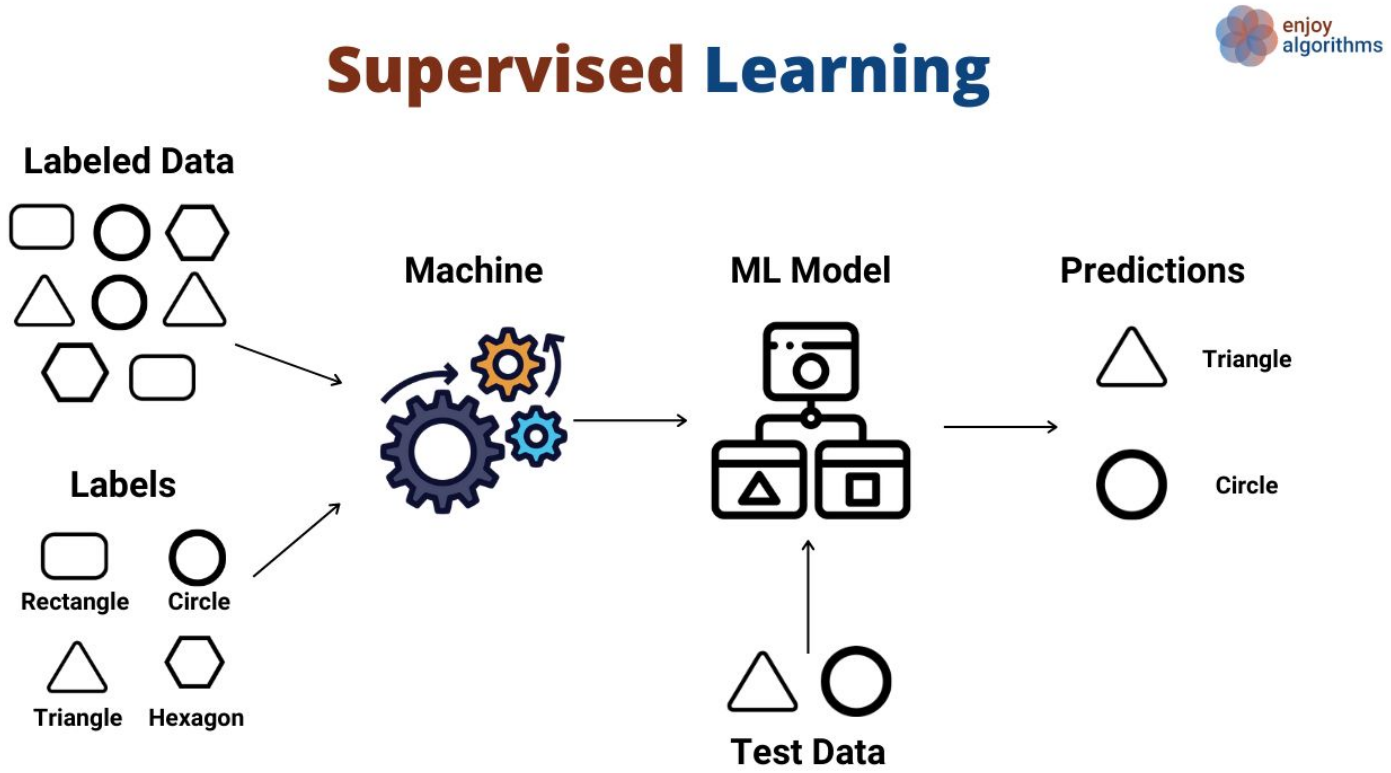
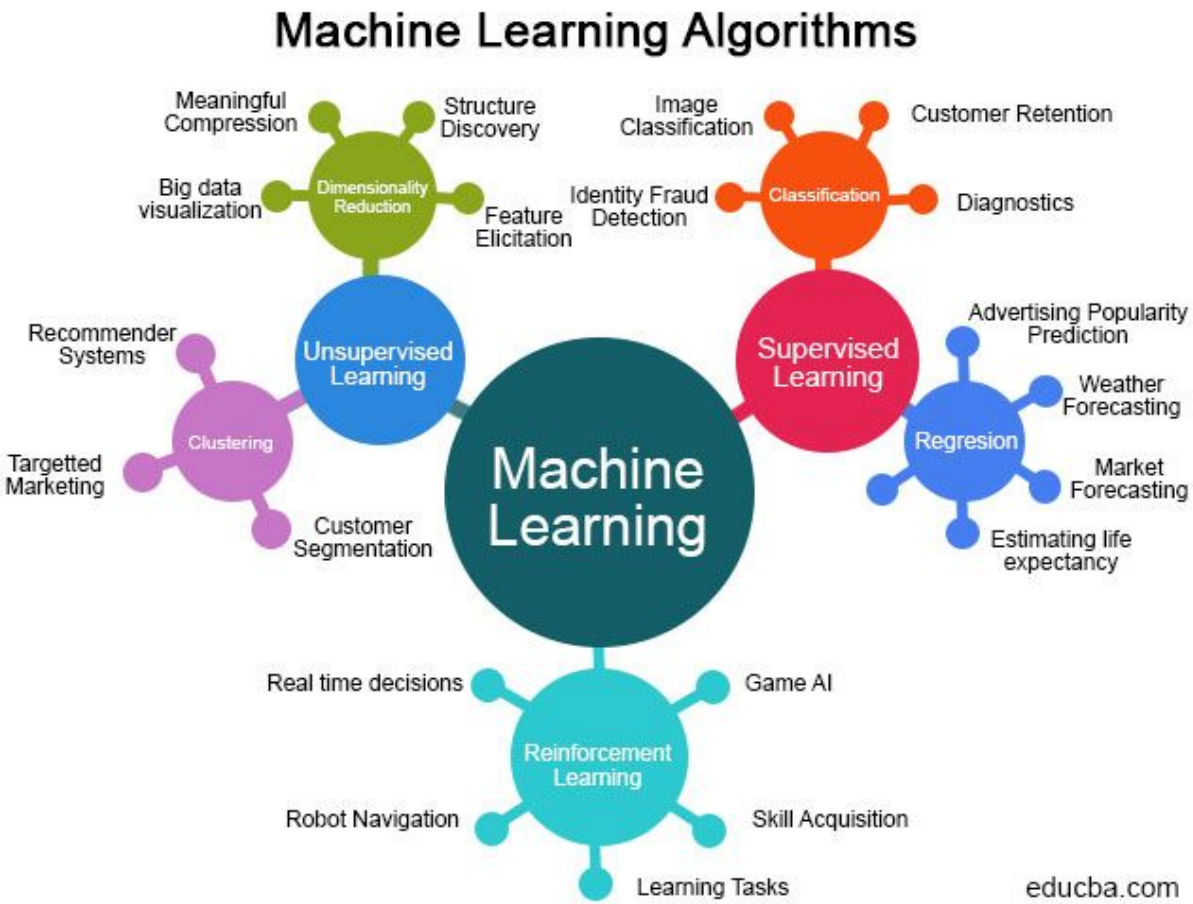
Metrics&Scaling for Classification/Clustering

