Conceptual elucidation of Machine Learning with Google Colab

Premanand S

Assistant Professor Jr
School of Electronics Engineering (SENSE)
VIT University - Chennai Campus

in association with
STAY LATE AND CODE (SLAC 2020)
Amrita Vishwa Vidyapeetham - Bengaluru Campus

March 25, 2021

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Roadmap

- Data
- Data Characteristics
- Is Statistics is essential for Machine Learning?
- Importance of Statistics
- Basics of Statistical Terms / Data Exploration
- AI Vs ML Vs DL Vs DS
- Machine Learning - Intro
- History of Machine Learning
- Real time examples for Machine Learning
- Machine Learning - MEME!

Roadmap (Cont...)

- Types of Machine Learning
- Algorithms for each types of Machine Learning
- Preferable languages used for Machine Learning
- Why Python?
- Different IDE's for Machine Learning
- Important libraries for Machine Learning
- Numpy - Intro & Hands-on
- Pandas - Intro & Hands-on
- Matplotlib - Intro
- Scikit-learn - Intro

Roadmap (Cont...)

- Steps for Machine Learning
- Metrics - Regression
- Metrics - Classification
- Algorithm Explanation - Support Vector Machine
- Algorithm Explanation - K-Nearest Neighbors
- Algorithm Explanation - Decision Tree
- Algorithm Explanation - Random Forest
- Algorithm Explanation - Principal Component Analysis
- Algorithm Explanation - K-Means Clustering
- Some key jargon of Machine Learning

Roadmap (Cont...)

- Regression - Exercise
- Classification - Exercise
- Trending domains for Machine Learning
- Other important tools / technology for Machine Learning and Data Science
- Some best Course for Machine Learning across the globe
- Tips to improve Machine Learning coding & knowledge
- Deep Learning

So DATA?

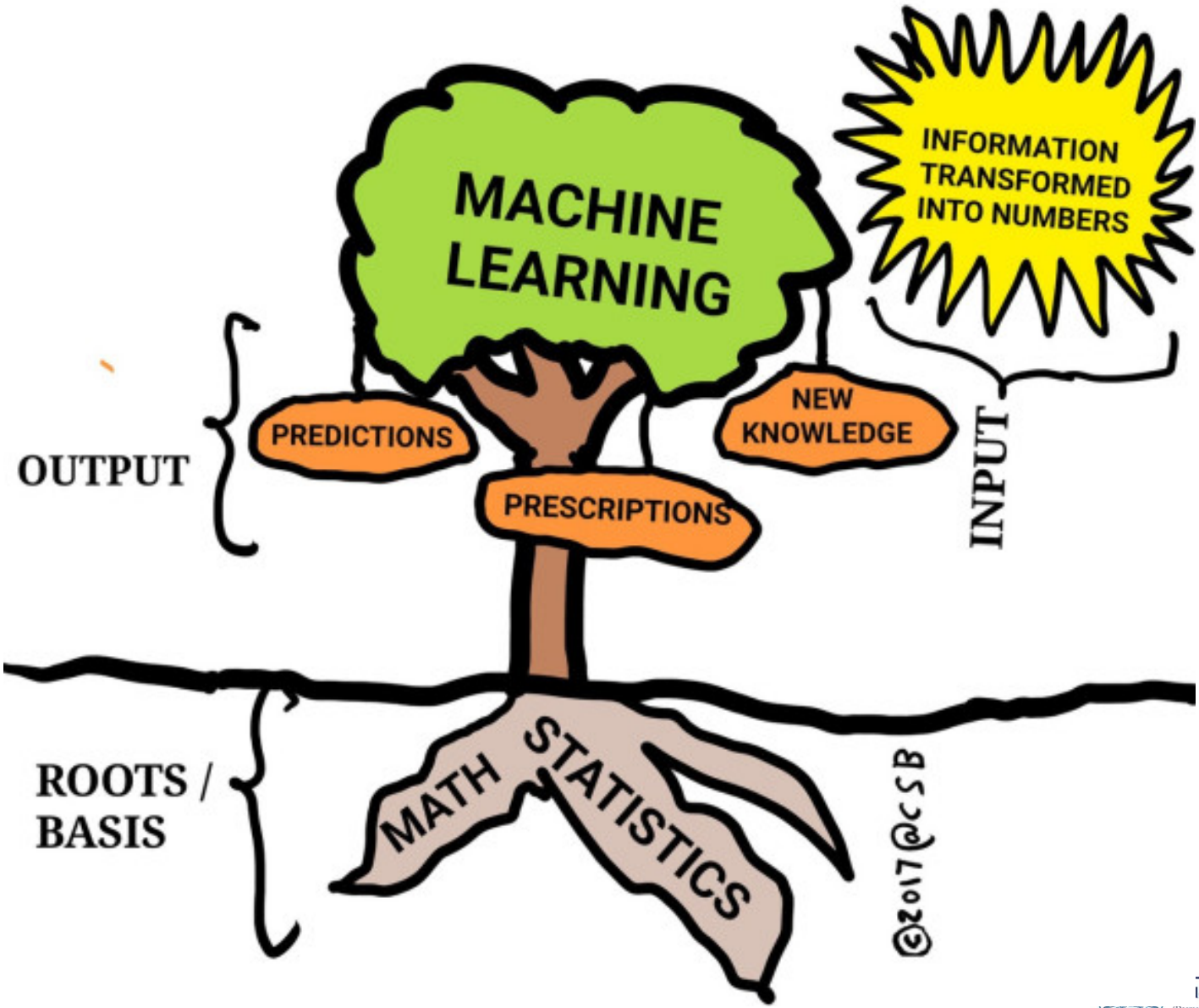
- **Signal** - Information
- **Source** - Smart Phone, Laptop, Sensors, Smart devices & so on...
- **Types** - Image, Audio, Video, Pdf, Word, Excel & so on...
- **Application** - Communication, Social Media, E-Commerce, Sports, Groceries, News, Entertainment, Games & so on...

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

DATA Characteristics

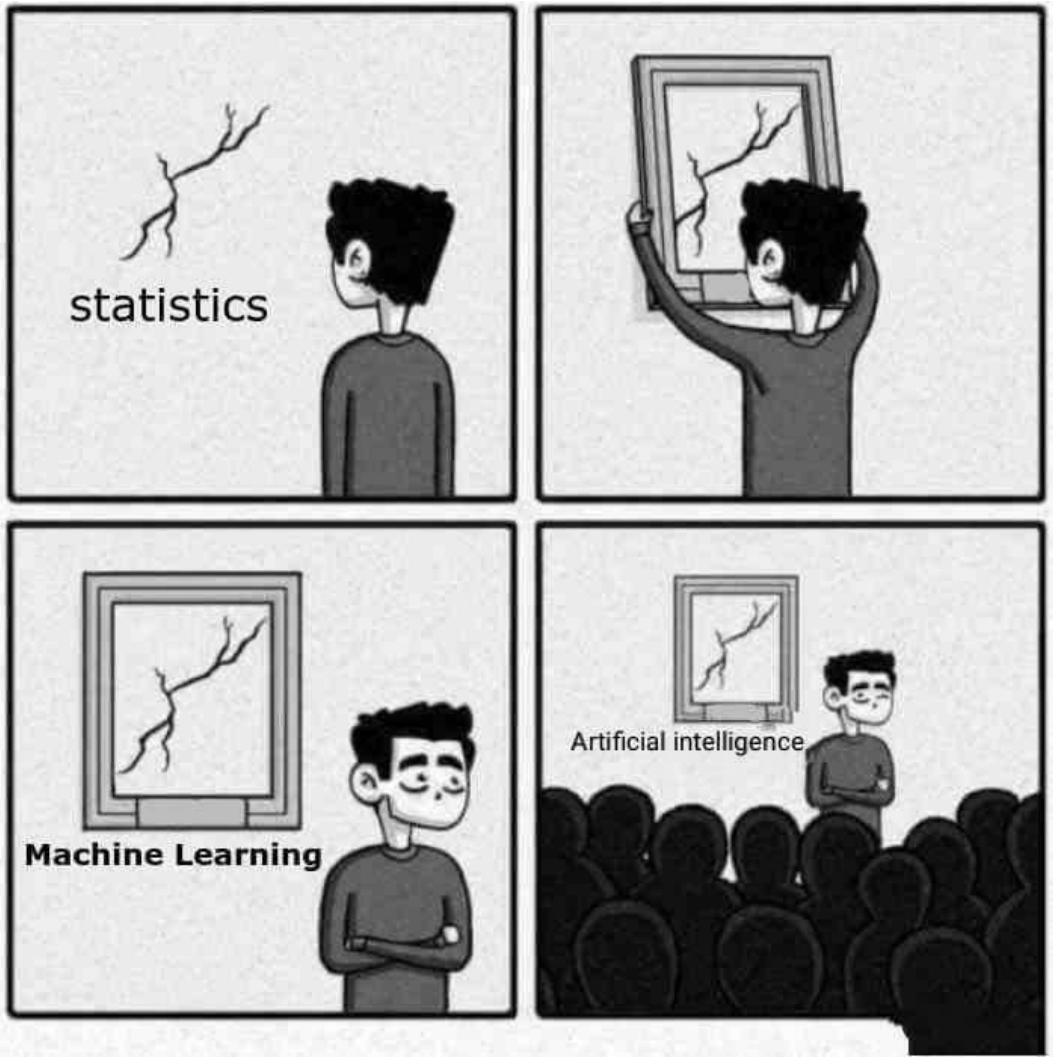
- **Volume** - Amount of data - Bytes - KB - MB - GB - TB - PB - EB - ZB - Till 2020 (20 ZB)
- **Velocity** - Different Source generate data every day
- **Variety** - Different format
- **Veracity** - Uncertainty of data
- **Value** - Usefulness

Is Statistics is essential for Machine Learning?



VIT
lore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Statistics Vs Machine Learning - MEME Explanation



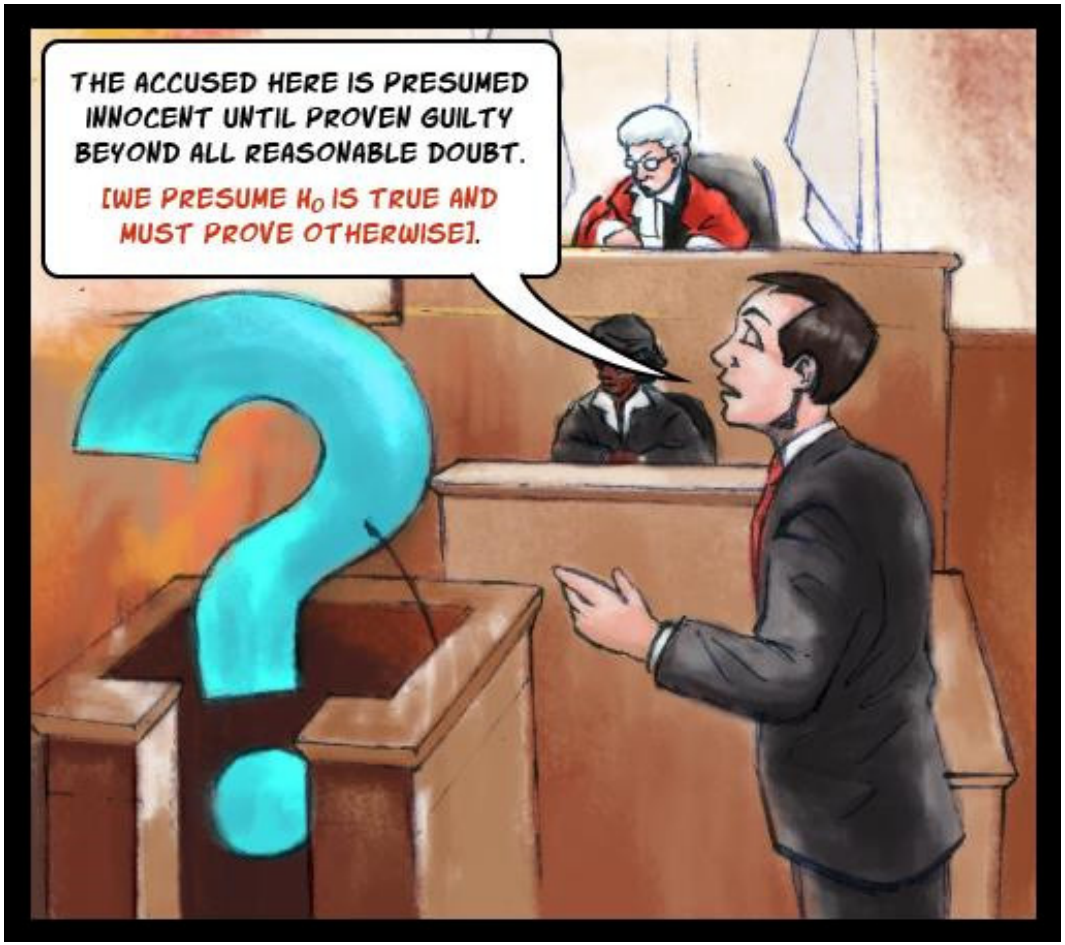
Importance of Statistics

- Dataset + Problem = Some Pattern (Algorithm!?)
- Dataset + Machine Learning (Library) = 'Something' (Outcome)

Basics of Statistical Terms / Data Exploration

- Central Tendency (Mean, Median, Mode)
- Population & Sample
- Measure of Spread (Range, Quartile, Variance & Standard Deviation)
- Descriptive statistics (IQR, Skewness & Kurtosis)
- Inferential statistics (Point of estimation - Confidence intervel & Margin of Error)
- Types of data (Categorical & Numerical)
- Matplotlib & Seaborn Visualization
- Covariance - Correlation
- Central Limit theorem
- Hypothesis Testing

Hypothesis Testing



Hypothesis Testing

- What is Hypothesis Testing?
- Why do we use it?
- What are the basics of Hypothesis?
- Which are important parameters of Hypothesis testing?
- What are the testing process?

What is Hypothesis Testing?

- Statistical decisions using data
- Ex1: In a particular class, girls are taller than boys
- Ex2: Avg marks of this subject in this particular class is 42.5 %
- In Ex1 How we can predict girls are taller?
- In Ex2 How we got 42.5% ?
- Mathematical conclusion - Statistics - True!

Why do we use Hypothesis Testing?

- Essential procedure?!
- Evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data
- Ex: Subjects in a semester - Pass / Fail scenario

What are the basics of Hypothesis?

