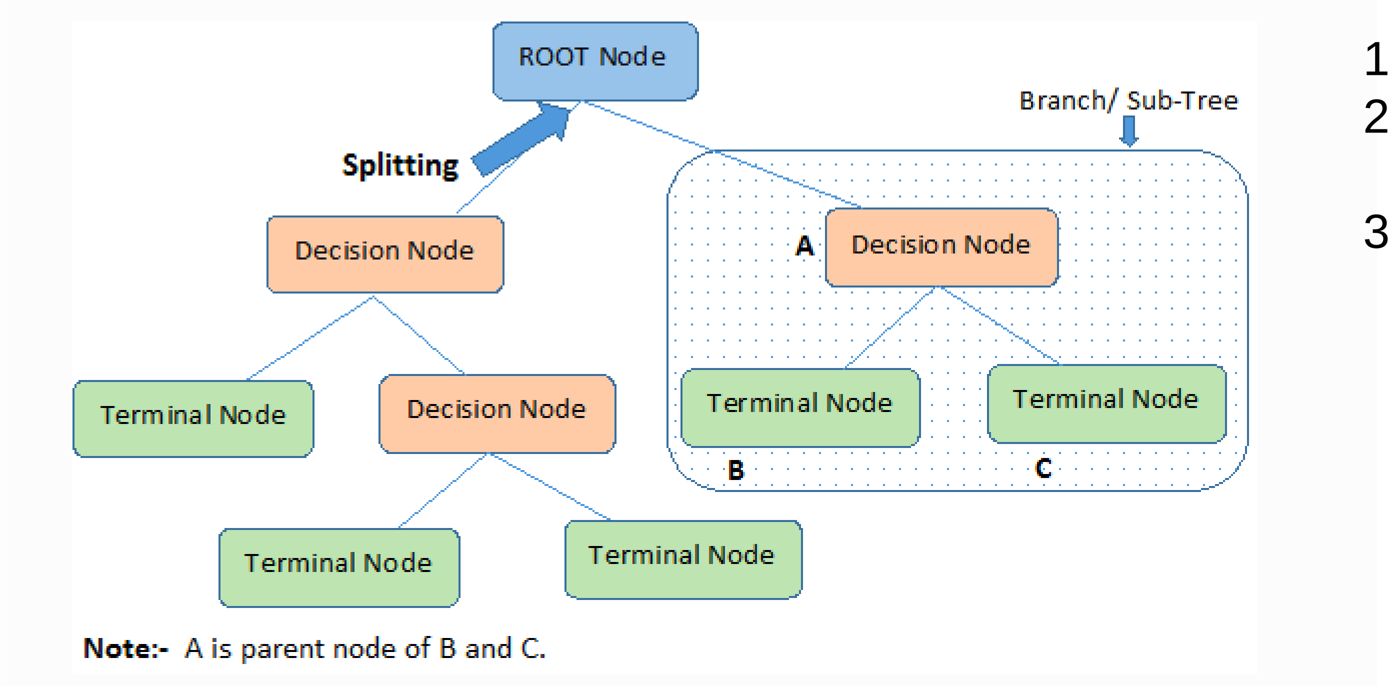




Classification : SVC
Regression : SVR

DECISSION TREE

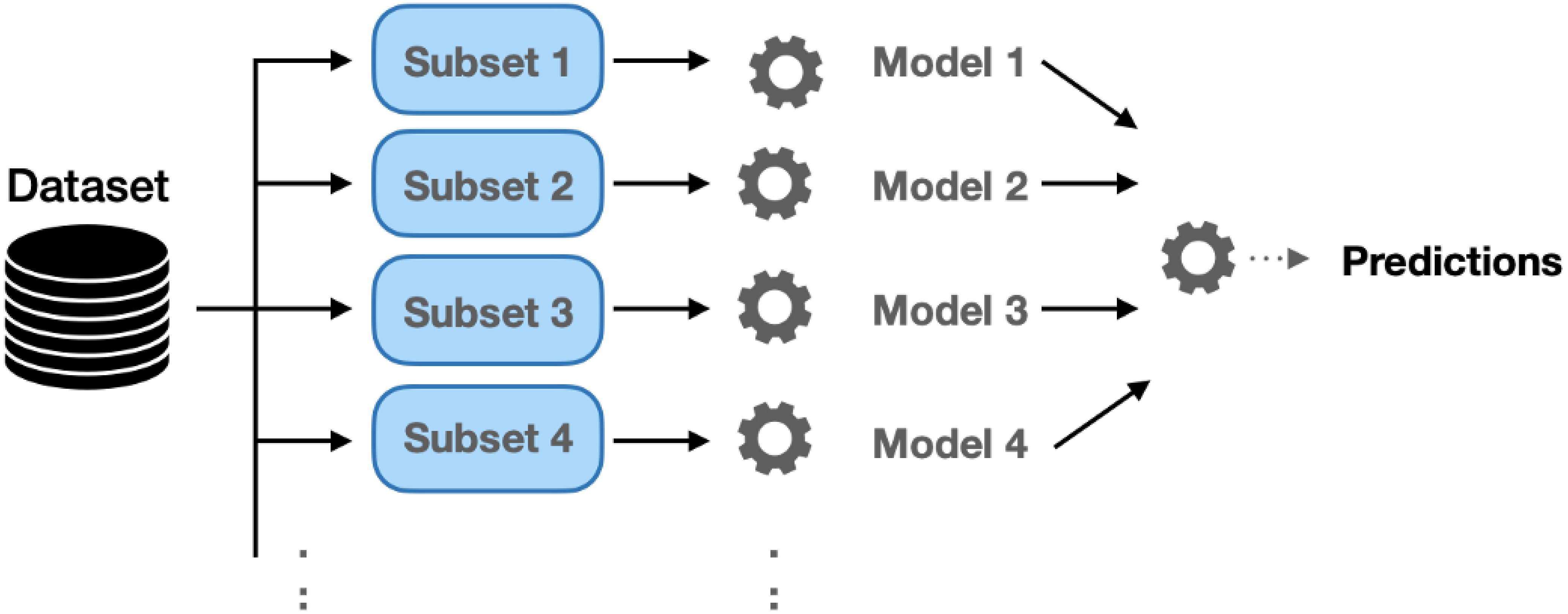
- Classification menggunakan alur berupa pohon keputusan



1. Menentukan root node