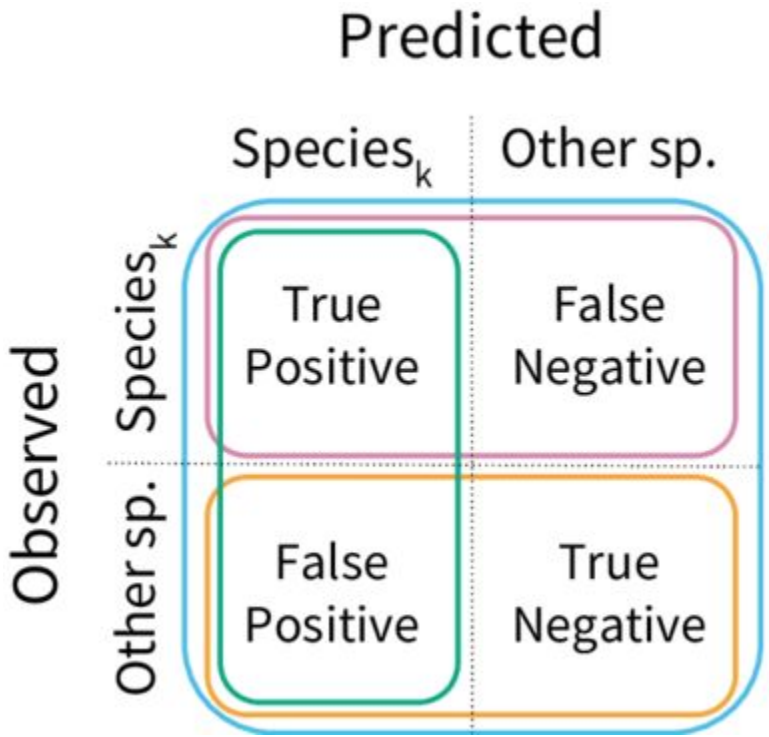
Supervised Learning Algorithm



Big Picture


Machine Learning Algorithms
Meaningful Compression
Structure Discovery
Big data visualization
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Identity Fraud Detection
Classification
Diagnostics
Recommender Systems
Unsupervised Learning
Clustering
Targetted Marketing
Customer Segmentation
Machine Learning
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Regresion
Market Forecasting
Estimating life expectancy
Real time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks
educba.com

Metrics



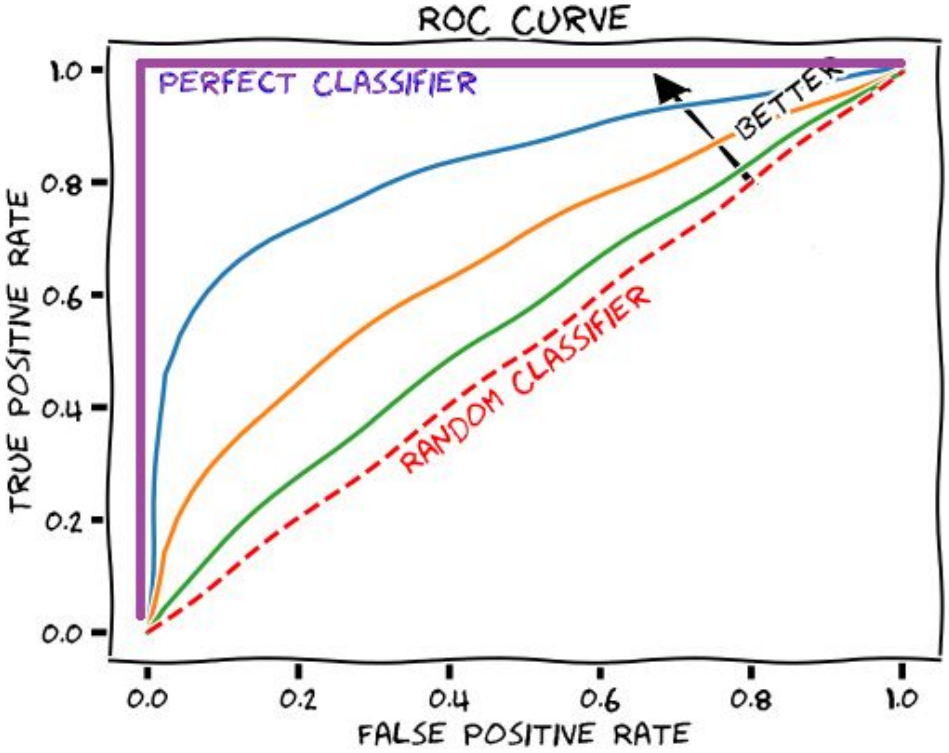
Accuracy $= \dfrac{TP + TN}{TP + TN + FP + FN}$

