KASKUS

Machine Learning 101

August 30, 2017

Muhammad Idrus F
Data Analyst, GDP Labs
muhammad.i.fachruddin[at]gdplabs.id

gdp LABS

OUTLINE

1. Introduction and Definition
2. Type of Machine Learning Problem
3. Common Algorithms
4. Machine Learning Implementation in Industries
5. How to Start Learn ML

INTRO & DEFINITION

History

Definition


MACHINE LEARNING
MACHINE LEARNING EVERYWHERE!
memegenerator.net

"Can machines think?"... The new form of the problem can be described in terms of a game which we call the 'imitation game."

(Alan Turing)

Machine Learning Timeline

1950 — Turing Test

1952 — Game Checker
Arthur Samuel

1957 — Perceptron
F. Rosenblatt

1967 — Nearest Neighbour

1970 - 1980 — AI Winter

1980 — NN Breakthrough
Rediscovery
Backpropagation

1990s — Knowledge-driven to Data-driven
SVM, RF, RNN, LSTM

2000s — Competitive ML
imageNet, MNIST, Kaggle, Netflix Price

2010s — Deep Learning

gdp LABS KASKUS

Machine Learning Definition

“Field of study that gives computers the ability to learn without being explicitly programmed” .

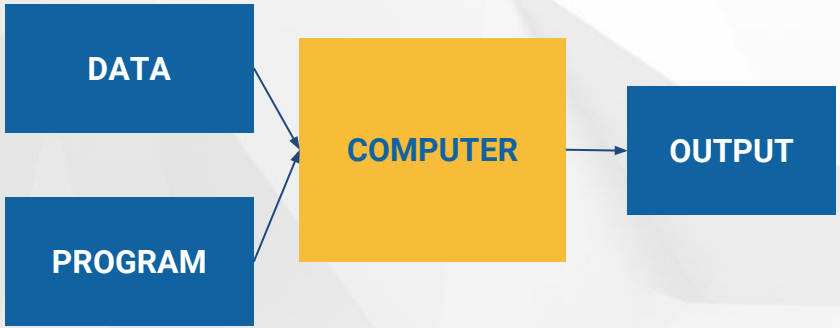
-- Arthur Samuel (1959)

“Software apps are programmed, intelligent apps are trained (with big data)” .