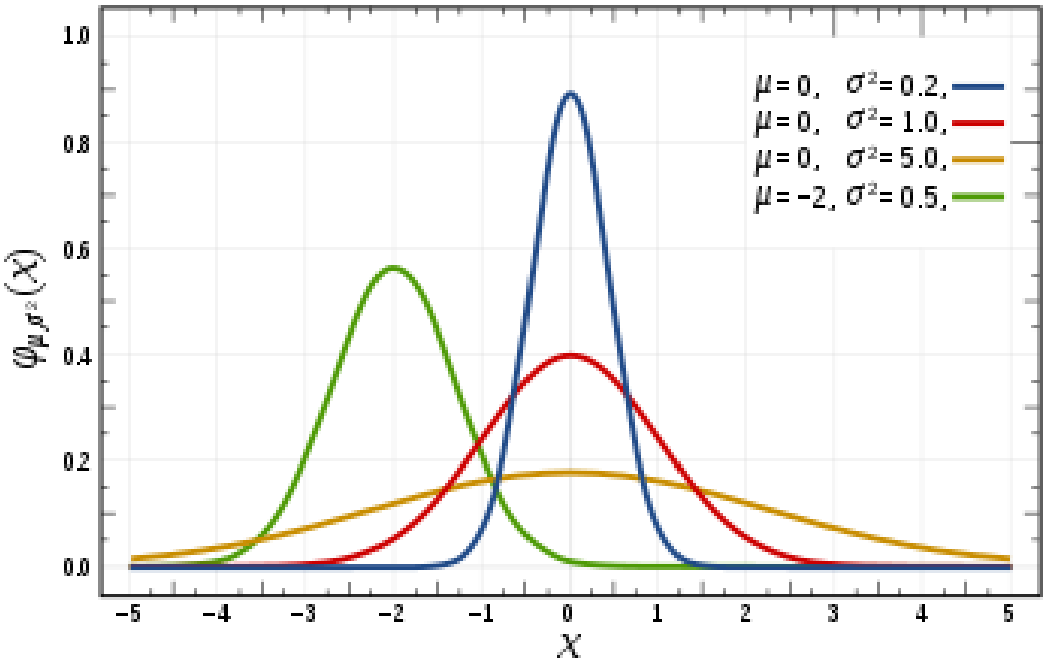


Figure: Normal Curve images with different mean and variance

What are the basics of Hypothesis?

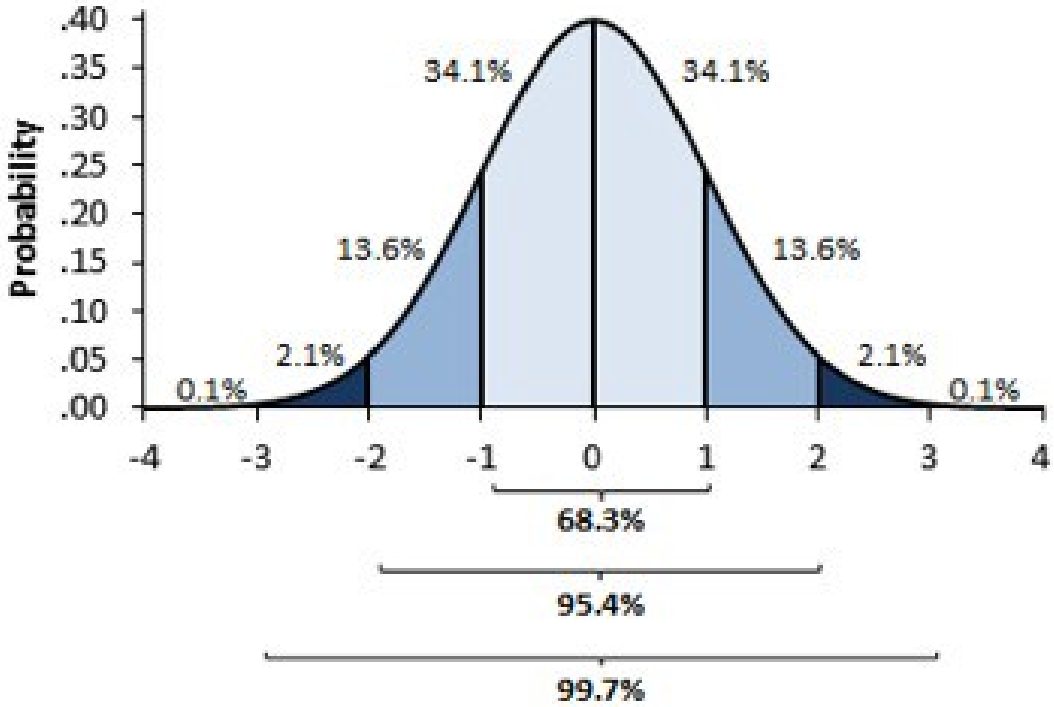


Figure: Standardised Normal curve image and separation on data in percentage in each section

What are the basics of Hypothesis?

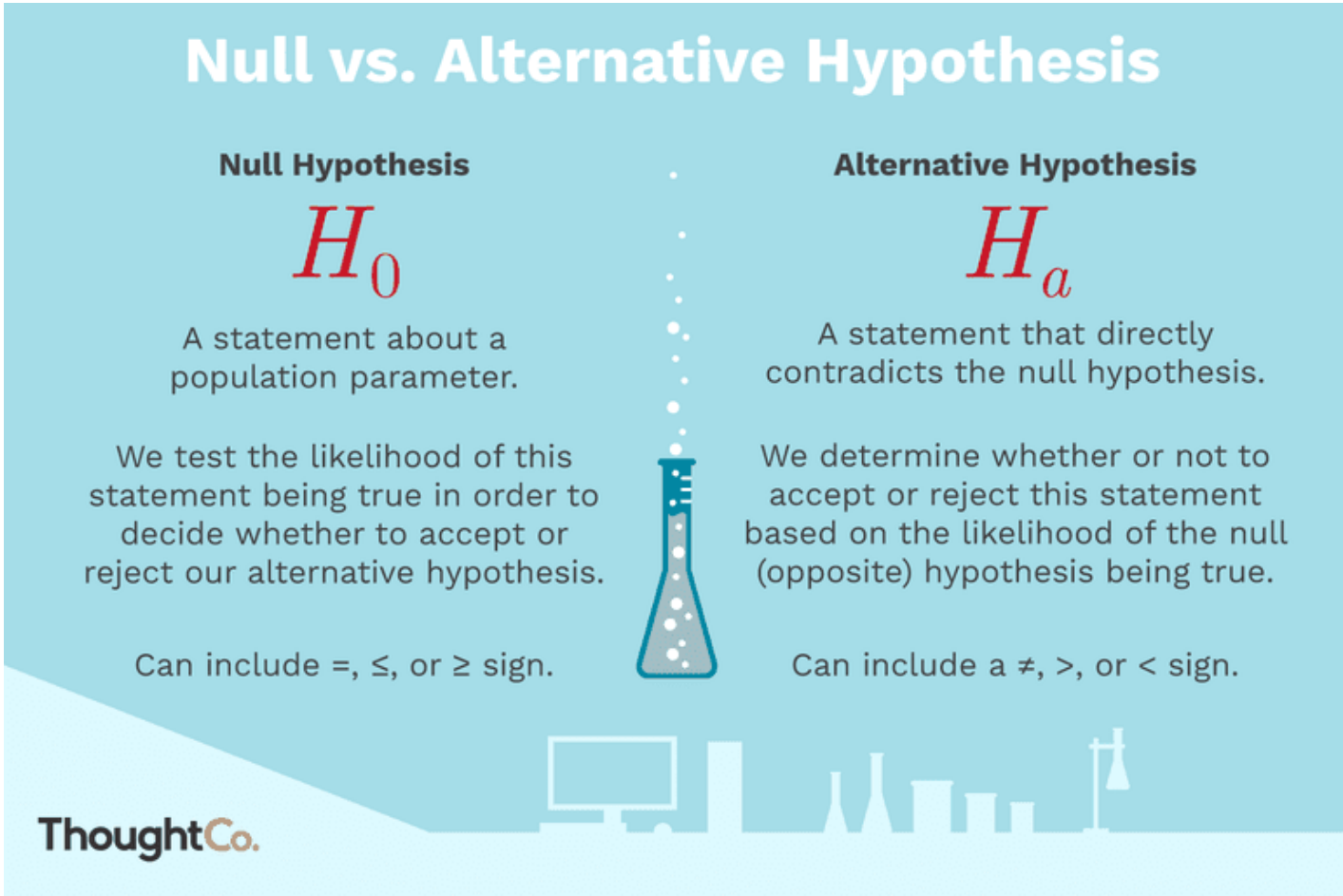
- Normalization (Mean = Median = Mode) / Normal curve

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Standard Normalization (Mean = 0, Std Deviation = 1)

$$x_{new} = \frac{x - \mu}{\sigma}$$

Which are important parameter of Hypothesis testing?



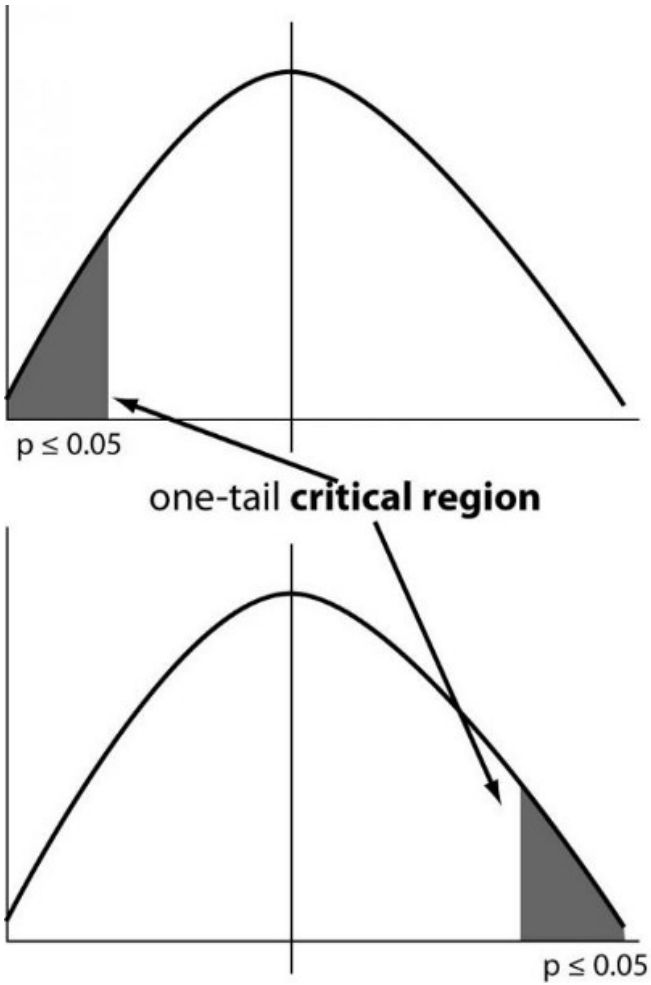
Null Vs Alternate Hypothesis

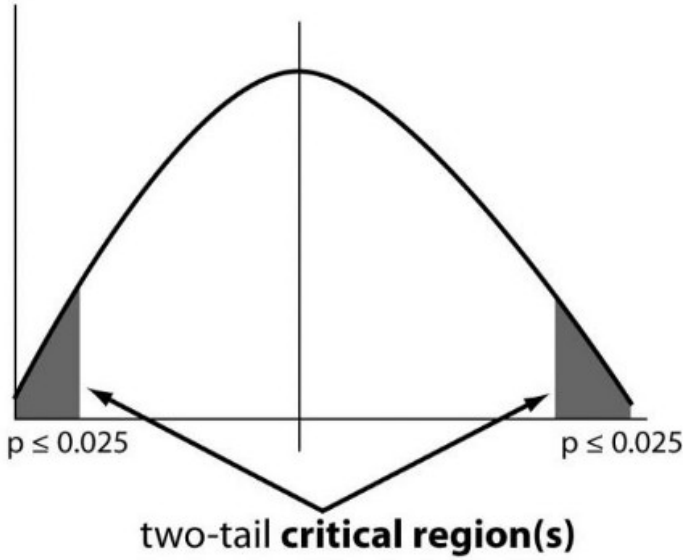
- **Null Hypothesis** - general statement - no association among groups
- Ex: ABC Bicycle company production = 500 unit / day
- **Alternate Hypothesis** - Contrary to the null hypothesis - Observations are the result of a real effect
- Ex: ABC Bicycle company production != 500 unit / day

Null and Alternate Hypothesis

- **Level of significance**- Degree of significance in which we accept or reject the null-hypothesis
- 100% is not possible - 5% significance is considered (Alpha - 0.05) - Output should be 95% in each sample
- **Type I Error** - When we reject the null hypothesis, although that hypothesis was true - Alpha (Critical region)
- **Type II Error** - When we accept the null hypothesis but it is false - Beta (Acceptance region)
- **One Tailed test** - Region of rejection is on only one side of the sampling distribution
- **Two Tailed test** - critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values

One Tailed Vs Two Tailed test





P value

- Probability of finding the observed, results when the null hypothesis of a study question is true
- P-value is less than the chosen significance level then you reject the null hypothesis
- Ex: **Coin is fair (H0) or tricky (H1)** (alpha = 0.05)
  1st (T) = 50 %
  2nd (T) = 50/2 = 25 %
  3rd (T) = 25/2 = 12.5%
  4th (T) = 12.5/2 = 6.25 %
  5th (T) = 6.25/2 = 3.125 %
  6th (T) = 3.125/2 = 1.5625 %

Degree of freedom

- Maximum number of logically independent values, which are values that have the freedom to vary, in the data sample
- Commonly discussed in relation to various forms of hypothesis testing
- Ex: Consider 5 no. - no relation between them (3, 8, 5 & 4, ?) and mean will be 6 - 5th = 10 - DoF will be 4