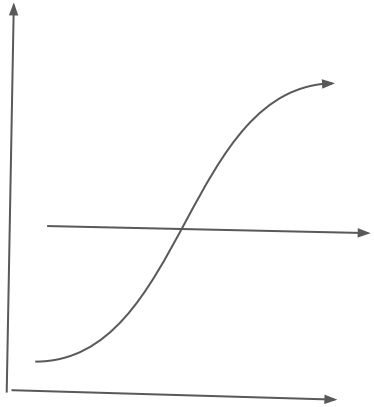
Specificity $= \dfrac{TN}{TN + FP}$

Precision $= \dfrac{TP}{TP + FP}$

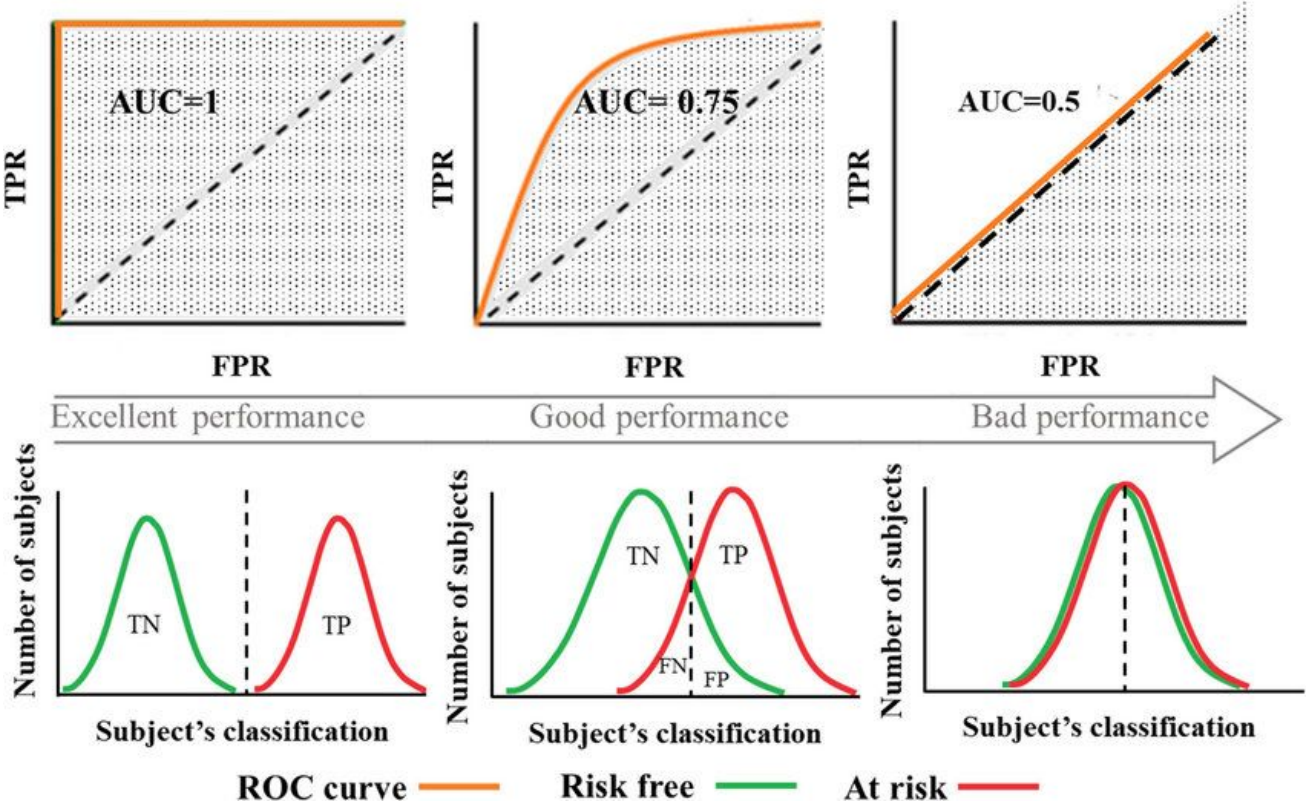
Recall $= \dfrac{TP}{TP + FN}$

Metrics


ROC CURVE
PERFECT CLASSIFIER
BETTER
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
0.0
0.2
0.4
0.6
0.8
1.0