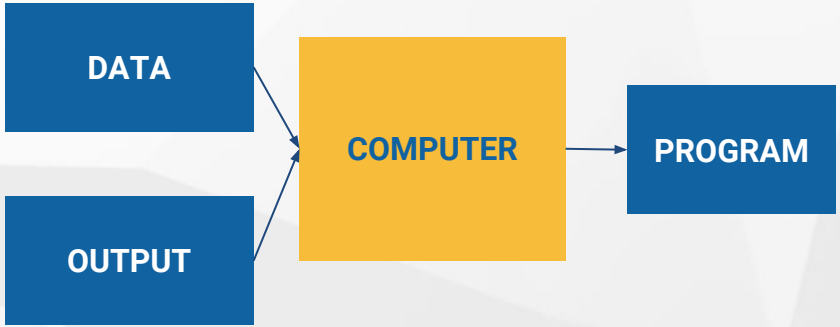
-- Carlos Guestrin

Machine Learning Definition

Traditional Programming

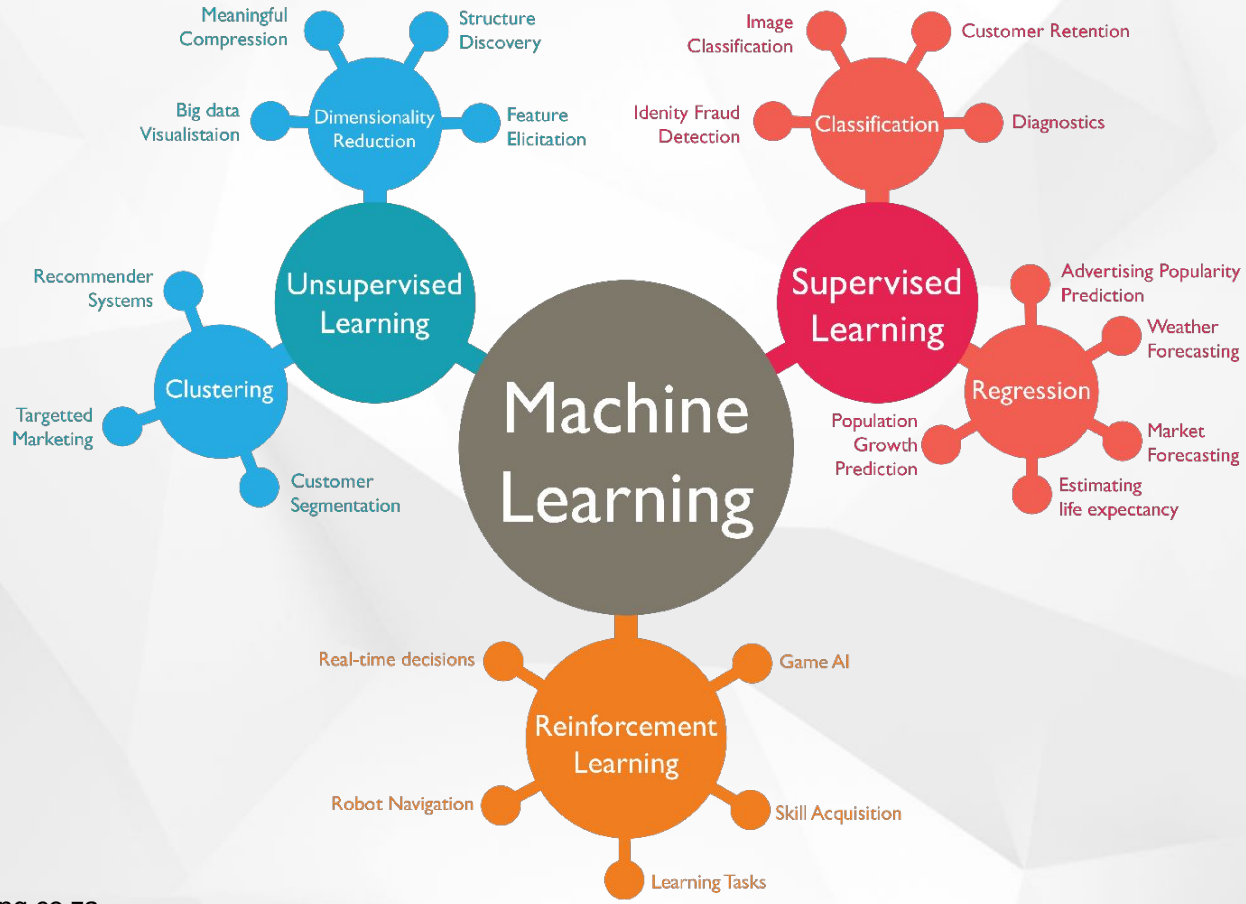


Machine Learning



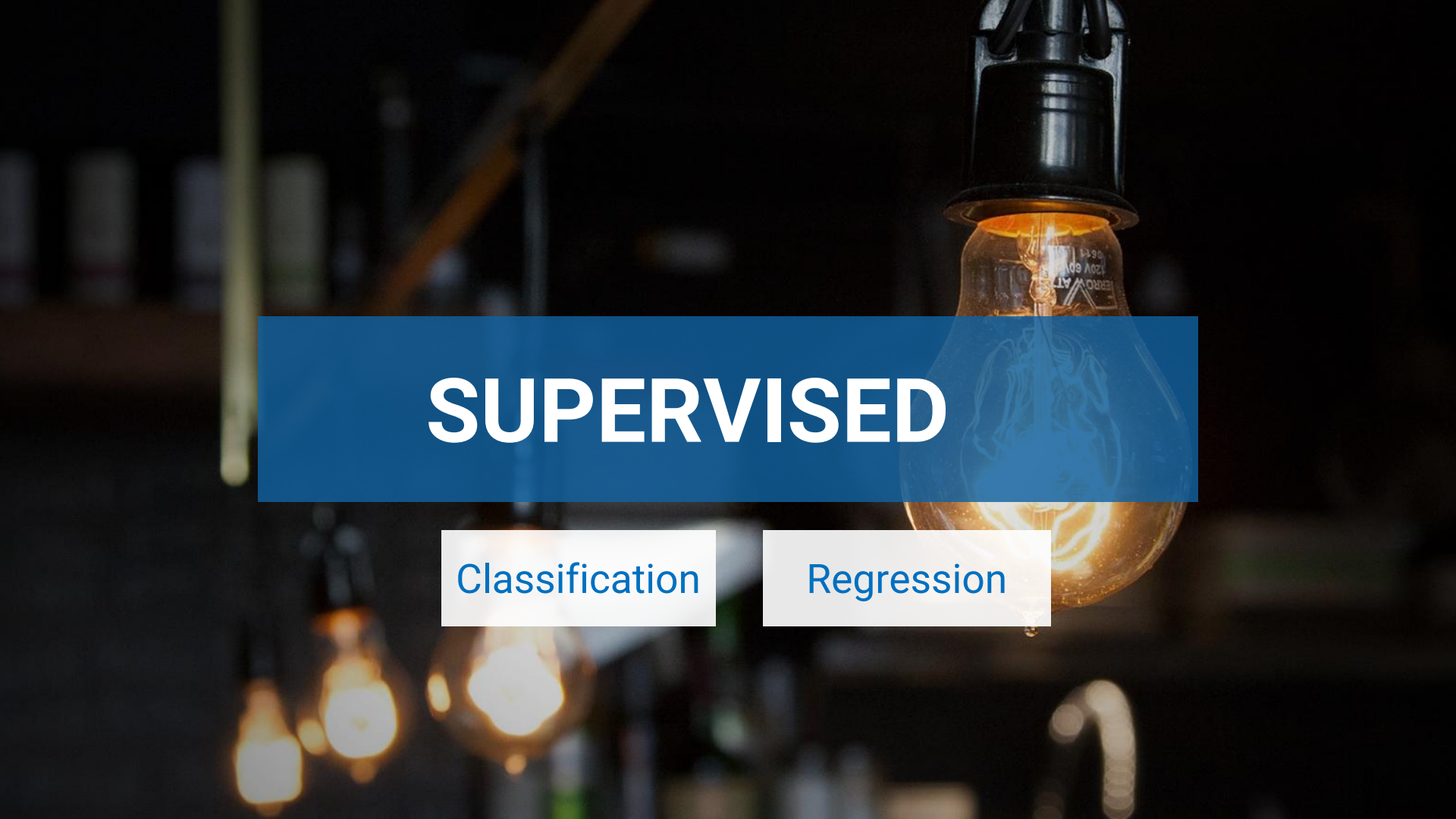
Pedro Domingos,
Machine learning Class

Machine Learning Problem



Picture : www.isaziconsulting.co.za

Supervised Learning

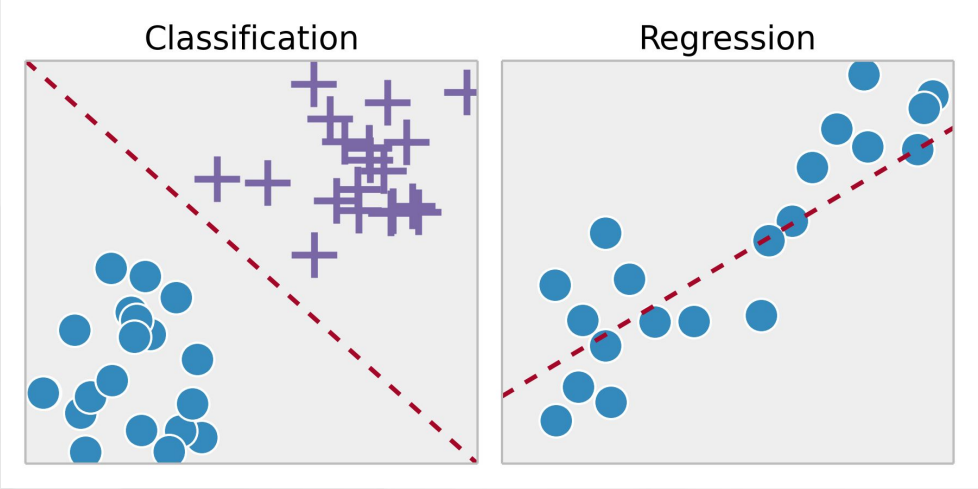


SUPERVISED

Classification

Regression

Supervised Learning



Credit : ipython-books.github.io

Finding a function ***f*** that maps a set of points **X** to a set of labels **Y**, based on given data (x_i , y_i).