Hypothesis testing type

- T-test (Student Test)
- Z-test
- ANOVA-test (F-test)
- Chi-Squared-test

AI Vs ML Vs DL Vs DS

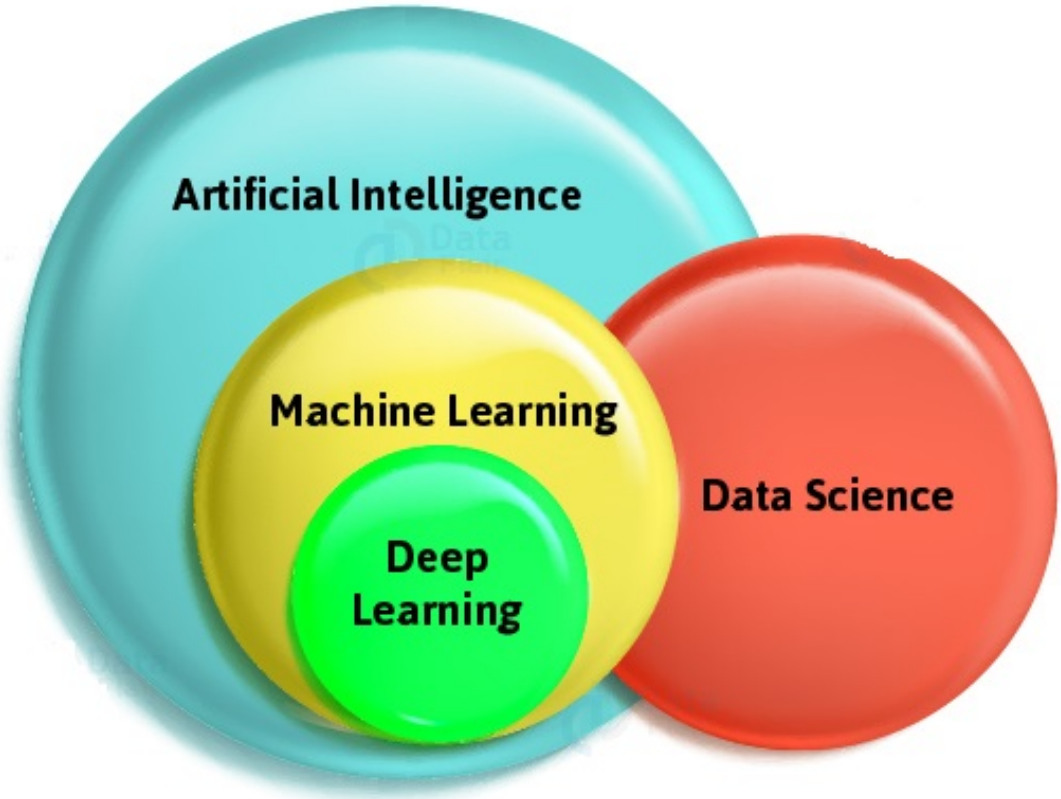


Figure: Conceptual understanding

Machine Learning - Intro

- **General Intro** Machine Learning, means it can access the data and use it to learn for themselves without any programming.
  "Machine Learning is the field of study that gives computers, the ability to learn without being explicitly programmed. — **Arthur Samuel, 1959**"
- **Engineering Intro** - A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. — **Tom Mitchell, 1997**
- **Layman Intro** - Child learning of alphabet
  "A baby learns to crawl, walk and then run. We are in the crawling stage when it comes to applying machine learning. -**Dave Waters**"

History of Machine Learning

Real time examples for Machine Learning

- Facial Recognition
- Virtual Reality headsets - MIT
- Speech to text (Iphone users)
- Robo dog - Spot! - Spot - Dance for Uptown! (Reinforcement Learning)
- Amazon, Flipkart, Netflix, Audible (E-Commerce) (Recommender System)
- Medical Images and Signals (ECG, PPG, EEG, EMG...) (Time Series)

Machine Learning - MEME!

Interviewer: What's your biggest strength?

Me: I'm an expert in machine learning.

Interviewer: What's 9 + 10?

Me: Its 3.

Interviewer: Not even close. It's 19.

Me: It's 16.

Interviewer: Wrong. Its still 19.

Me: It's 18.

Interviewer: No, it's 19.

Me: it's 19.

Interviewer: You're hired

Machine Learning - MEME!

Albert Einstein: Insanity Is Doing the Same Thing Over and Over Again and Expecting Different Results

Machine learning:

Types of Machine Learning

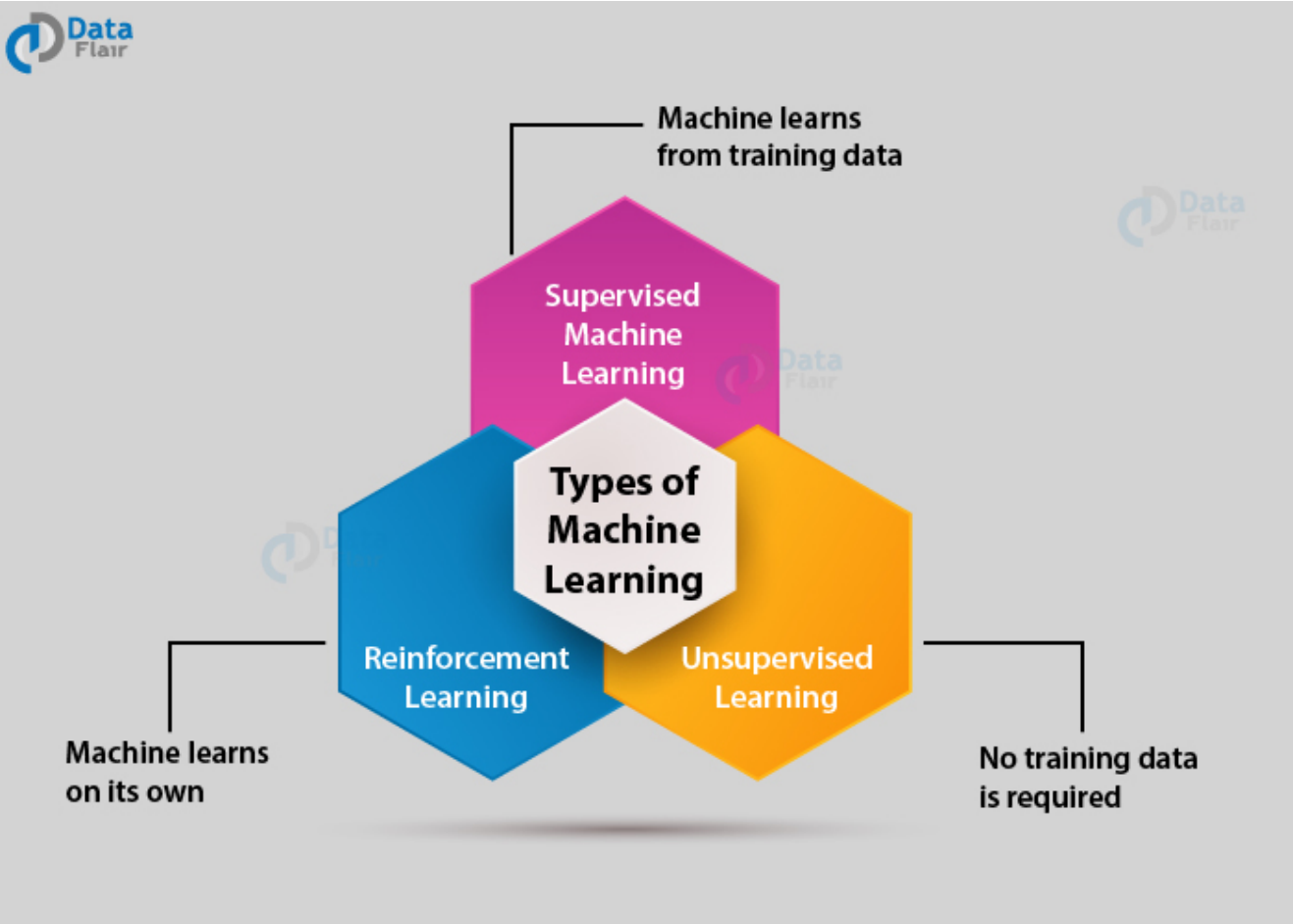


Figure: Broad classification of ML

Types of Machine Learning (Cont...)

- **Supervised Machine Learning** - Train me!
  - Classification
  - Regression
- **Unsupervised Machine Learning** - I am self sufficient in learning!
  - Clustering
  - Dimentionality Reduction
  - Association Rule
- **Reinforcement Learning** - My life My rules!

Algorithms for each types of Machine Learning

Supervised Machine Learning - Classification

- Support Vector Machine (Linear SVM)
- Kernel Support Vector Machine (Non-linear SVM)
- K-Nearest Neighbor (KNN)
- Logistic Regression
- Decision Tree classification
- Random Forest classification
- Naive Bayes classifier & many more...

Algorithms for each types of Machine Learning (Cont...)

Supervised Machine Learning - Regression

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Lasso Regression
- Ridge Regression
- Elastic Net Regression & many more...

Algorithms for each types of Machine Learning (Cont...)

Unsupervised Machine Learning - Clustering

- K-Means clustering
- Hierarchical clustering & many more...

Unsupervised Machine Learning - Association Rule Learning

- Apriori
- Eclat & many more...

Unsupervised Machine Learning - Dimentionality Reduction

- Principal Component Analysis(PCA)
- Linear Discriminant Analysis(LDA)
- Kernel PCA & many more...

Reinforcement Learning

- Upper Confidence Bound
- Thompson Sampling

Preferable languages used for Machine Learning

Table: Tug of war between languages

Python	**R**	**Julia**
General purpose	Statistical analysis	Scientific computing
Good	Good	**speed** & **performance**
Huge community	Huge community	small community
200k libraries	15k libraries	3k libraries
In Billions	In Billions	13M downloads
-	-	**Compile just in time**
Jupyter, Pycharm	R Studio	Juno IDE
ijulia	-	-

So, Why Python ?

- Open source
- Platform Independence
- User Friendly and Easy to learn
- Vast community support
- Good Visualization options
- A great library ecosystem like Pandas, scikit.learn, numpy, scipy, keras, Tensorflow, Matplotlib, NLTK, scikit-image, PyBrain, Caffe, StatsModels
- Capability of interacting with almost all the third party languages and platforms.

Different IDE's for Machine Learning

- Pycharm
- Anaconda (Distribution) - Jupyter notebook, Spyder
- Colaboratory - Google
- Atom
- Sublime text 3
- IPython
- Visual Studio Code
- Rodeo

Important libraries for Machine Learning

- **Numpy** - Scientific computing
- **Pandas** - Data analysis and manipulation
- **Scipy** - Maths, Science, Engineering
- **spaCy** - NLP
- **Matplotlib** - Data Visulaization
- **Seaborn** - Statistical Visualization
- **Bokeh** - High end Visualisation
- **Scikit-learn** -Machine Learning library
- **StatsModels** - Statistics
- **SymPy** - Symbolic maths, computer algebra system
- **Keras** - Neural Netowork
- **Tensorflow** - Fast numerical computing

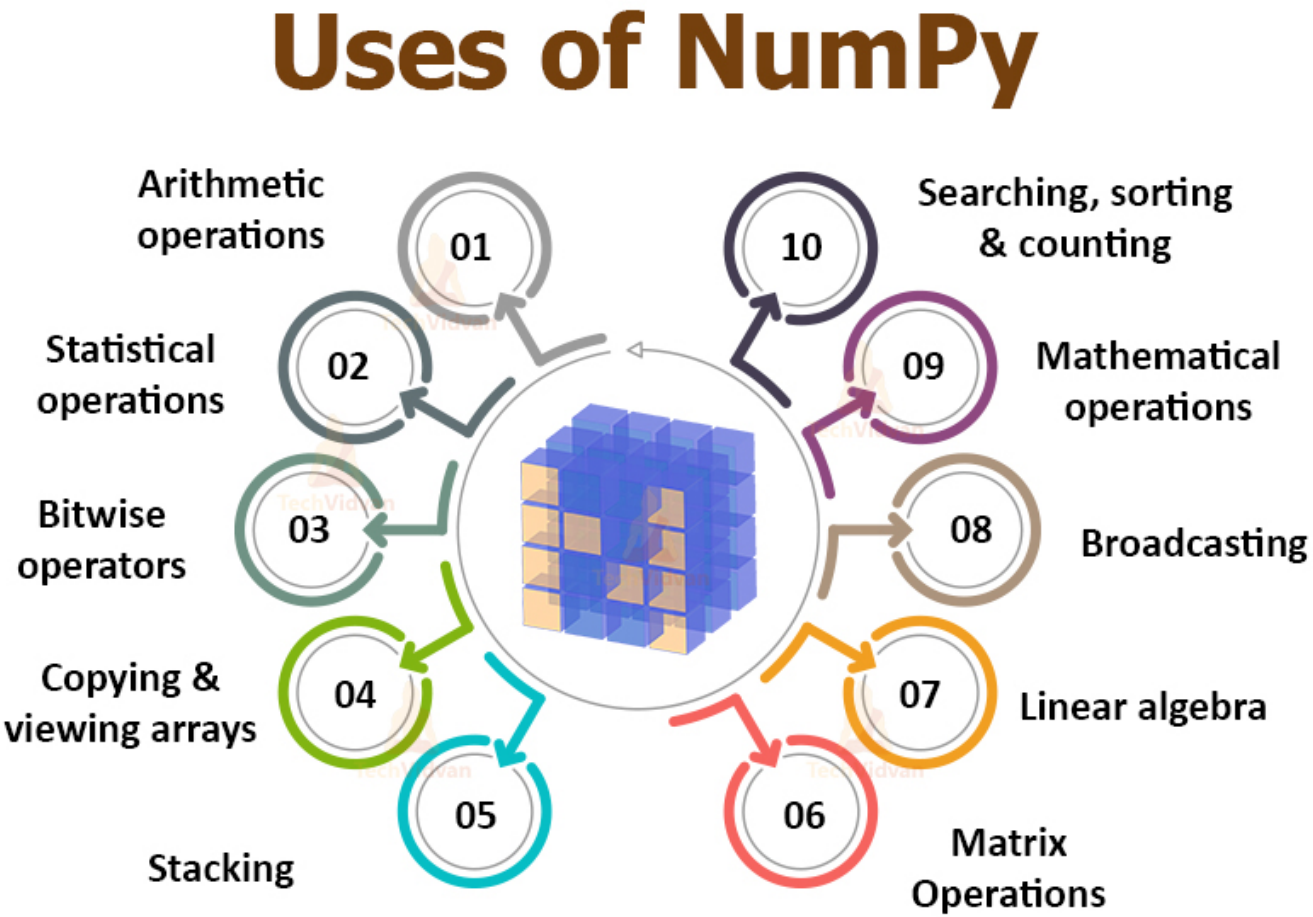


Figure: Usage of Numpy

Pandas - Intro

- Loading and Saving datasets
- Column insertion and deletion
- Data selection
- Column and Row renaming
- Row deletion
- Data sorting
- Handling missing values
- Handling duplicate datas
- Data Exploration
- Data Visualization