2. Menghitung Entropy dan Information Gain secara iterasi
3. Memilih atribut dengan Entropy paling rendah dan Information Gain paling tinggi

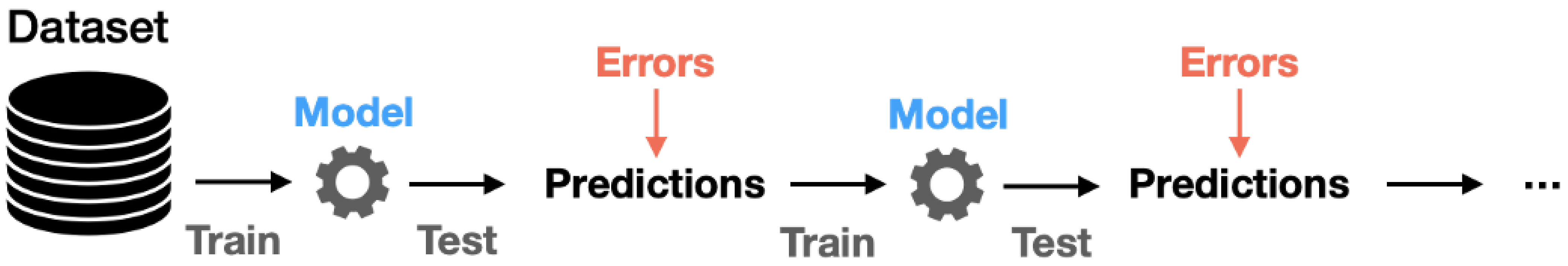
Classification : class mode of predicted nodes
Regression : mean of predicted nodes