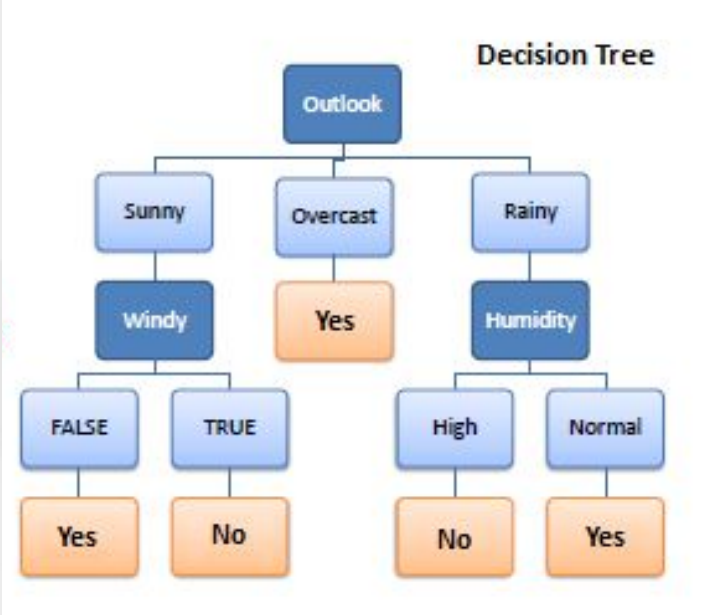
- ***Classification***
  Supervised problem when the target variable (***y***) is categorical.
  Ex : Spam filtering, digit recognition
- ***Regression***
  Supervised problem when the target variable (***y***) is any real (continuous) value.
  Ex: Predict stock market, predict PV of a website, predict ads revenue.

Supervised Algorithms

Popular algorithms:

1. **Tree based model**
   - Decision Tree
2. **Instance Based Learning**
   - k-Nearest Neighbour (kNN)
3. **Ensemble method**
   - Bagging (Random Forest)
   - Boosting (GBT, XGBoost)
4. **Support Vector Machine**
5. **Naive Bayes**
6. **Artificial Neural Network (ANN)**

Tree Based Model : Decision Tree



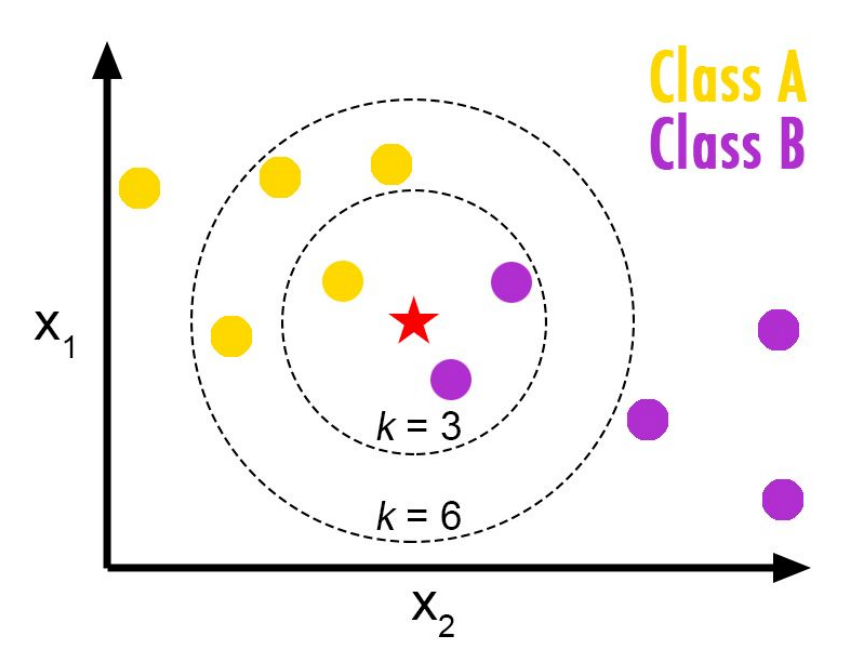
Credit : saedsayad.com

Algorithm : ID3

1. Calculate entropy of the target and attributes
2. Calculate information gain based on entropy of target and every attribute
3. Choose attribute with the highest information gain as the decision node
4. Divide dataset by its branches and repeat the process on every branch
   a. Branch with entropy 0 is a leaf node
   b. Branch with entropy > 0 needs further splitting
5. Run recursively in the non-leaf branches until all data is classified

Instance-based Learning : *k* Nearest Neighbour

A case is classified by a majority vote of its k neighbors.



Credit : proft.me

Algorithm :

- Choose *k*
- Assign a new case (unlabeled data) based on majority class of its *k* nearest neighbour
- If *k* is 1, use distance

Distance functions

Euclidean	$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{k}\lvert x_i - y_i\rvert$
Minkowski	$\left(\sum_{i=1}^{k}(\lvert x_i - y_i\rvert)^q\right)^{1/q}$

Credit : saedsayad.com

Ensemble

Combine ‘weak’ classifier to get better result.