Scikit-learn - Intro

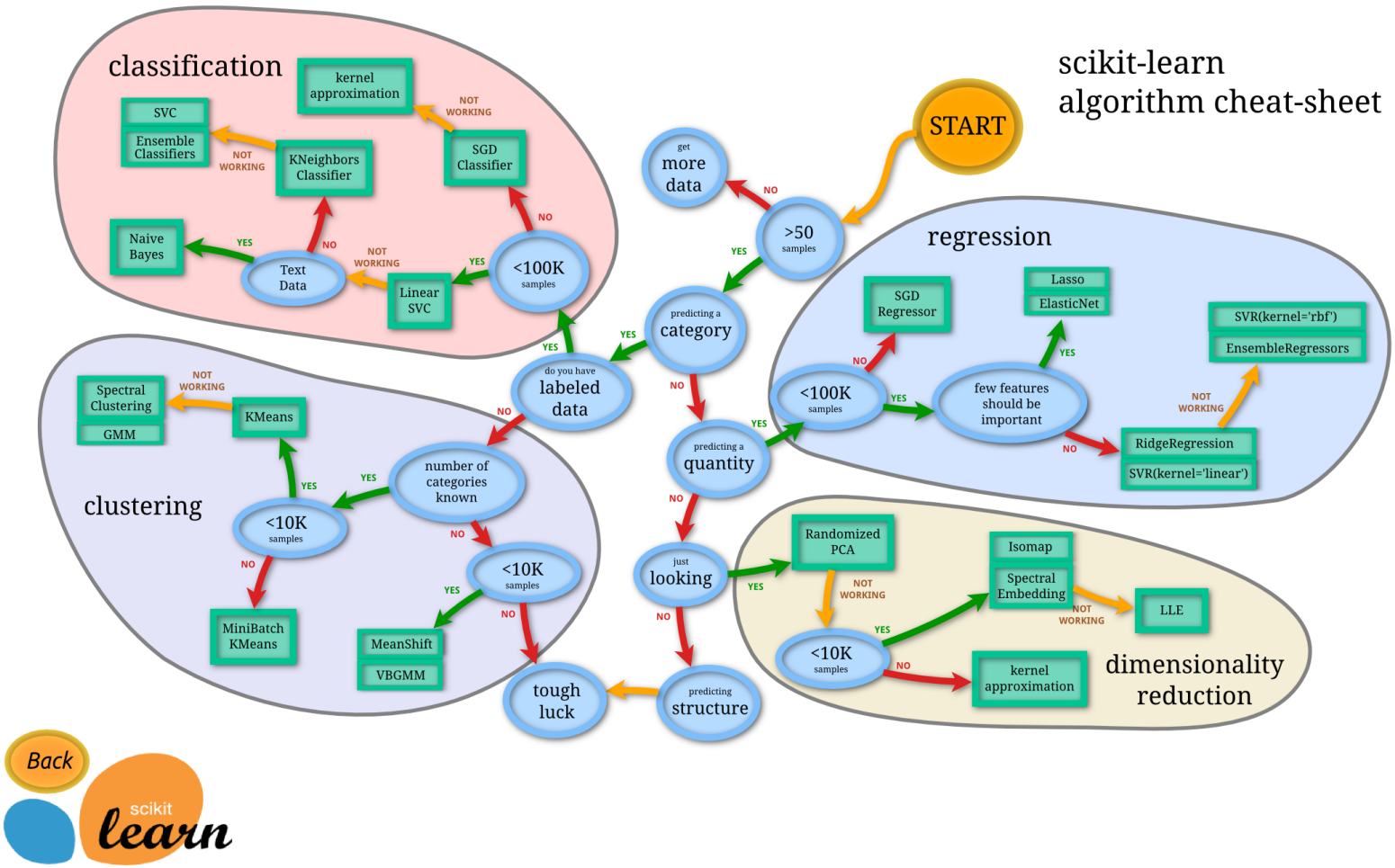


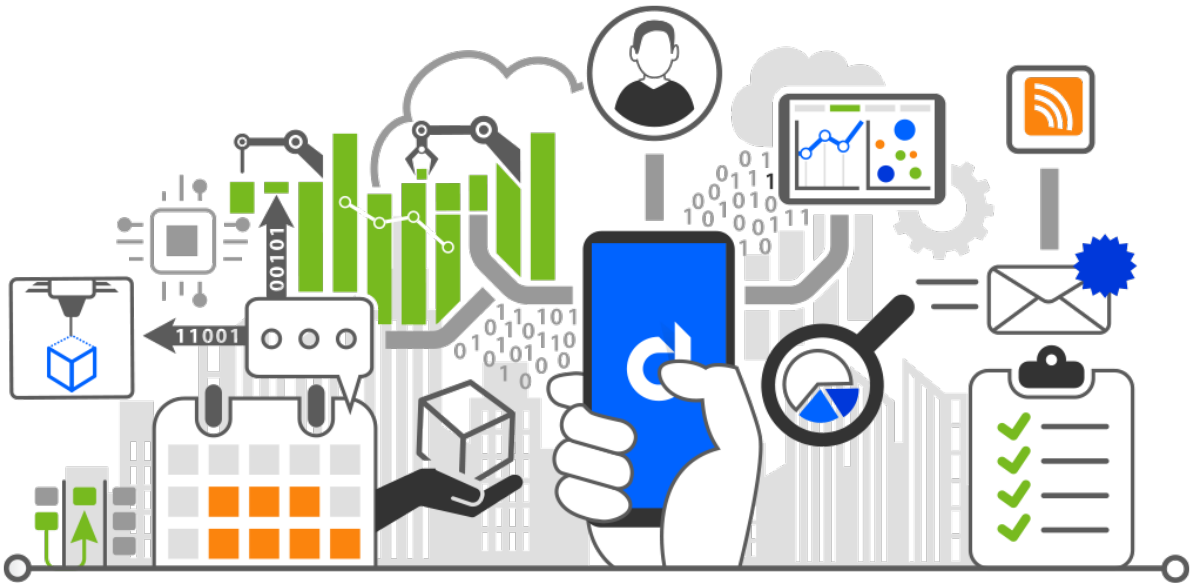
Figure: scikit-learn API reference

Steps for Machine Learning

- Data Collection or Acquisition
- Importing Libraries
- Loading Datasets
- Pre-processing and Exploratory Data Analysis (EDA)
- Splitting of Datasets
- Feature Scaling
- Feature Selection
- Dimensionality Reduction
- Modelling
- Metrics

Need an interesting read, Glimpse to Machine Learning- Hit me!

Step 1: Data Collection or Acquisation



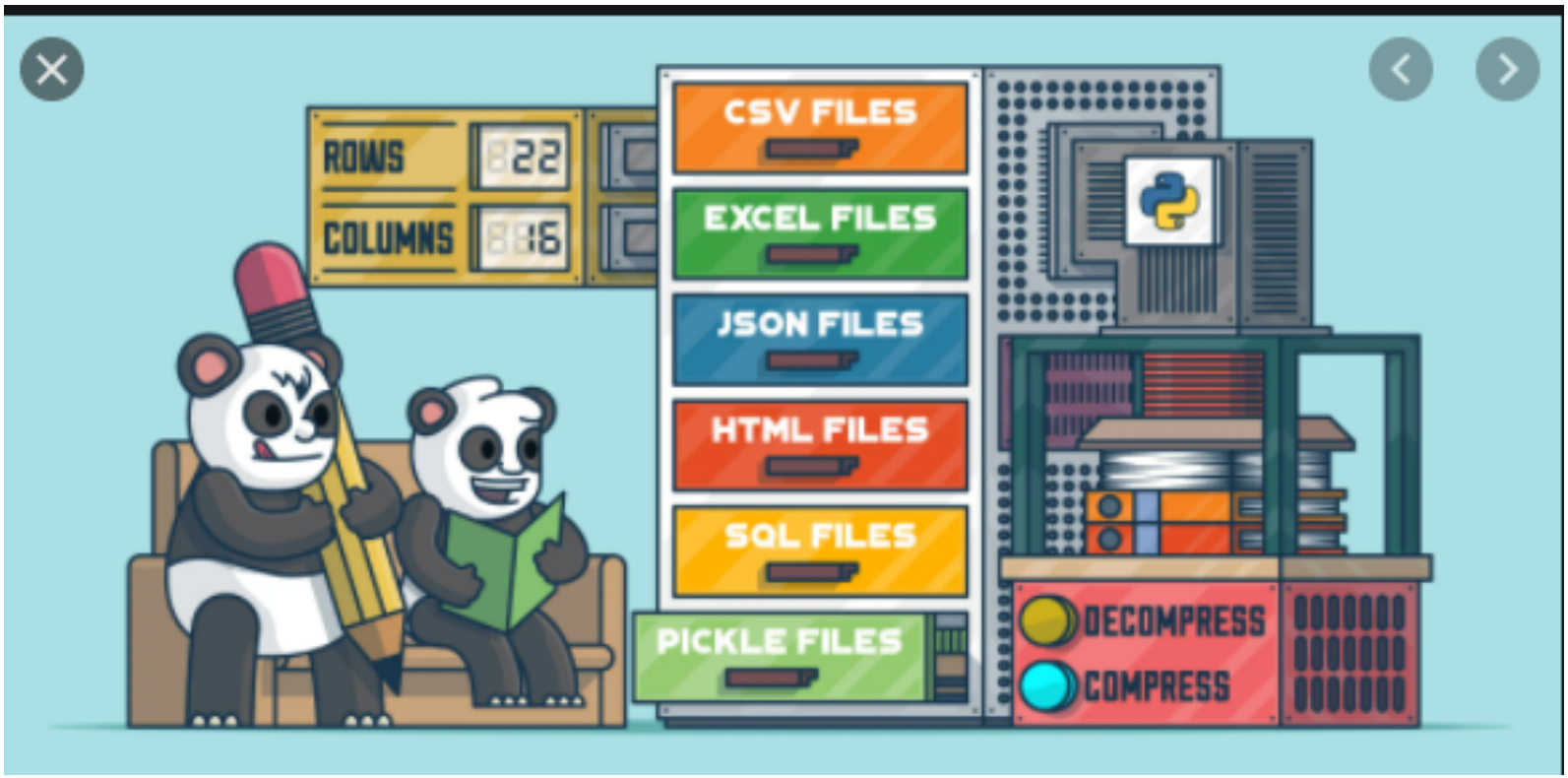
- By using Sensors, Medical devices like ECG, PPG...
- Google dataset search - link
- UCI Machine Learning Repository - link
- CMU libraries - link
- OpenML - link
- Fivethirtyeight - link
- Physionet - link
- Kaggle datasets - link
- Data.gov - link
- Academic torrents - link
- Awesome dataset by github - link

Step 2: Importing libraries

- Either installing or importing pacakages
- Through PIP install! - Through conda install

Step 3: Loading Datasets

- .csv, .json, .xlsx, .xml, .docx, .txt, .pdf, .png, .jpg, .mp3, .mp4



Step 4: Pre-Processing & Exploratory Data Analysis

- 60 - 70 work in this step
- Data information (.head, .tail, .shape, .columns)
- Overall data type (.info)
- Understanding basic statistics of data (.describe)
- Target details (.unique, .valuecounts)
- Checking for missing values (.isnull.sum)
- Solution for missing values (SimpleImputer)
- Outlier detection (Univariate: Box-plot, Grubbs test, Multivariate: PCA, Mahalanobis Distance, Cook's Distance...)
- Skewness (log transformation, square root transformation, box-cox transformation) and Kurtosis of data
- Correlation between features (.corr)
- Dependent and Independent variables
- Encoding categorical data for both dependent and independent variables (label encoder, OneHotEncoder, pd.getdummies)

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Step 5: Splitting of datasets

Step 6: Feature Scaling

- Normalization Vs Standardization
- Normalization (z score, min-max, scaling to unit length, logarithmic scale) and Standardization
- Feature transformation (Scaling (Minmax scaler, standardscaler, normalizer, robustscaler), Discretization, Binning

Step 7: Feature Selection

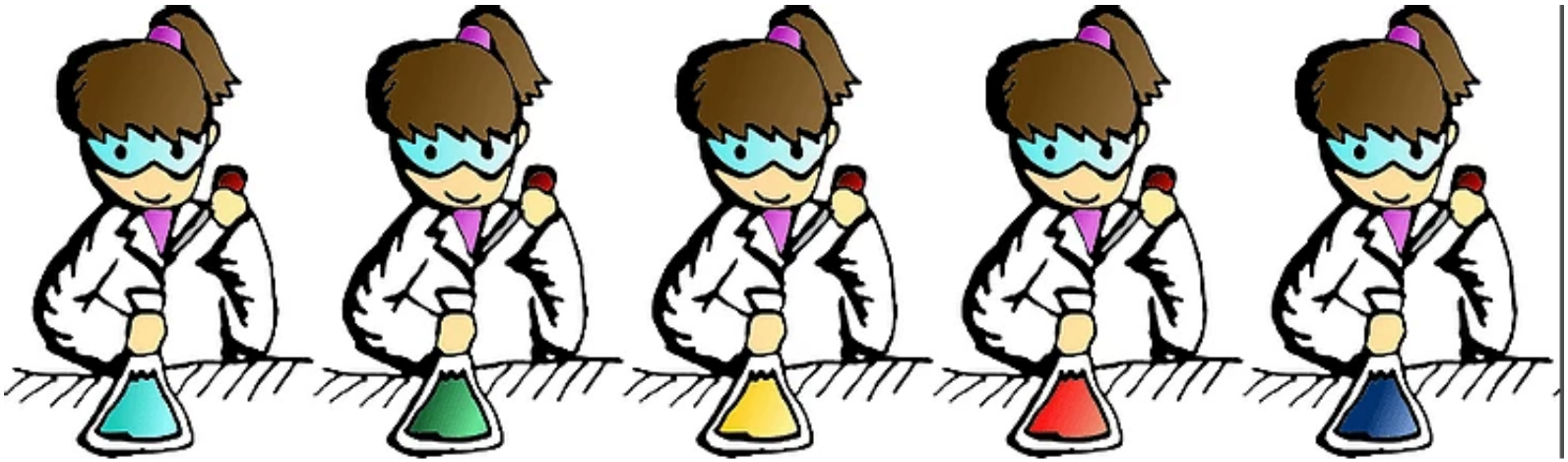
If you put garbage in, you will only get garbage to come out

