W1D2 – Algoritma & Data Set

Algoritma Machine Learning yang umum dipakai :

1. Unsupervised Learning
2. Supervised Learning
3. Reinforcement Learning

**Data Set**

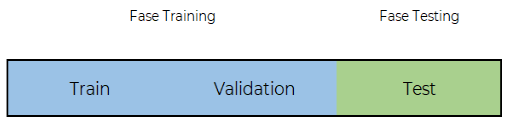


* **Instance:** setiap baris disebut dengan instance. Biasanya instance merupakan satu buah data hasil pengamatan yang terpisah dari data pengamatan lainnya.
* **Feature**: setiap kolom disebut dengan feature. feature menunjukkan komponen dari pengamatan pada data tersebut. Feature biasanya terdiri keseluruhan input data, atau dibagi menjadi input dan output data (label)
* **Dataset:**keseluruhan tabel atau data
* **Data Type**: tipe data yang dimiliki oleh feature. Tipe data ini bisa string, integer, real, ordinal, dsb. Pada machine learning, tipe data ini dikelompokan menjadi dua, yaitu bilangan real dan kategorikal.

**Pembagian Data Set**

1. Unsupervised Learning menggunakan seluruh data set untuk melakukan training
2. Supervised learning tidak menggunakan seluruh data set, melainkan membagi data set tersebut kedalam 3 kategori :
   1. Training Dataset : Dataset digunakan untuk melatih model
   2. Test Dataset : Dataset untuk menguji model, biasanya untuk evaluasi apakah model kita sudah bisa mengolah data
   3. Validation Dataset : melakukan validasi pada dataset yang kita miliki saat kita masih melakukan perubahan pada parameternya

Rekomendasi pembagian dataset :

Train set 60% , Validation Set 20%, Test Set 20%

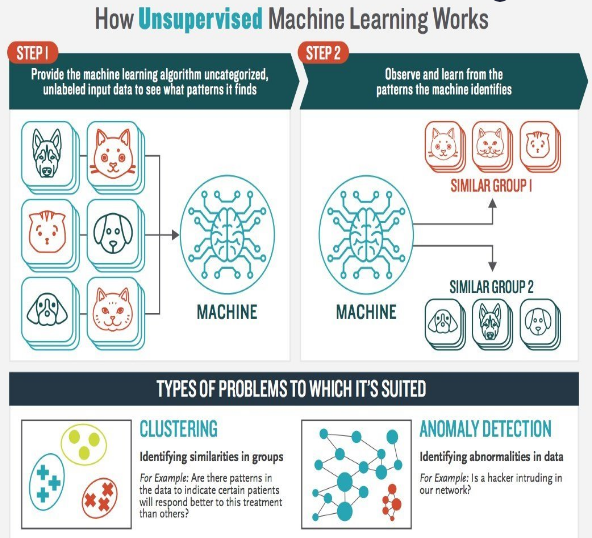
Hal yang perlu diperhatikan dalam pembagian dataset : Kesetimbangan dataset

* **Unsupervised Learning** = Mesin akan belajar bebas tidak ada tuntunan karena mesin hanya akan menerima data input tanpa adanaya data output yang digunakan untuk penuntun.

Unsupervised learning akan membagi data tersebut tergantung pada pola yang terlihat dan manusialah yang menginterpretasikan pembagian tersebut/ Menjelaskan (seperti di Python Jan)

* + Association
  + Clustering

Alogitma yang termasuk unsupervised learning adalah

* + K-Means
  + Hierarchical Clustering
  + DBSCAN
  + Fuzzy C-Means
* 
* Supervised Learning : Data yang ada input dan Output (Label).

Label menjadi penuntun/supervisor.

Model Supervised Learning yang menggunakan data terlabel (y) untuk belajar melakukan mapping fungsi. Fungsi ini akan menghubungkan label/ output (y) dengan input (x),

Y = f(x)

* + Classification
  + Regression

Algoritma yang termasuk supervised learning

* + Support Vectoir Machine
  + Linier Regression
  + Logistic Regression
  + Naïve bayes
  + Decision Tree
  + K-Nearest Neighbor