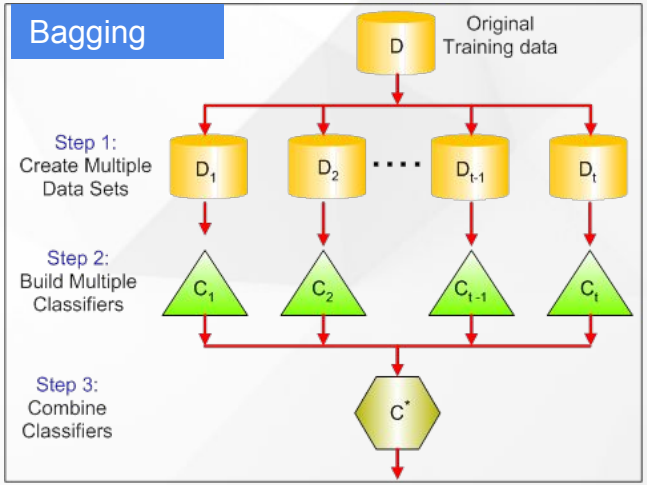
If each voter has probability p of being correct and the majority of voters being correct is M. Then p > 0.5 will imply M > p. M approach to 1 for all p > 0.5 as the number of voters approaches infinity.

Marquis de Condorcet theorem

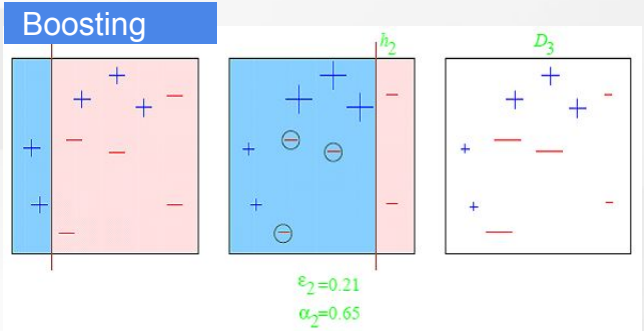
Ensemble

Type of ensemble method :

- **Bagging (Bootstrap aggregating)**
  Build many classifiers from smaller sampling then combine (average or vote) the result.
  Ex : Random Forest, AdaBoost



- **Boosting**
  Improve result by boosting the weak part of latest classification result.
  Ex : Gradient Boosting, XGBoost

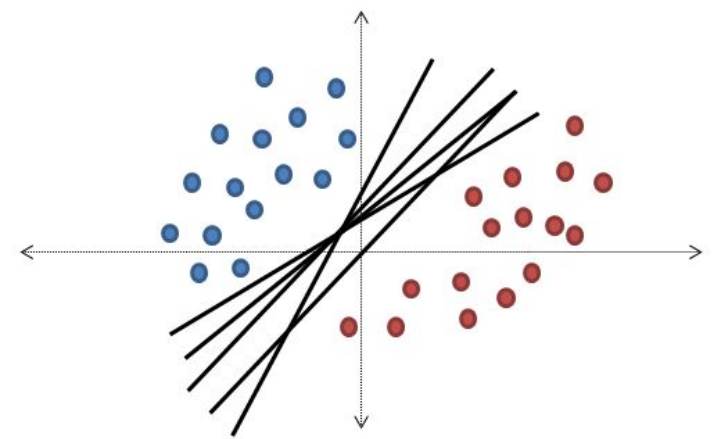


Credit : analytics vidhya

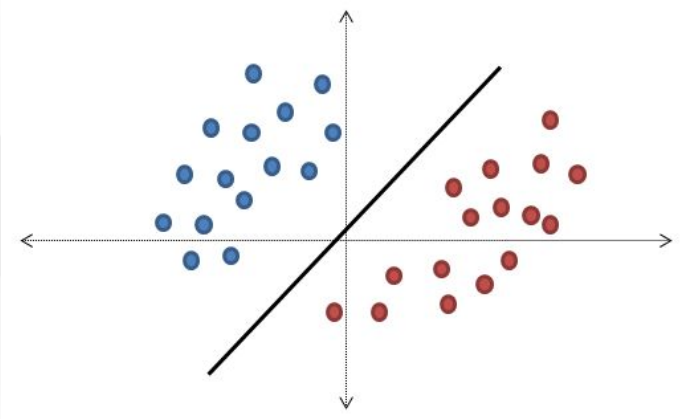
Support Vector Machine

Performs classification by finding the hyperplane that **maximizes the margin** between the two classes.

- We can choose infinite hyperplane (line) to separate two class (a)
- SVM finds hyperplane with the largest margin from each class by using supporting vector (b)



(a)

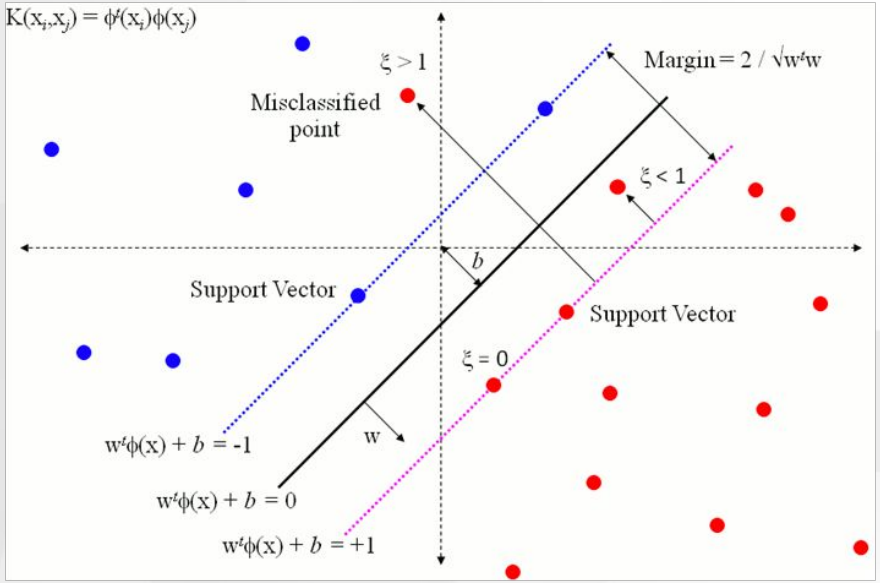


(b)