- Random forest classifier, chi-2, select from model, variance threshold, correlation threshold, Pearson's correlation (heatmap), chi squared, ANOVA f value, maximal information coefficient (MIC)
- wrapper based – forward search, backward search, recursive feature elimination (RFE)(rfe.support, rfe.ranking)
- sequential feature selector (SFB, SBS, SFFS, SBFS)
- Embedded methods – lasso regularization in linear regression, select k best in random forest, Gradient Boosting Machine (GBM) (univariate and multi variate feature selection)

Step 8: Dimentionality Reduction

- What are Dimensionality Reduction Techniques?
- The Curse of Dimensionality
- Importance of Dimentionality Reduction
- Feature Selection
- Feature Extraction

Step 9: Modelling



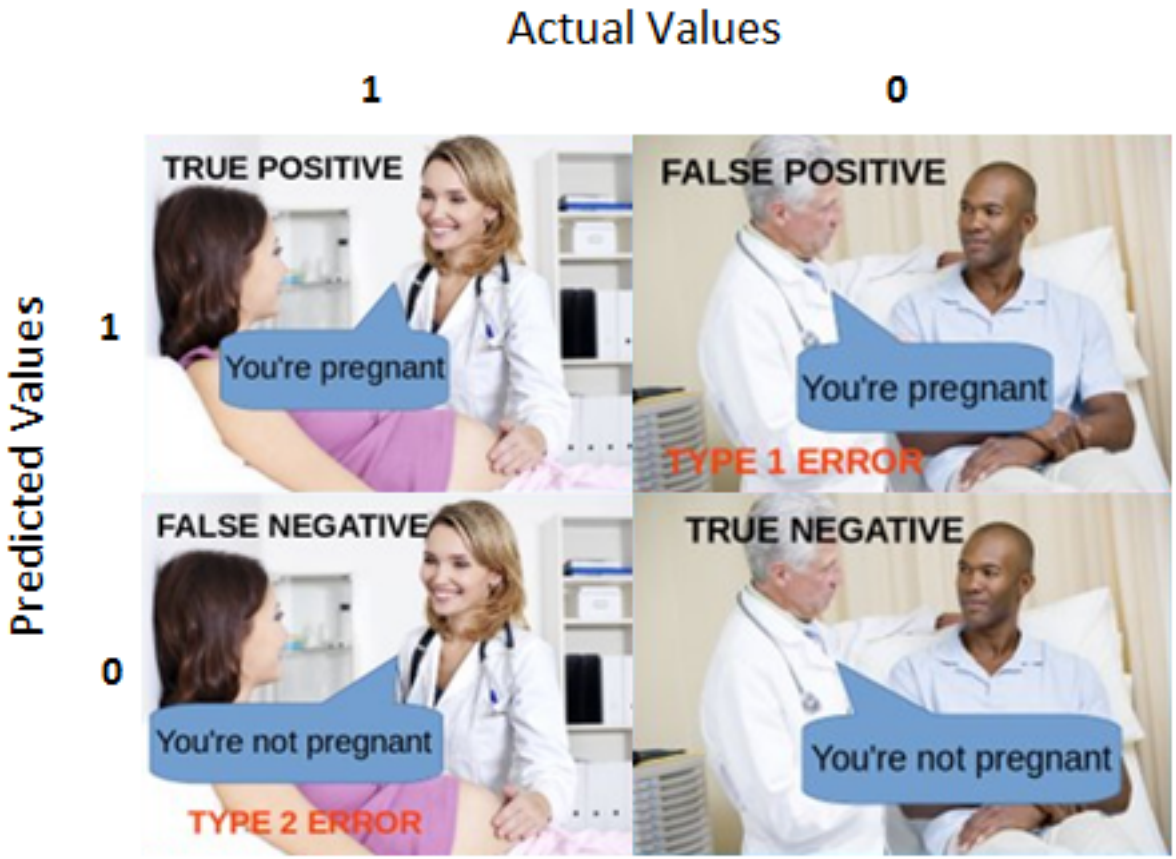
- Regression / Classification algorithms
- Fit & Predict
- Hyper Parameter!
- Grid search
- Cross validation

Step 10: Metrics

Metrics - Regression

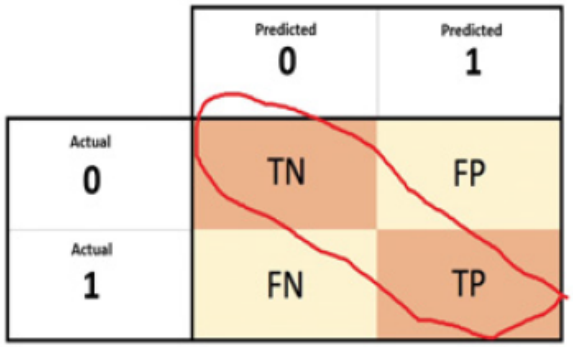
S No	Term	Criterion
1	R-Squared	High
2	Adj R-squared	High
3	F-Statistics	High
4	Std.Error	Close to zero
5	t-Statistics	>1.96 <0.05
6	AIC (Akaike Info Crit)	Low
7	BIC (Bayesian)	Low
8	Mallows cp	Should be close to no of target
9	MAPE (Mean Abs Per Err)	Low
10	MSE (Mean Squ Err)	Low
11	MPE (Mean Per Err)	Low
12	Min-Max Acc	High

Metrics - Classification - Confusion matrix 1

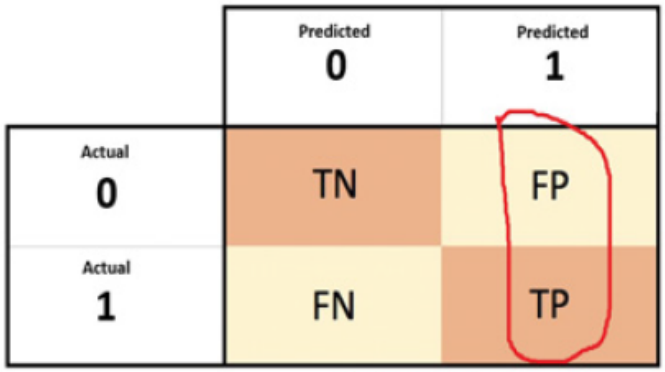


Metrics - Classification - Confusion matrix 2

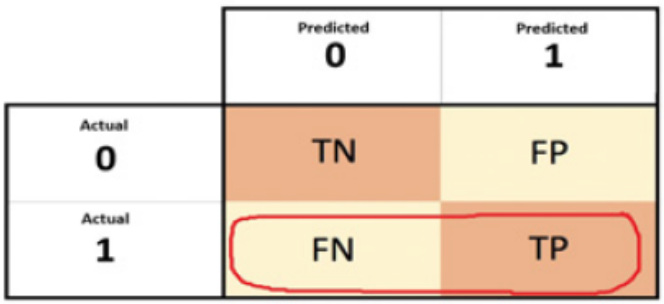
- f1 score = 2*((precision*recall)/(precision+recall))



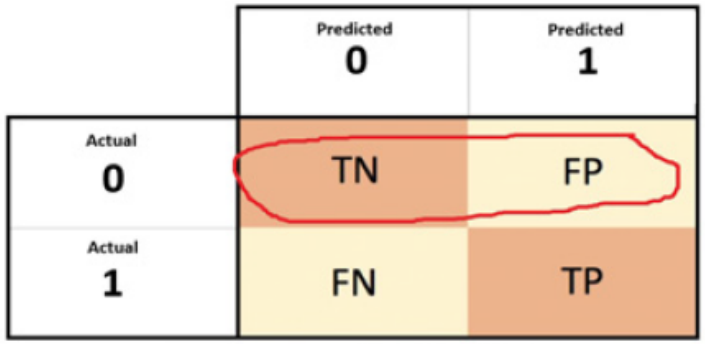
$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$



$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$



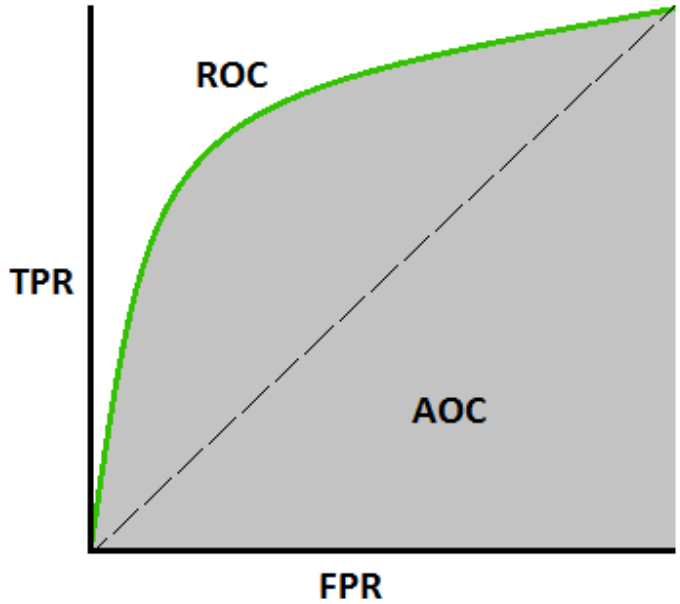
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$



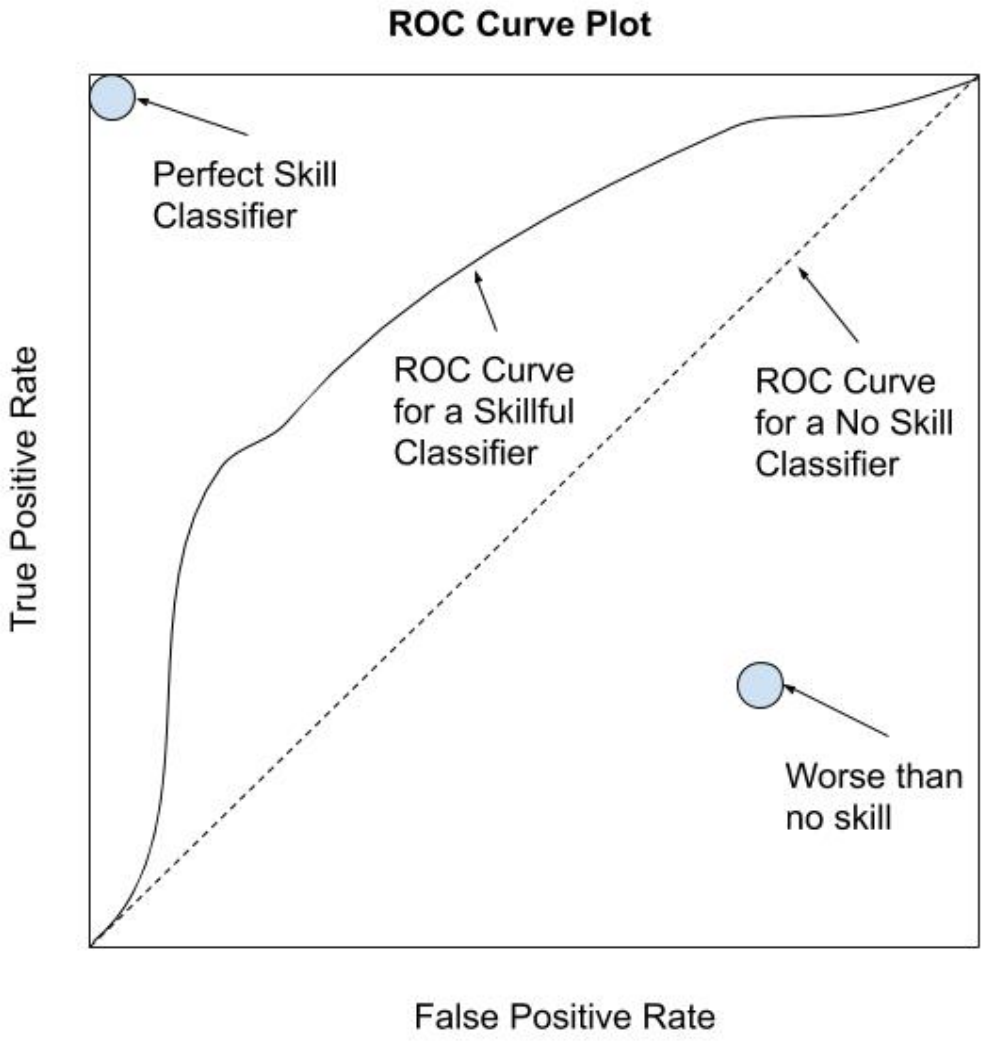
$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Metrics - Classification - AUC-ROC

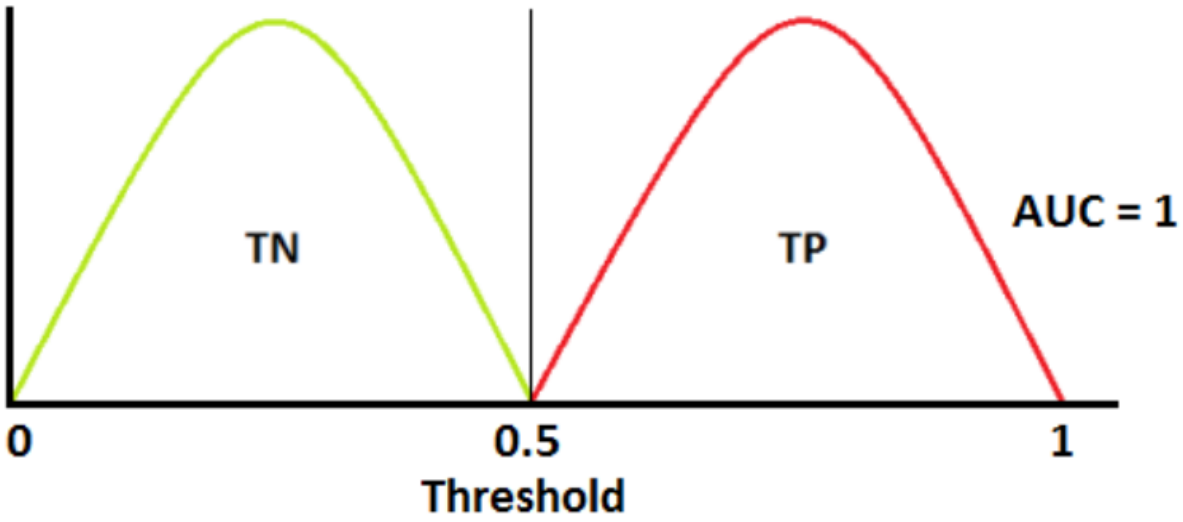
- ROC - Probability curve
- AUC - Degree or Measure of separability
- TPR / Recall
- Specificity
- FPR = 1 - Specificity
- Sensitivity, Specificity, Threshold, TPR, FPR