Metrics


AUC=1
AUC= 0.75
AUC=0.5
TPR
FPR
Excellent performance
Good performance
Bad performance
Number of subjects
TN
TP
FN
FP
Subject's classification
ROC curve
Risk free
At risk

Feature Scaling

Normalization or Standardization

- **Feature Scaling means scaling features to the same scale.**
- **Normalization scales features between 0 and 1, retaining their proportional range to each other.**

new value

original value

Normalization

$$X' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Standardization scales features to have a mean (u) of 0 and standard deviation (a) of 1.**

new value

original value

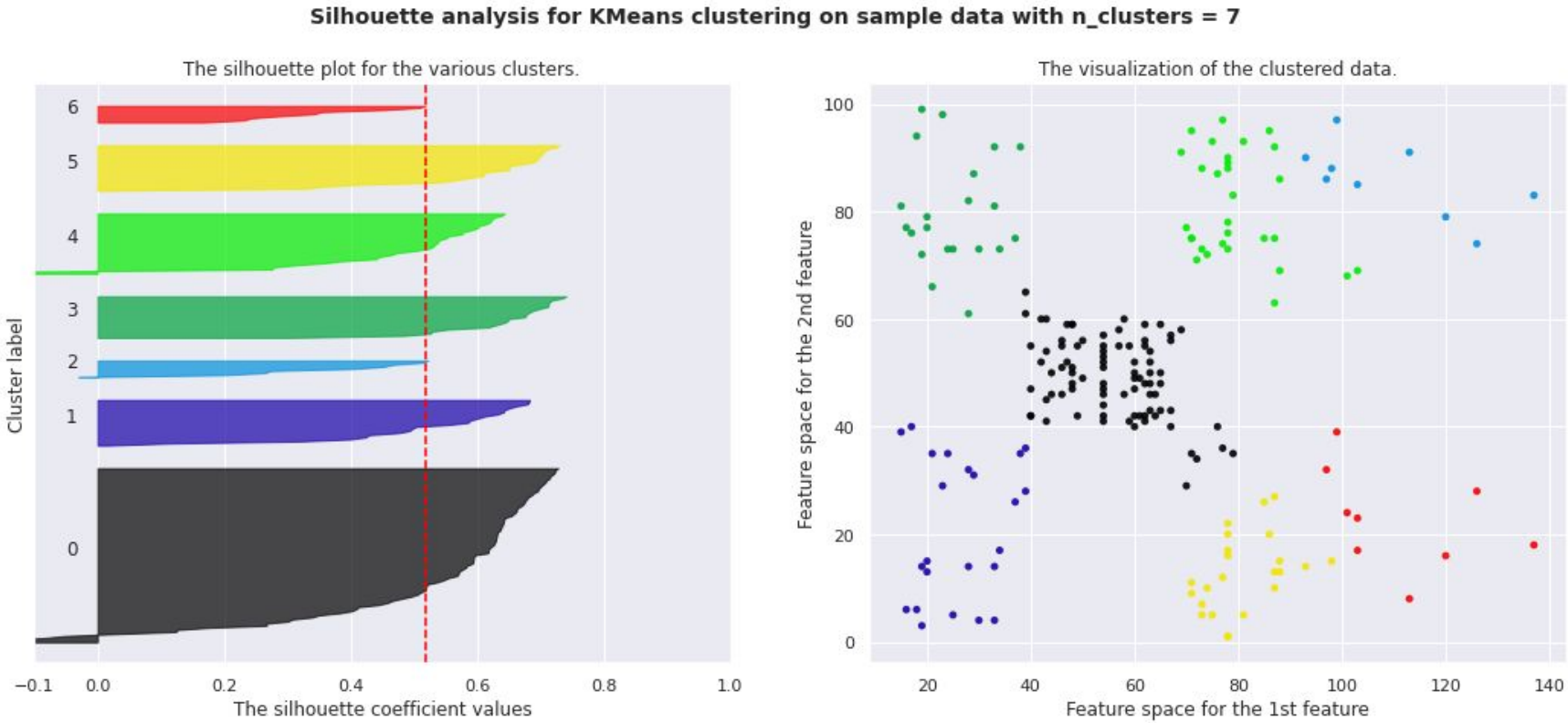
Standardization

$$X' = \frac{x - u}{a}$$

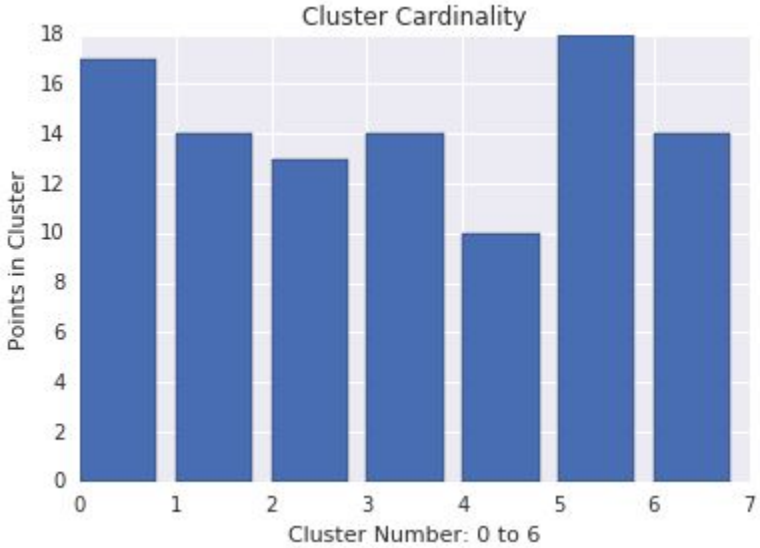
mean

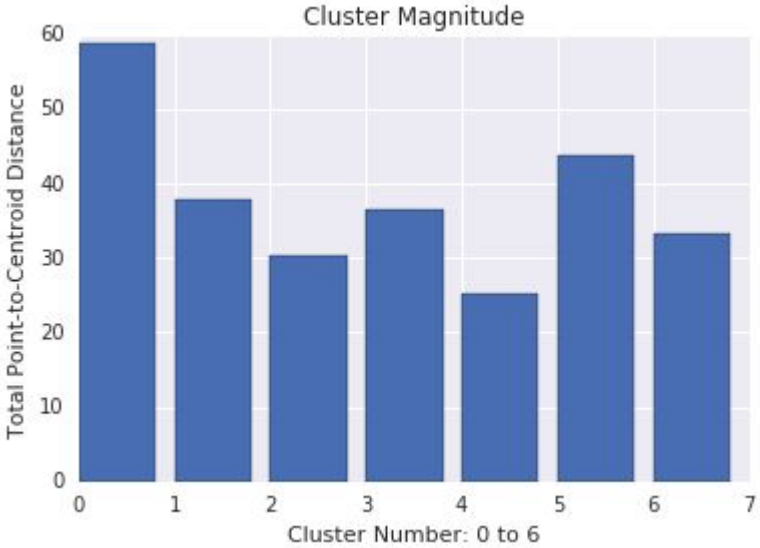
standard deviation

Metrics for Clustering - Silhouette



Metrics for Clustering - K-means





Metrics for Clustering - K-means

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use

$$minimize\left(\sum_{k=1}^{k} W(C_k)\right)$$

where C_k is the kth cluster and $W(C_k)$ is the within-cluster variation