Credit : stat.exchange

Support Vector Machine

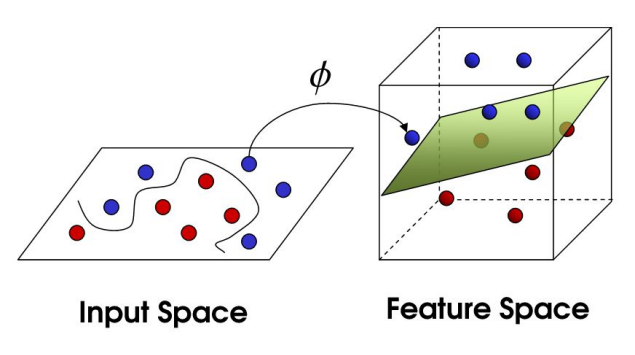
Linear hyperplane



Credit : stat.exchange

SVM has **Kernel** to solve non-linearity problem

- Mapping data to higher dimension
- Type of kernel :
  - Linear
  - Polynomial
  - Radial Basis

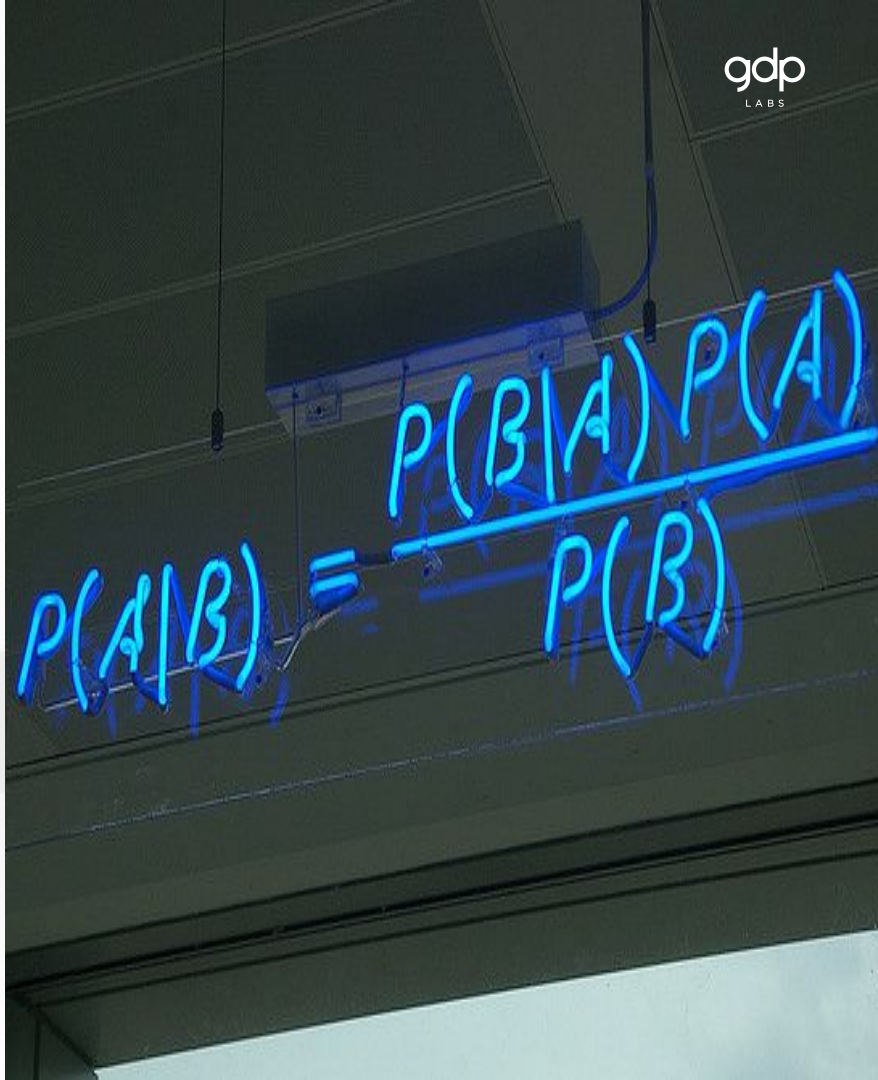


Credit : Linkedin

Naive Bayes

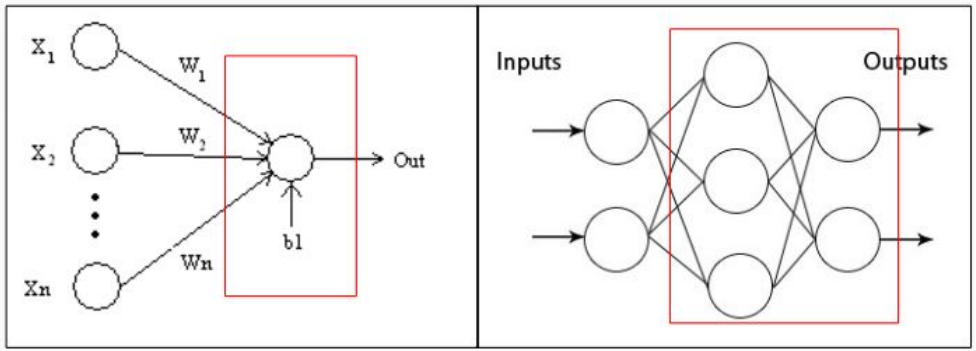
The Naive Bayesian classifier is based on **Bayes' theorem** with independence assumptions between predictors.

- Fast to compute due its independent assumption
- Mainly used for large data. Ex: sentiment analysis
- Quite promising result



Artificial Neural Network (ANN)

An artificial neuron is a computational model inspired in the natural neurons



Credit : Deduction Theory

Perceptron training

- Simple neural function
- Linear classifier

Multi-layer perceptron

- Has the same structure of single perceptron with one or more hidden layer
- Can handle non-linear problem