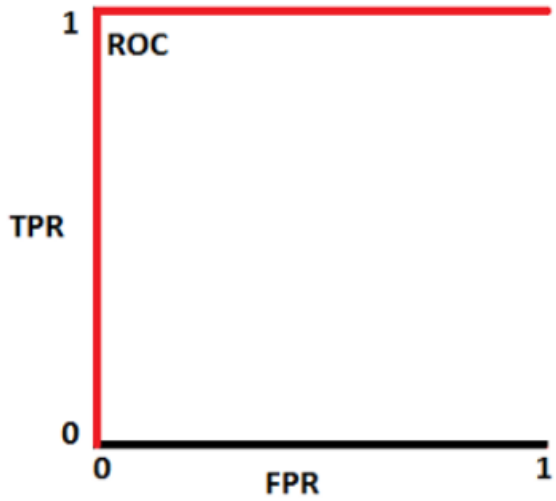


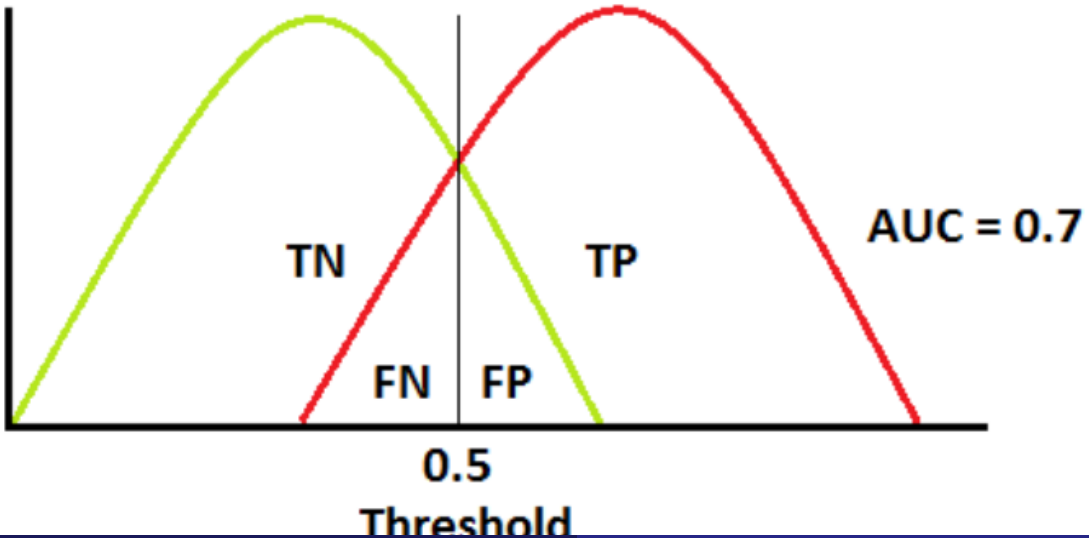
Metrics - Classification - ROC

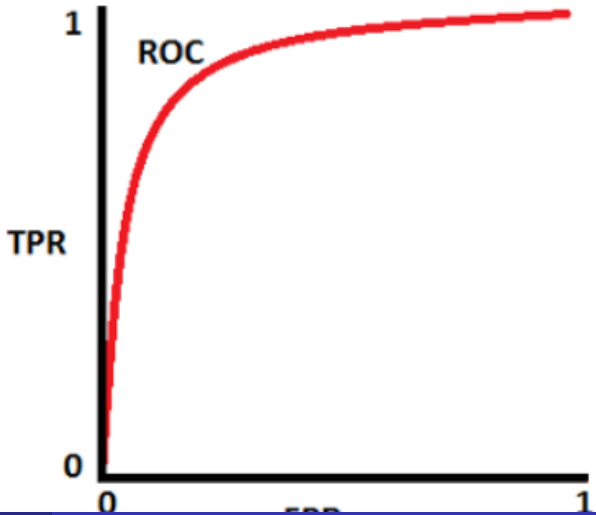


Metrics - Classification - AUC-ROC

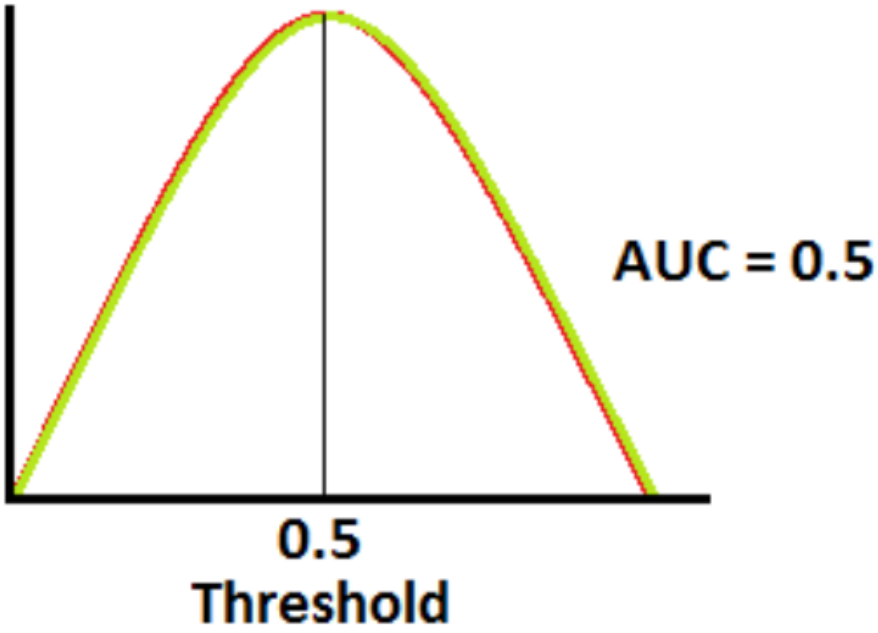


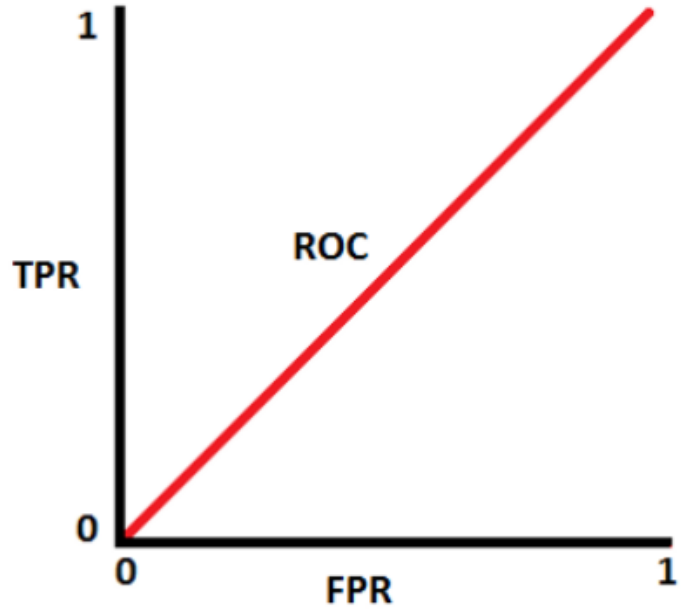


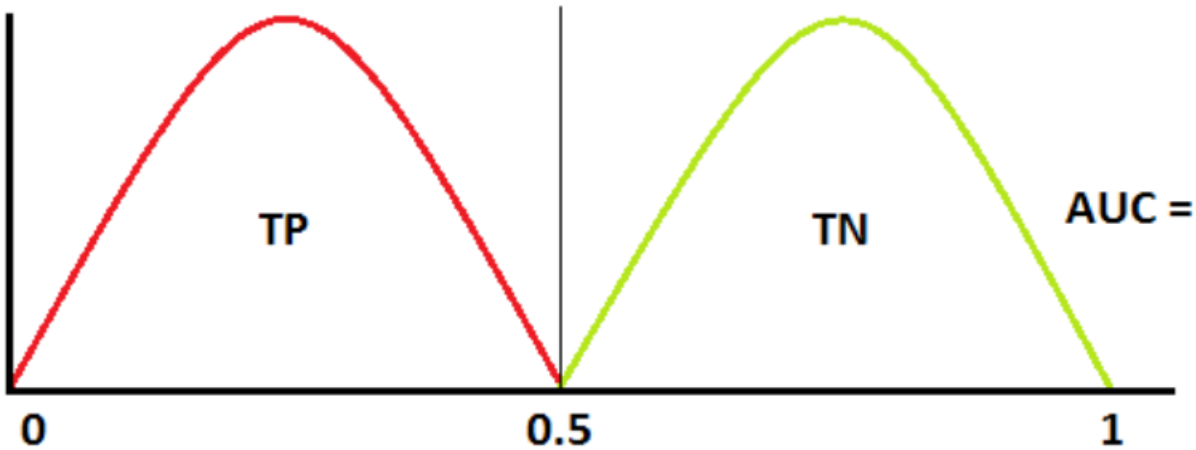


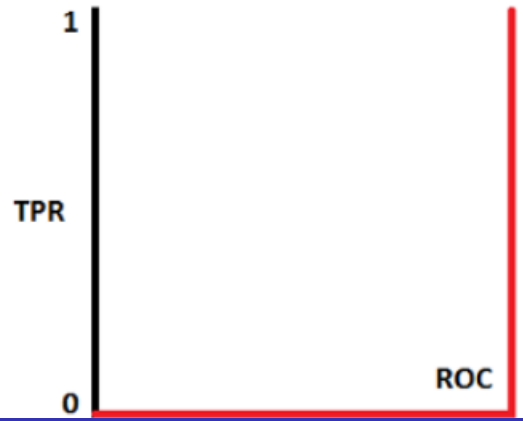


Metrics - Classification - AUC-ROC

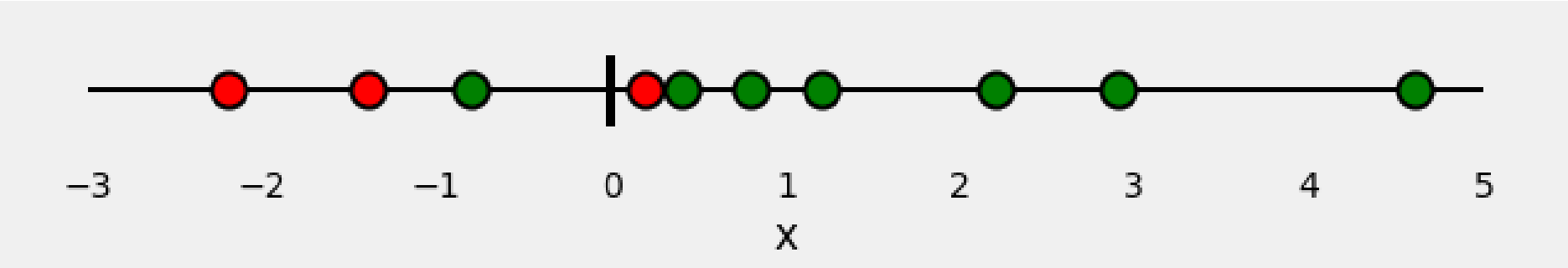


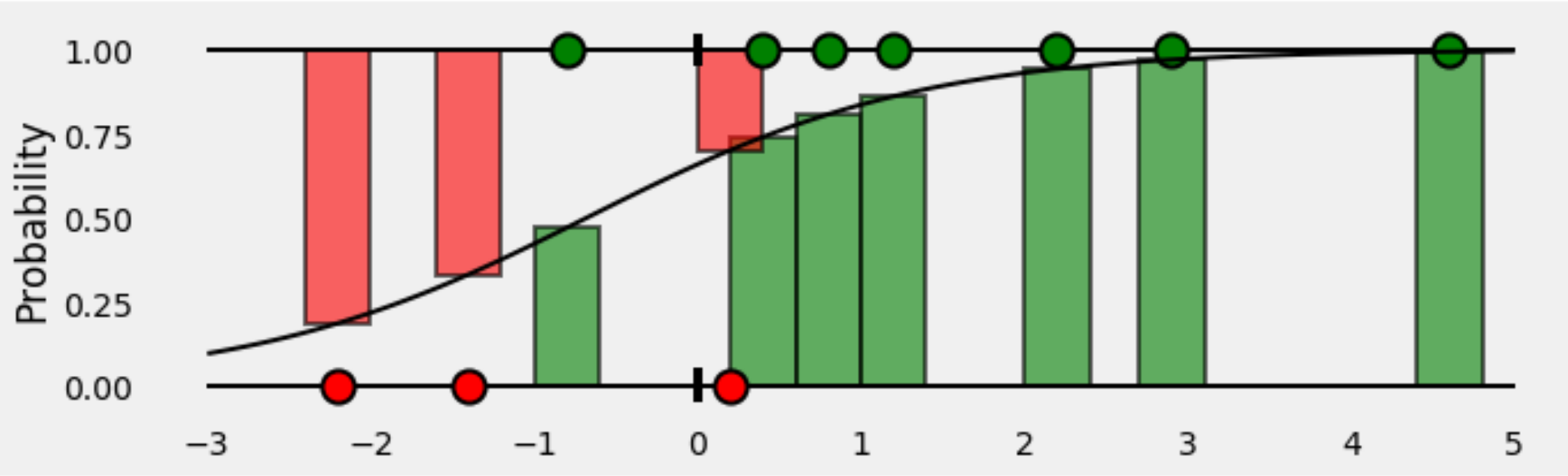






Metrics - Classification - Log Loss





Algorithm Explanation - Support Vector Machine

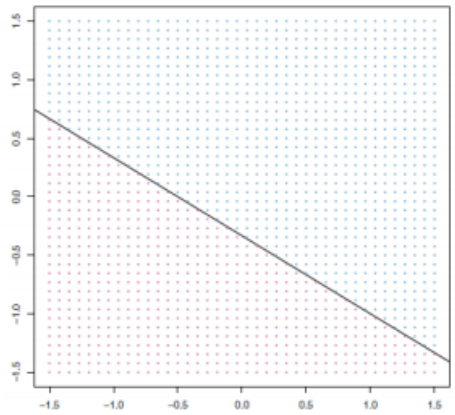
Support Vector Machine

- Supervised learning
- Classification & Regression analysis
- The goal of the SVM algorithm is to create the best decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- **Out-of-the-box** classifier
- For better understanding, we go by
  - Maximal Margin Classifier
  - Support Vector Classifier
  - Support Vector Machine

SVM - Maximal Margin Classifier

1. One Dimensional space

2. Two Dimensional space



3. Three Dimensional space

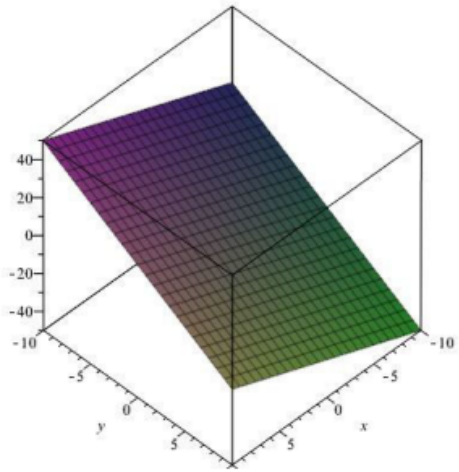


Figure: 1D Vs 2D Vs 3D Space

SVM - Maximal Margin Classifier (Contd...)