ENSEMBLE - BAGGING



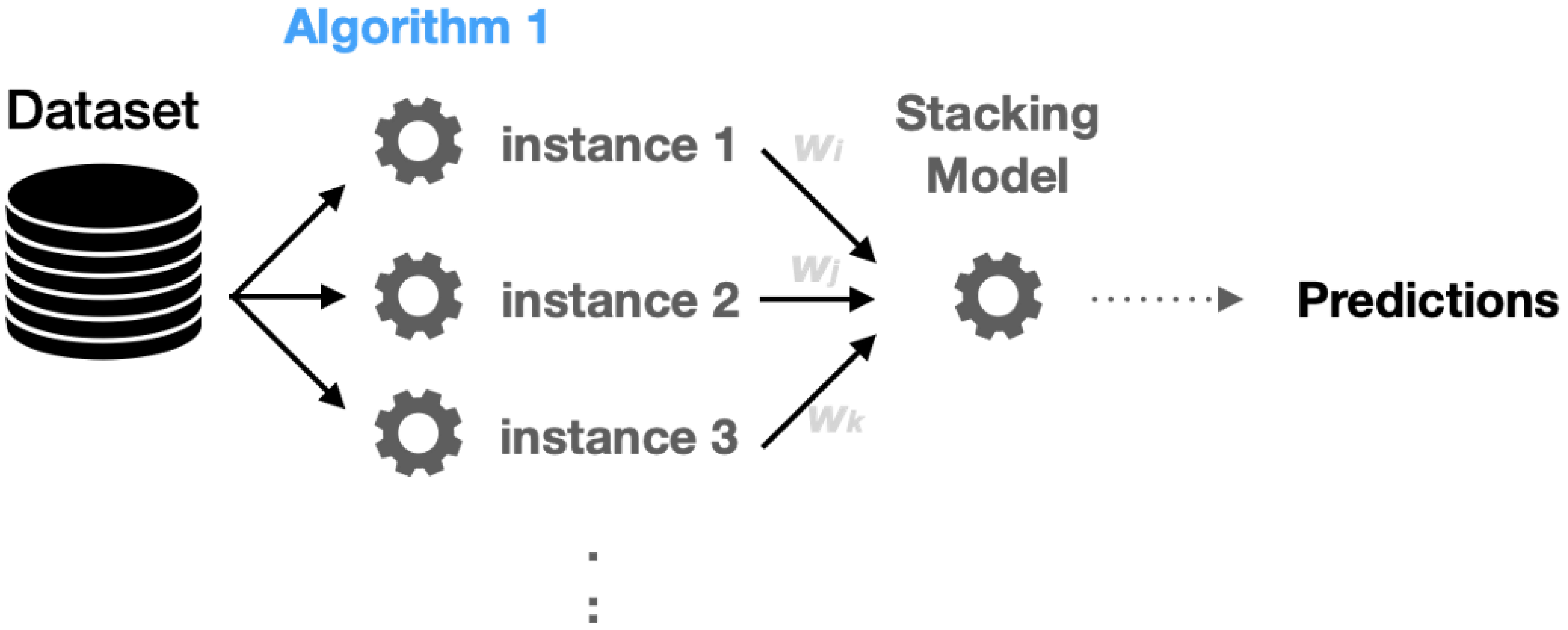
Classification : RandomForest (mode)
Regression : RandomForest (mean)

ENSEMBLE - BOOSTING



Classification : AdaBoost, XGB (mode)
Regression : AdaBoost, XGB (mean)