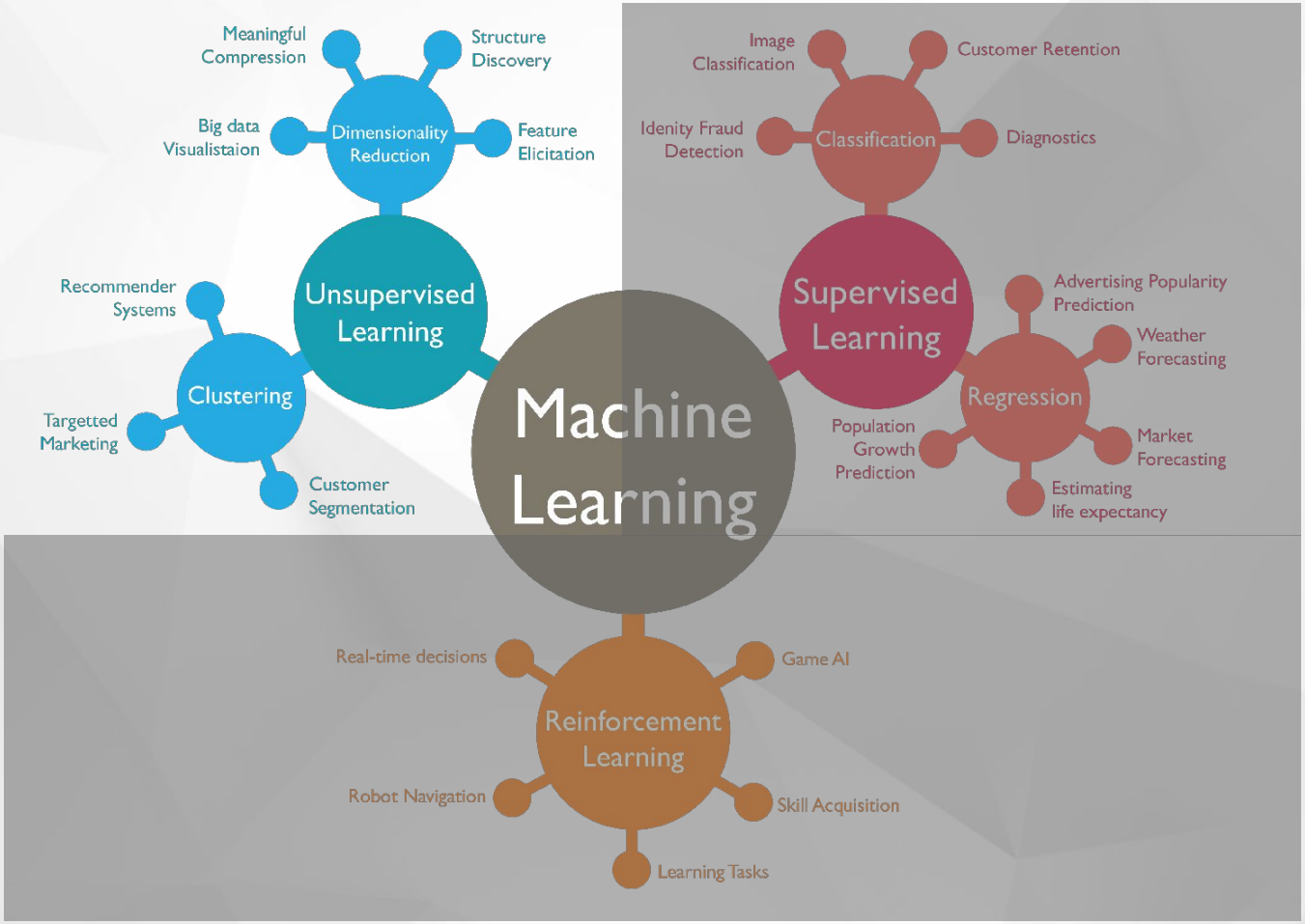
Backpropagation

- Algorithm used to adjust parameter in order to reduce error prediction
- Propagate error backward to adjust the weight both in internal (hidden layer) and external (output layer)

Deep Learning

- Extension of NN
- NN with more than 2 layers

Unsupervised Learning



UNSUPERVISED

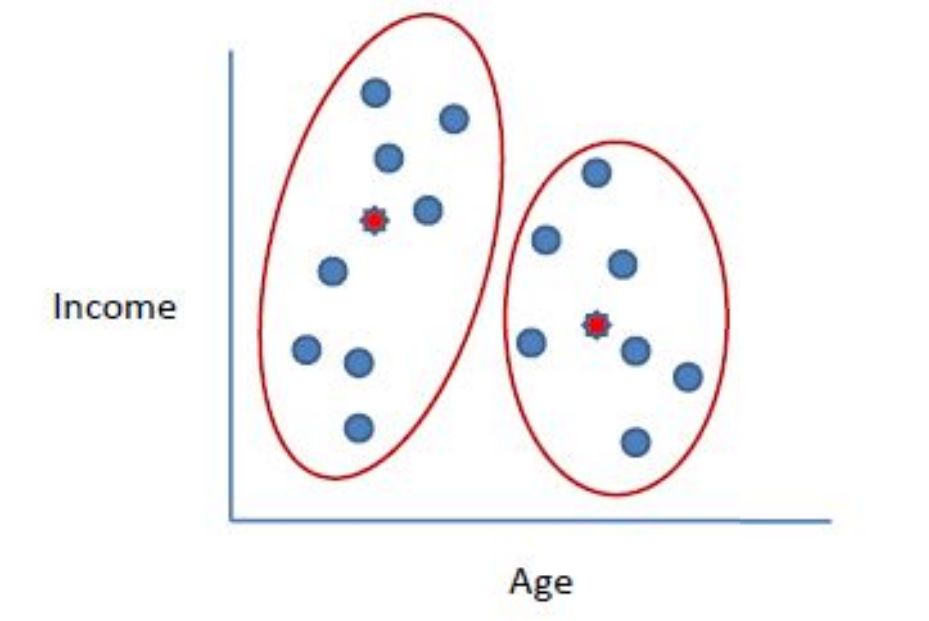
Clustering

Topic Model

Dimensionality Reduction

Clustering : K-Means

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that object in same group is as similar as possible and between group is as different as possible.