X1	X2	Category
60	82	Pass
20	42	Fail
...	...	...
91	72	Pass

Figure: Consider a dataset with 2 independent features (X1, X2) and 1 class or dependent feature (Category)

SVM - Maximal Margin Classifier (Contd...)

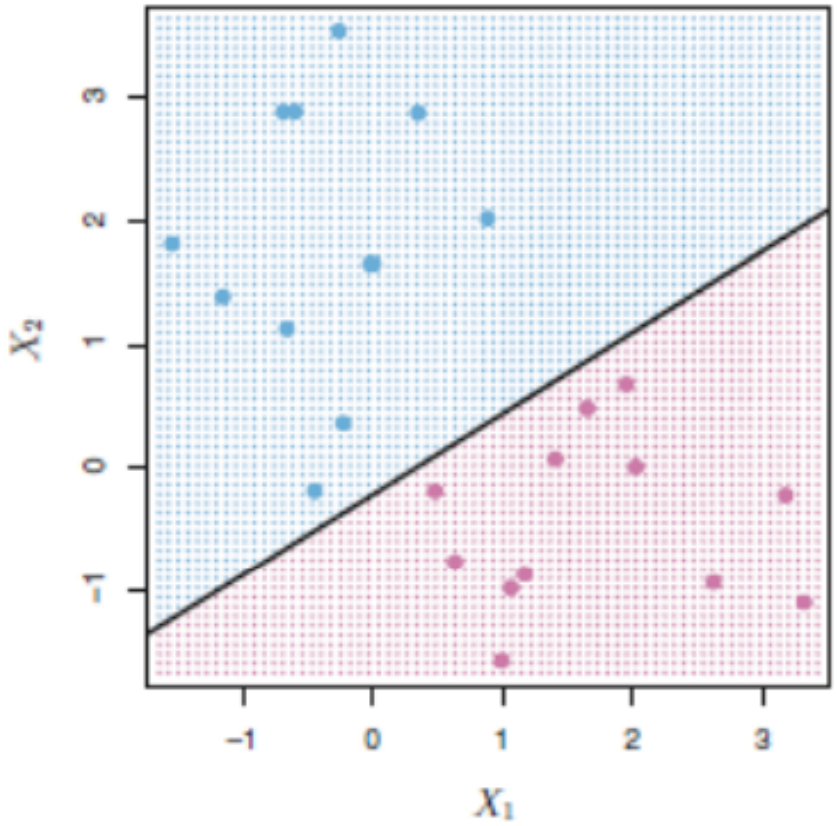


Figure: Let us assume, before mentioned dataset with 2 independent features (X1, X2) are plotted as 2D space or graph and separated by class category (Pass or Fail)

SVM - Maximal Margin Classifier (Contd...)

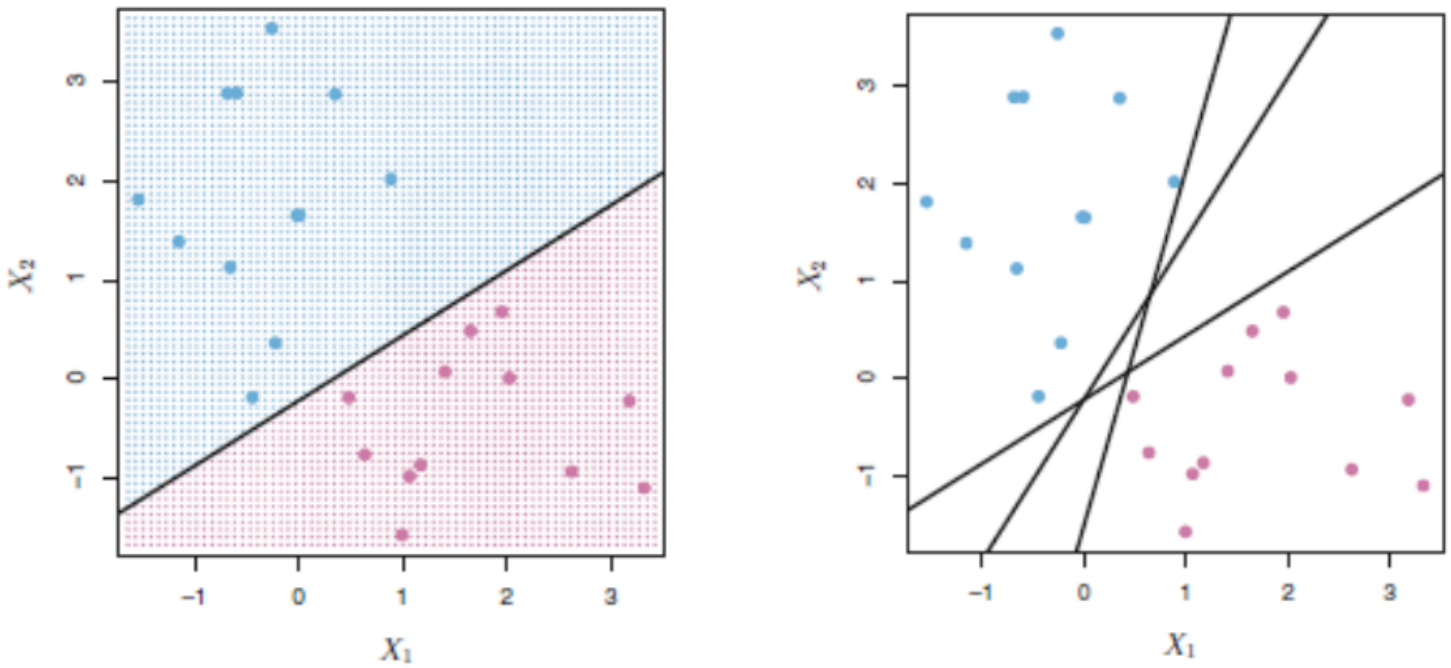


Figure: There are many possibilities to separate the two category class - But which is best?

SVM - Maximal Margin Classifier (Contd...)

- Best Hyperplane - How?
  - Calculate perpendicular distance of observation
  - Maximum value of margin

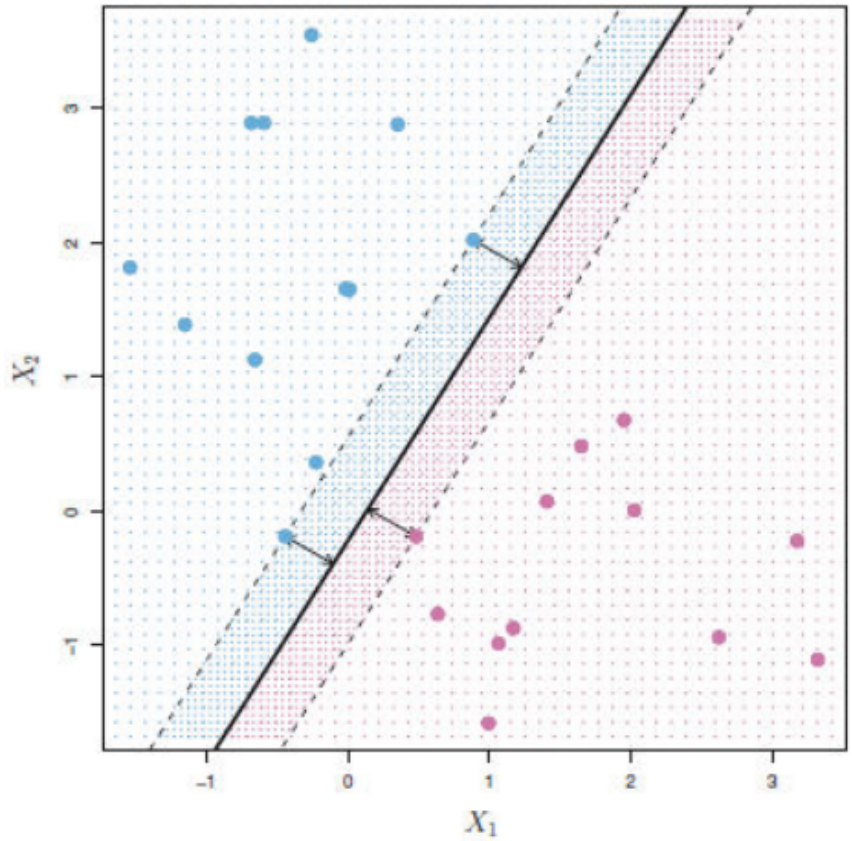


Figure: Hyperplane

H1 Vs H2 Vs H3

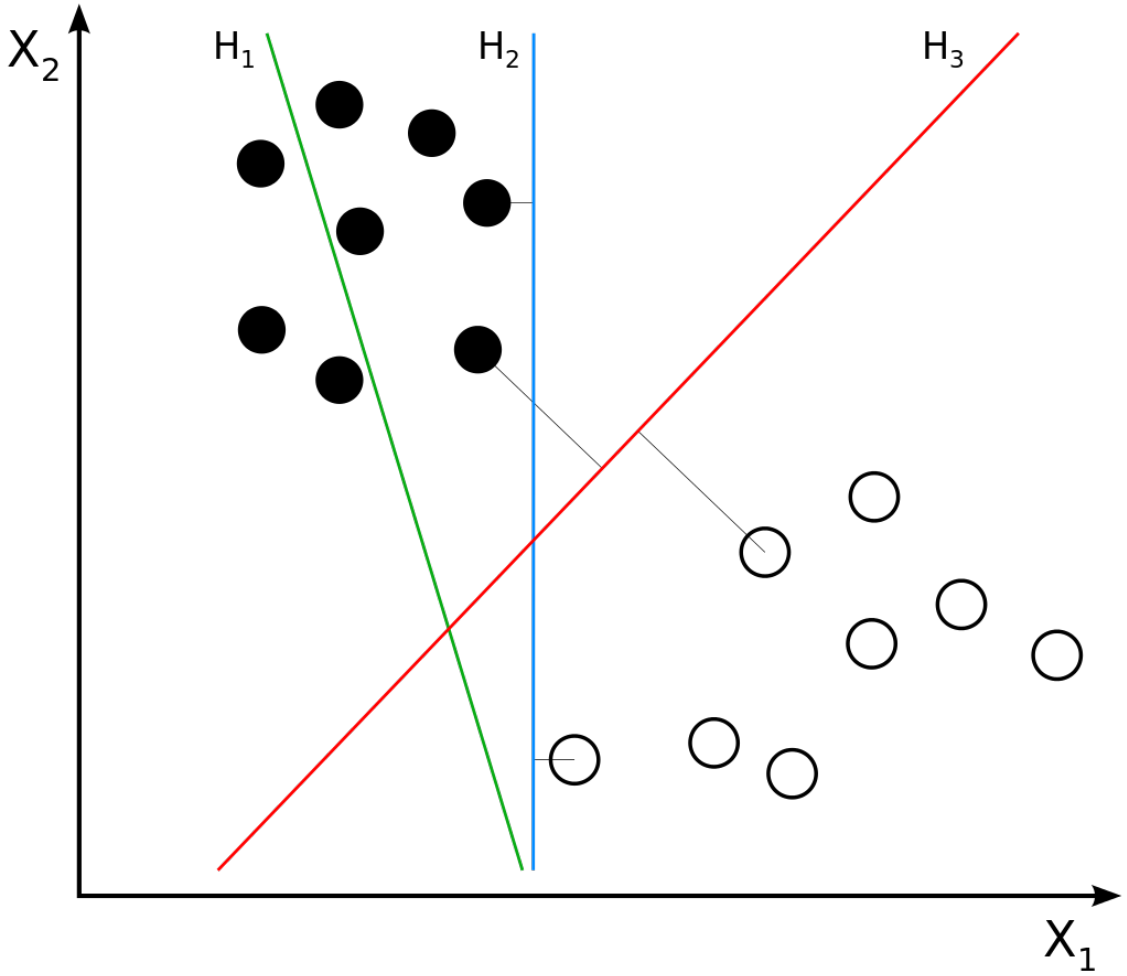


Figure: Hyperplane with maximum margin

SVM - Maximal Margin Classifier (Contd...)