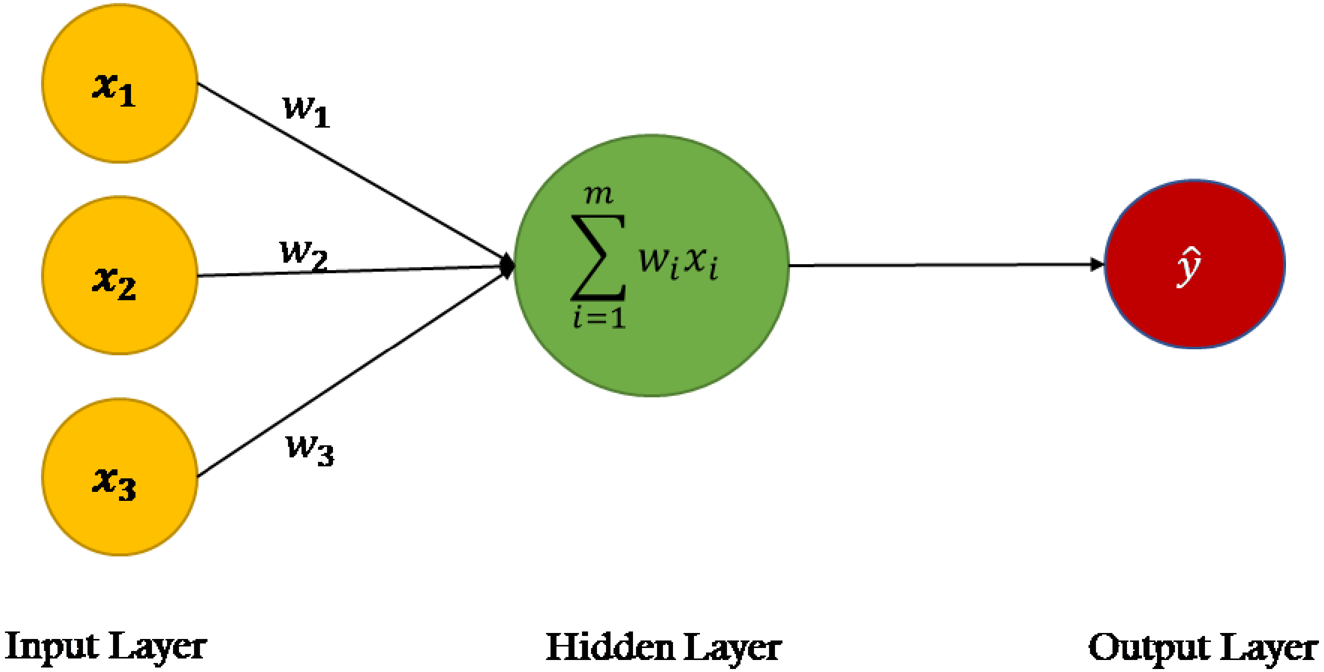
ENSEMBLE - STACKING



Classification : Any classification models
Regression : Any regression models

Linear Model Family

Linear
Logistic
Polynomial
SGD
Perceptron
Lasso |w|
Ridge w^2
Elastic Net



Activation function
- Sigmoid
- Tanh
- Relu
- Softmax
- etc

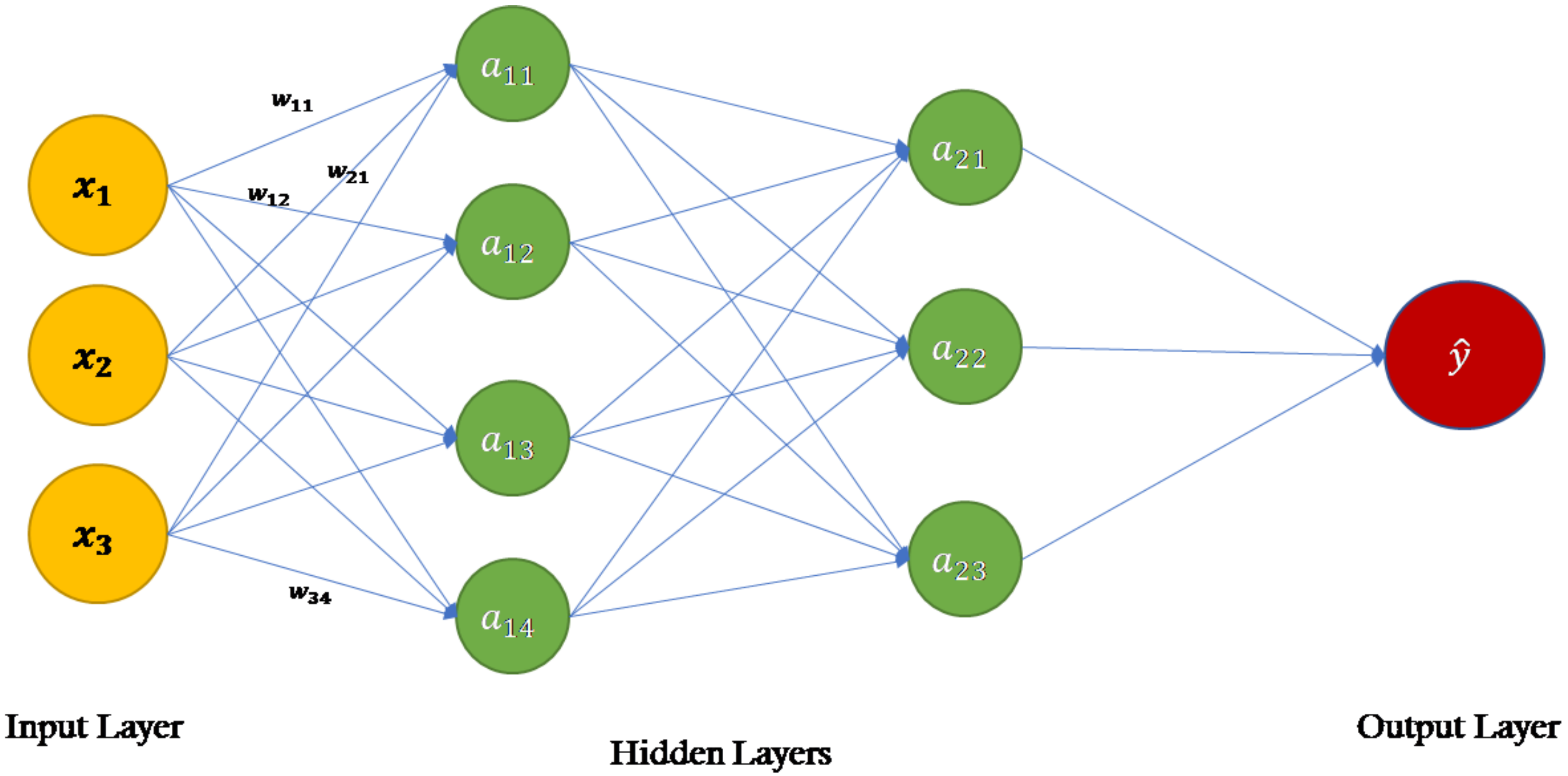
https://towardsdatascience.com/power-of-a-single-neuron-perceptron-c418ba445095