Credit : saedsayad.com

K-Means algorithm :

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centers.
3. Assign objects to their closest cluster center according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

Others algorithms :

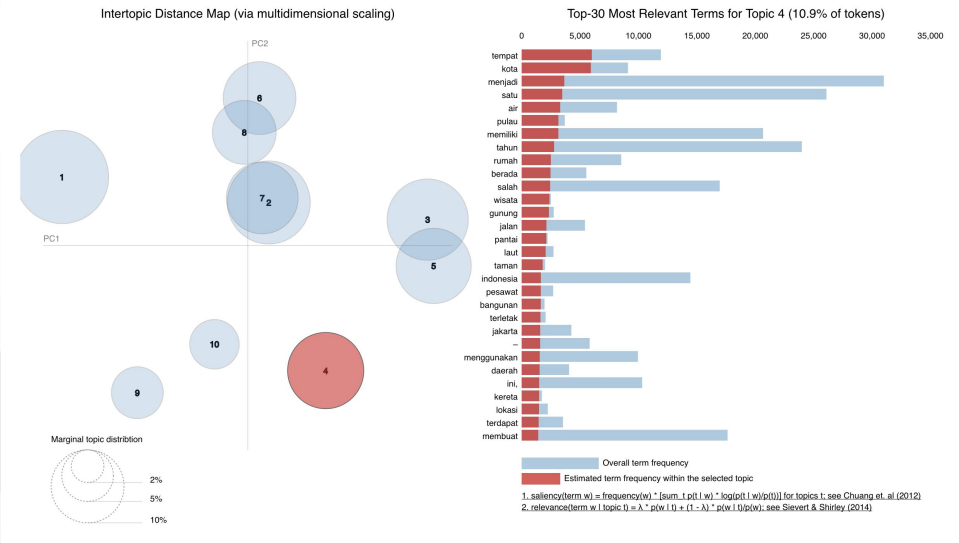
- Hierarchical clustering
- K-prototypes for mixed attributes (numeric & categorical)
- DBSCAN

Topic Model : LDA (Latent Dirichlet Allocation)

LDA represents documents as **mixtures of topics** that spit out words with certain **probabilities**.

Applications

- Understanding set of large documents
- Automated article tagging
- Recommendation system : LDA-based recommendation system



Application of topic modelling at Kaskus, grouping contents in The Lounge forum into several topics/categories.

Credit : blog.echen.me

Dimensionality Reduction

gdp LABS

Definition

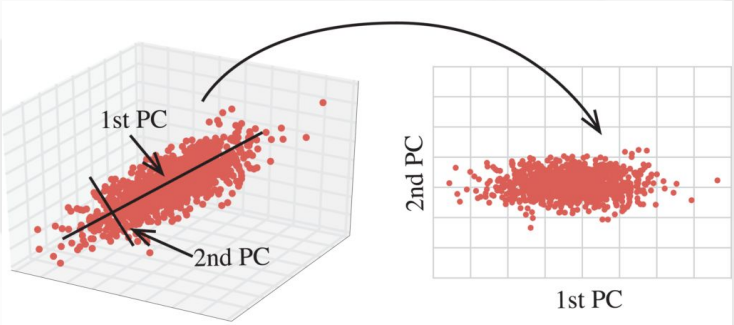
Process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely. (analytics vidhya)

Algorithm

- Principal component analysis
- Factor analysis
- Canonical correlation analysis
- Low variance filter

Benefits

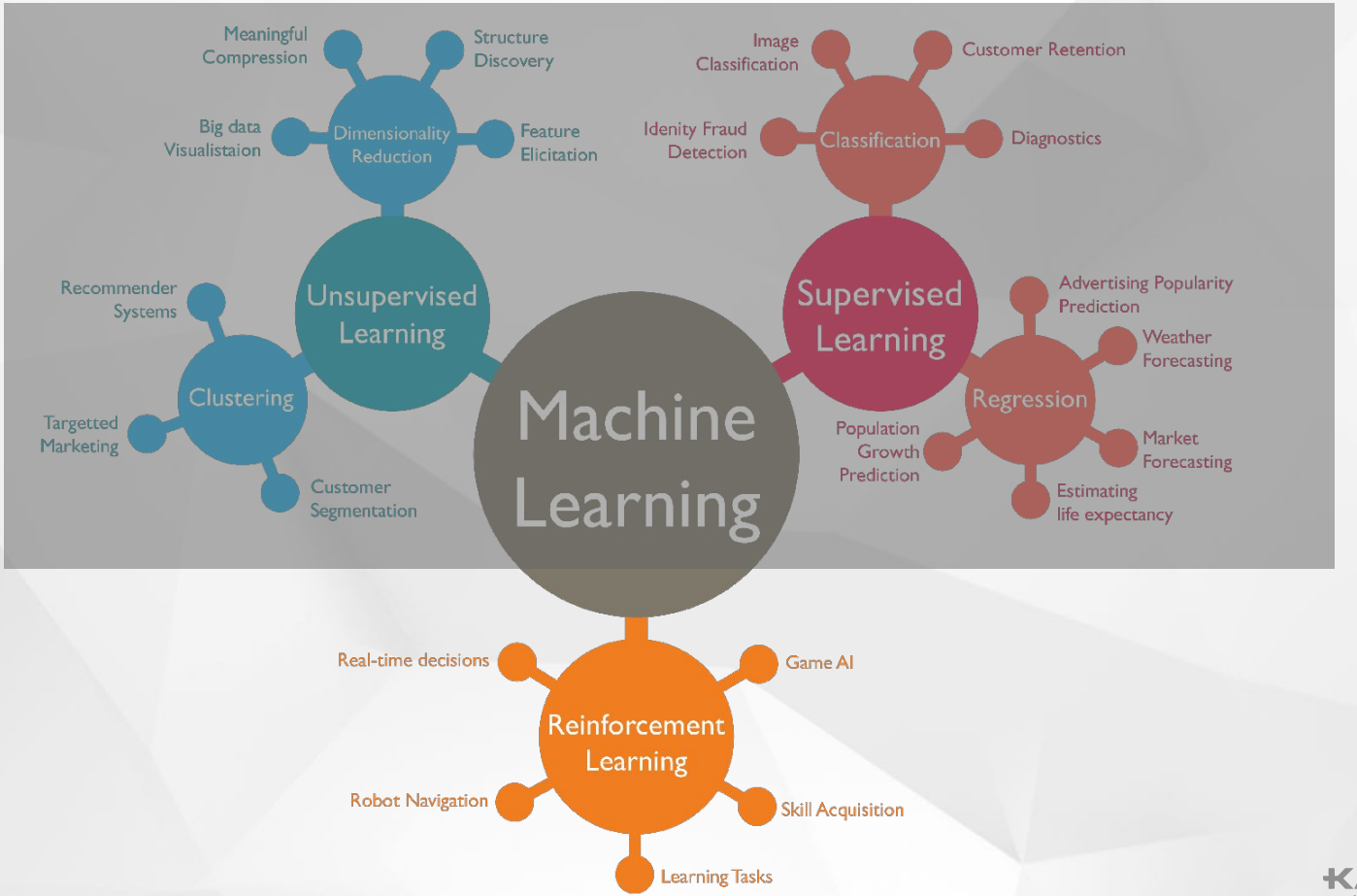
- Reducing storage space
- Fastens the time required for training
- Improving performance i.e it takes care of multicollinearity problem
- Reducing dimension of data to 2D or 3D may allow us to plot and visualize.