Classification : Softmax on output layer
Regression : no actiovation on output layer

ARTIFICIAL NEURAL NETWORK

Multilayer Perceptron (MLP)

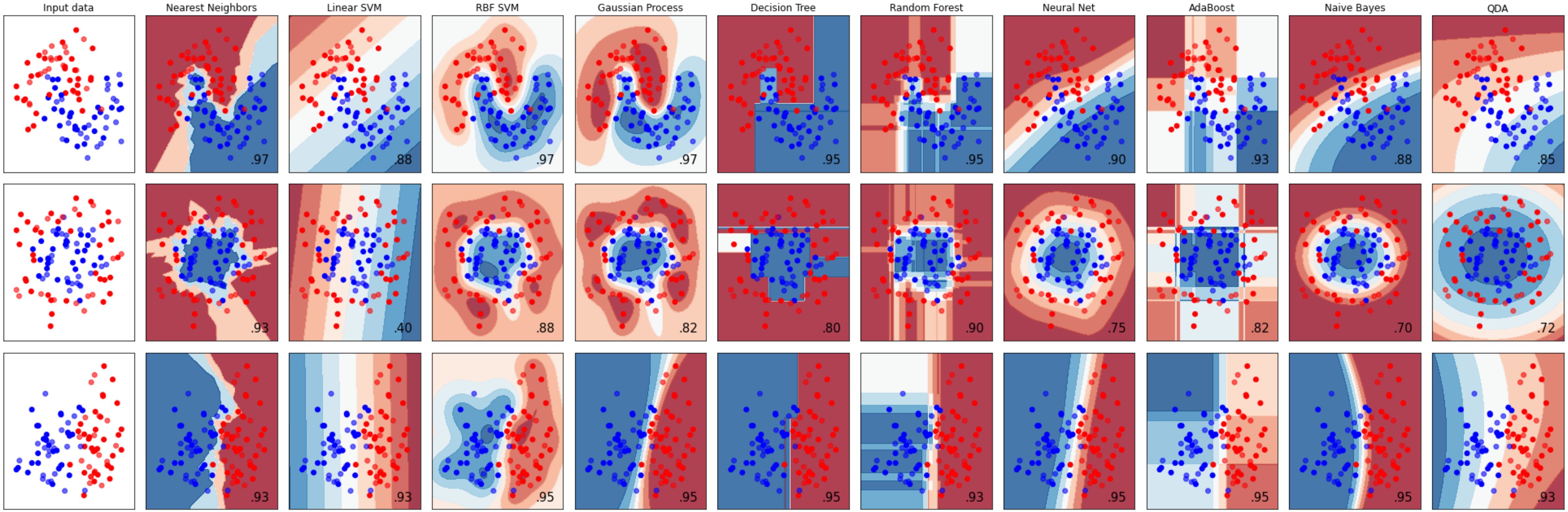
Neural Network



Multilayer Perceptron

https://towardsdatascience.com/power-of-a-single-neuron-perceptron-c418ba445095

CLASSIFICATION ALGORITHMS



https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Let’s Coba

Buka notebook di google colab


TERIMA KASIH