Credit : kaggle

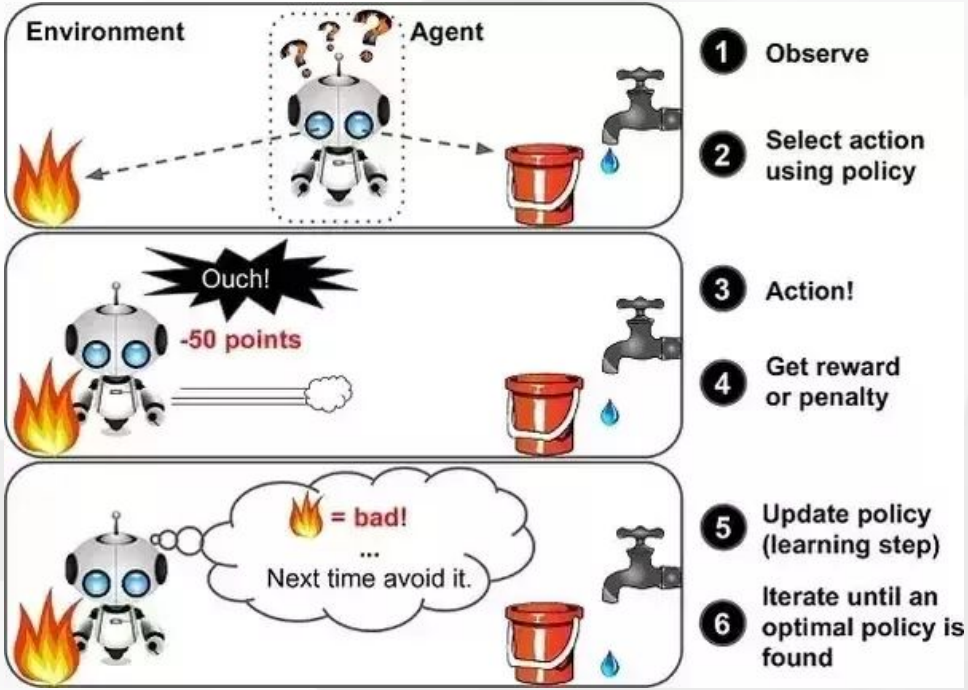
REINFORCEMENT LEARNING

Reinforcement Learning



Reinforcement Learning

Concerned with how ***agents*** ought to take ***actions*** in an ***environment*** so as to maximize some notion of cumulative ***reward***. (Wikipedia)



Credit : marutitech.com

Reinforcement Learning

Algorithms:

- Markov decision process
- A/B Testing
- Multi Arm Bandit (MAB)
- Q-learning

Applications:

- Robotics
- Real time decision
- Game theory and multi-agent interaction

Alpha Go beat Lee Sedol

Credit : Quartz

Example

ML implemented in Kaskus & GDP Labs

HOW TO START

How to start learn machine learning?

1. **Learn Statistics and algebra** (at least basic)
2. **Learn python or R**
3. **Take course** on Coursera, Udacity, etc. Recommended : Machine learning coursera by Andrew Ng
4. **Join community** (offline/online)
5. **Practice**
6. **Compete** (Kaggle, Analytics vidhya, etc)

THANK YOU

gdp LABS

KASKUS
kaskus.id

career.catapa.com/GDPLabs/jobs

gdp LABS

- Help **SISTER COMPANIES**
- Incubate **STARTUPS**
- Constantly **LEARNING**

Founded in
JUNE 2012

140 PEOPLE & GROWING

WON **99** INTERNATIONAL & NATIONAL OUT OF **255** COMPETITIONS

CONTINUOUS LEARNING

HARD WORK

SHARING

PEOPLE CENTRIC

- SOFTWARE ENGINEER
- UI/UX DESIGNER
- DATA SCIENTIST
- DATA ANALYST
- SYSTEM ENGINEER
- PRODUCT MANAGER

- JAKARTA
- BANDUNG
- YOGYAKARTA
- SURABAYA
- BALI

gdp LABS

gdp LABS

GDP Labs

@gdplabs

jobs@gdplabs.id

APPLY NOW!

JOIN OUR TEAM

BALI BANDUNG JAKARTA SURABAYA YOGYAKARTA

- Data Analyst (DA)
- Data Scientist (DS)
- Artificial Intelligent Engineer (AI)
- Graphic Designer (GD)
- Product Manager (PM)
- Software Development Engineer (SDE)
- System Engineer (SE)

career.catapa.com/GDPLabs/jobs

gdp LABS

GDP Labs

@gdplabs

jobs@gdplabs.id

APPLY NOW!

EMPLOYEE BENEFITS

Flexible working hours

Continuous Learning

Various Skills & Knowledge

NO CONTRACT

Training (Abroad/Local)

NO CONTRACT

Attend Conference (Abroad/Local)

career.catapa.com/GDPLabs/jobs

INTERNSHIP PROGRAM

gdp LABS

GDP Labs

@gdplabs

jobs@gdplabs.id

APPLY NOW!

Professional Developments

Practical Experience

Applied Best Practices

Chance to Get Full-time Offering

GDP LABS INTERNSHIP SCHOLARSHIP PROGRAM

Awarded to The Best Interns

No contract for the awardee

career.catapa.com/GDPLabs/jobs