- **Out-of-the-box** classifier ?
  - Observation fall on margin
  - Different from conventional ML

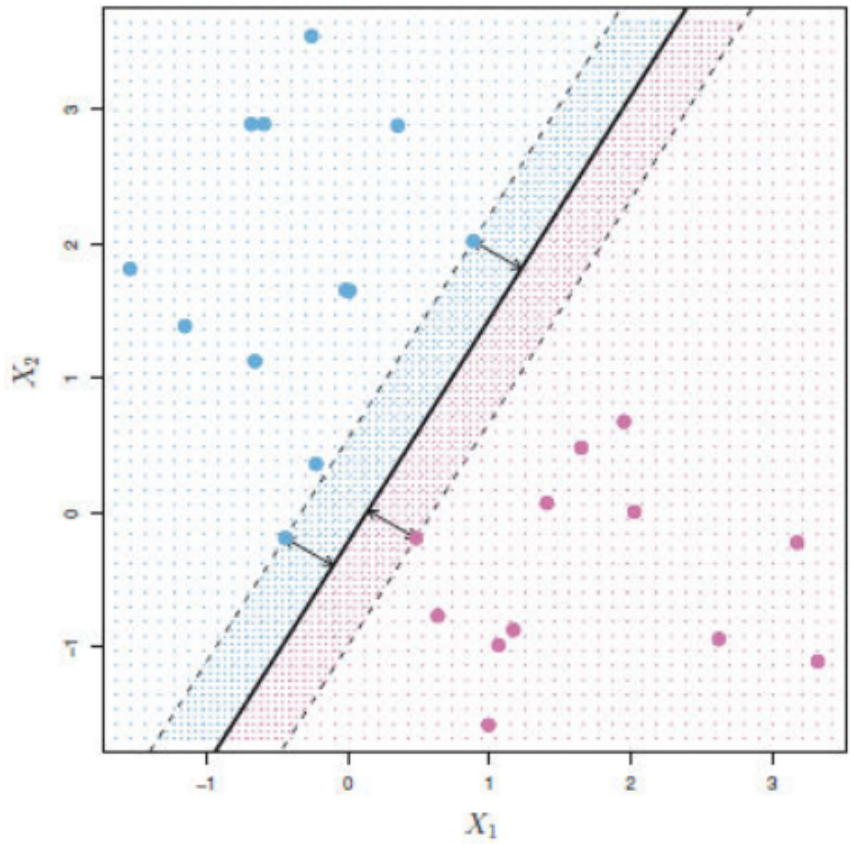


Figure: Hyperplane - Support Vector

SVM - Maximal Margin Classifier - Limitation

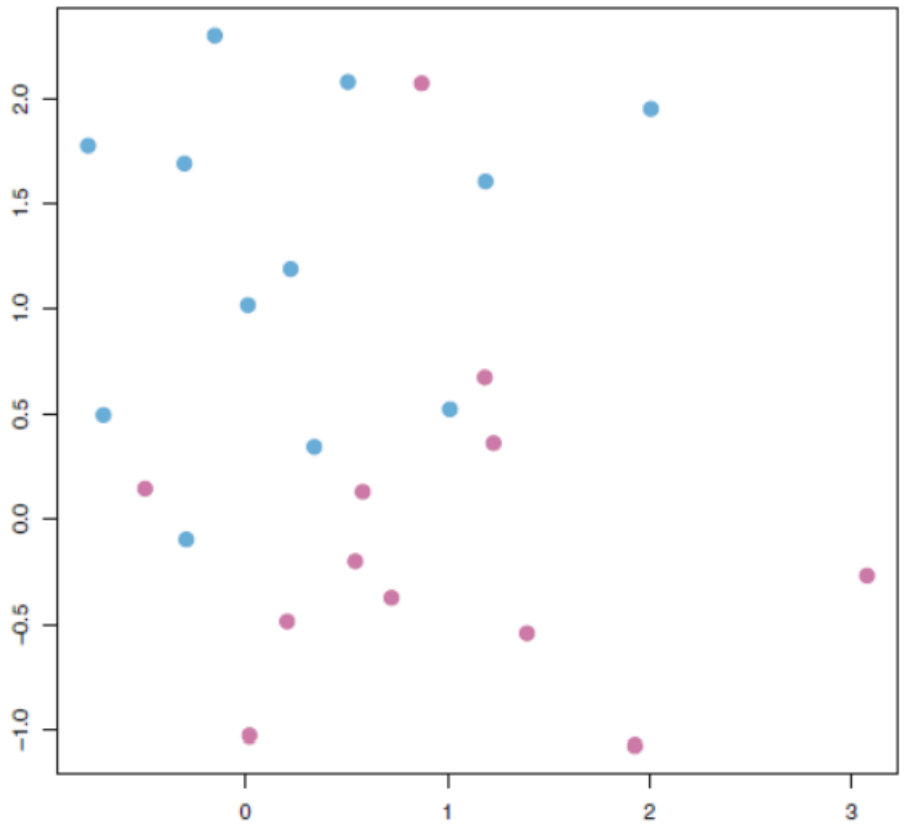


Figure: Linearly separable? - Yes / No

SVM - Maximal Margin Classifier - Limitation (Contd...)

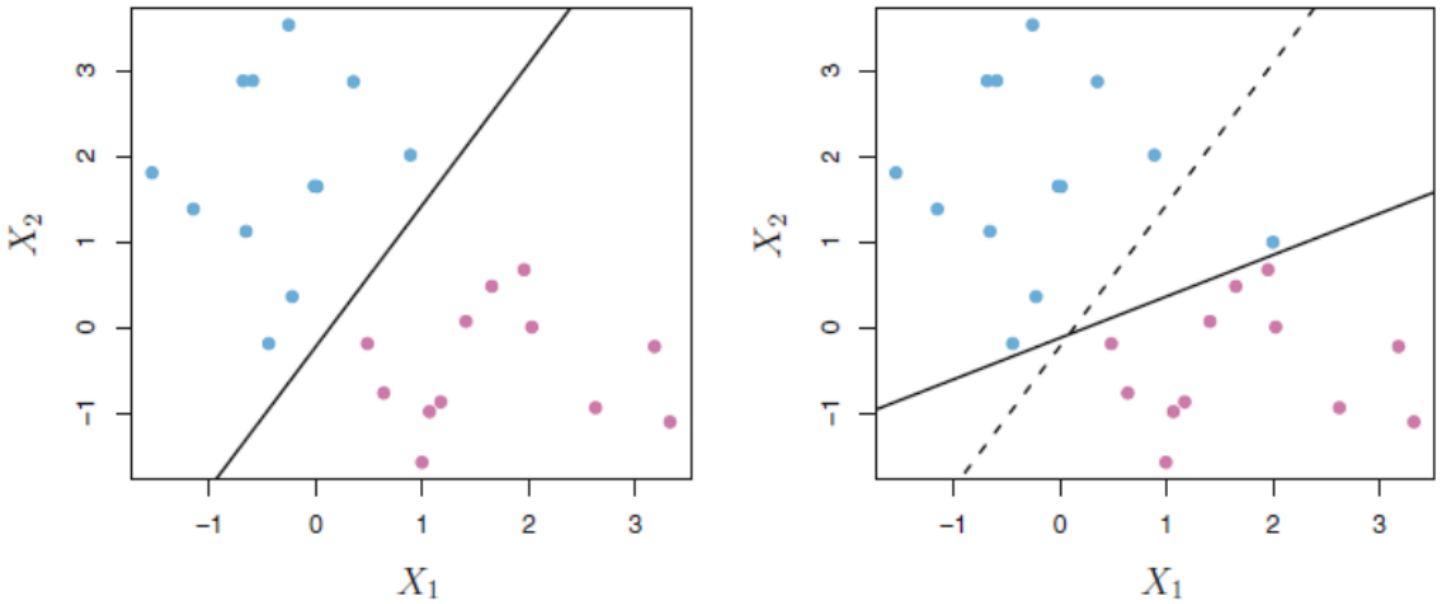
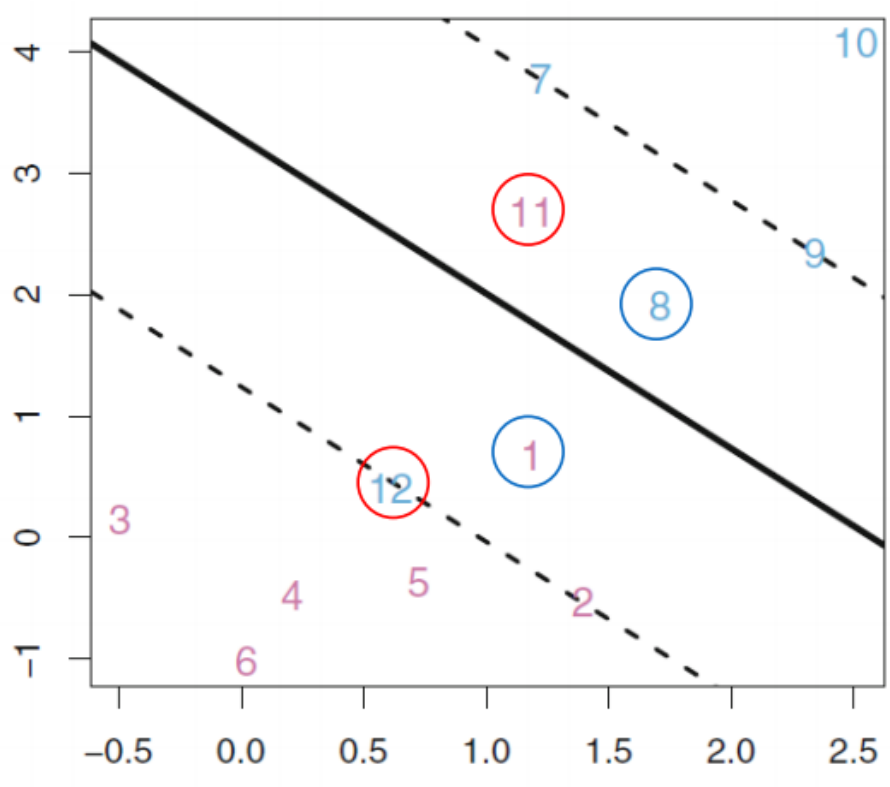


Figure: Support Vector - Sensitive!

Support Vector Classifier

- Non perfectly separable
- Robustness
  - Soft Margin Classifier
  - 1, 8 - Wrong side
  - 11, 12 - Wrong Classification



Support Vector Classifier - How?

- C
  - C small - Margin wide, Over-fitting
  - C large - Margin Narrow
  - Choose C wisely
  - SVM - C parameter

Support Vector Classifier- Limitation

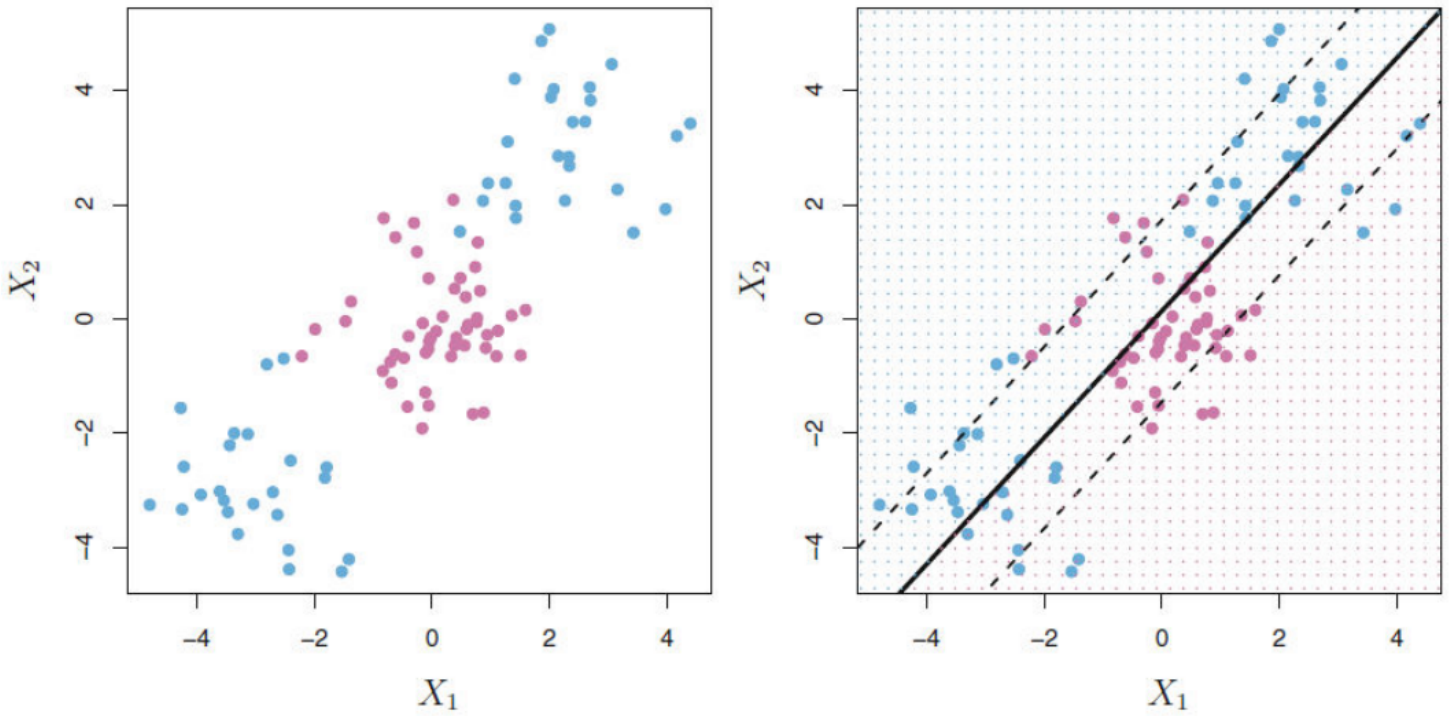
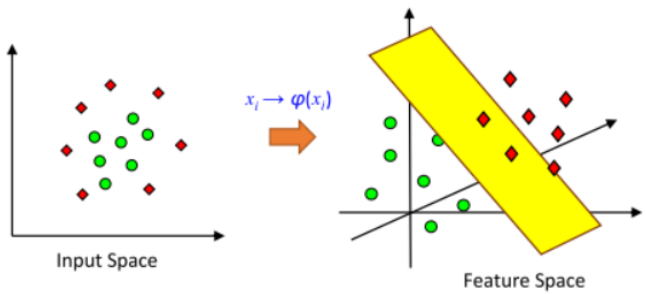


Figure: Hyperplane - Efficient? - If not how to increase?

Support Vector Machine - Atlast!

- SVM is an extension of SVC which uses kernels for non-linear condition - Hit!
- So Kernel?
  - Linear kernel
  - Polynomial kernel - hit!
  - Radial Basis Function (RBF) !



Kernels	Formula
linear	$k(x, y) = x.y$
sigmoid	$k(x, y) = \tanh(ax.y + b)$
polynomial	$k(x, y) = (1 + x.y)^d$
RBF	$k(x, y) = \exp(-a\|x - y\|^2)$
exponential RBF	$k(x, y) = \exp(-a\|x - y\|)$

Not satisfied yet! still need more?

Algorithm Explanation - K-Nearest Neighbors

Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.

K Nearest Neighbor

- Supervised learning
- Lazy learning algorithm since it doesn't have a specialized training phase
- Non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data

K Nearest Neighbor

- Load the data
- Initialize K to your chosen number of neighbors
- Calculate the distance between the query example and the current example from the data.
- Add the distance and the index of the example to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the first K entries from the sorted collection
- Get the labels of the selected K entries
- If regression, return the mean of the K labels
- If classification, return the mode of the K labels

K-Nearest Neighbors - Steps

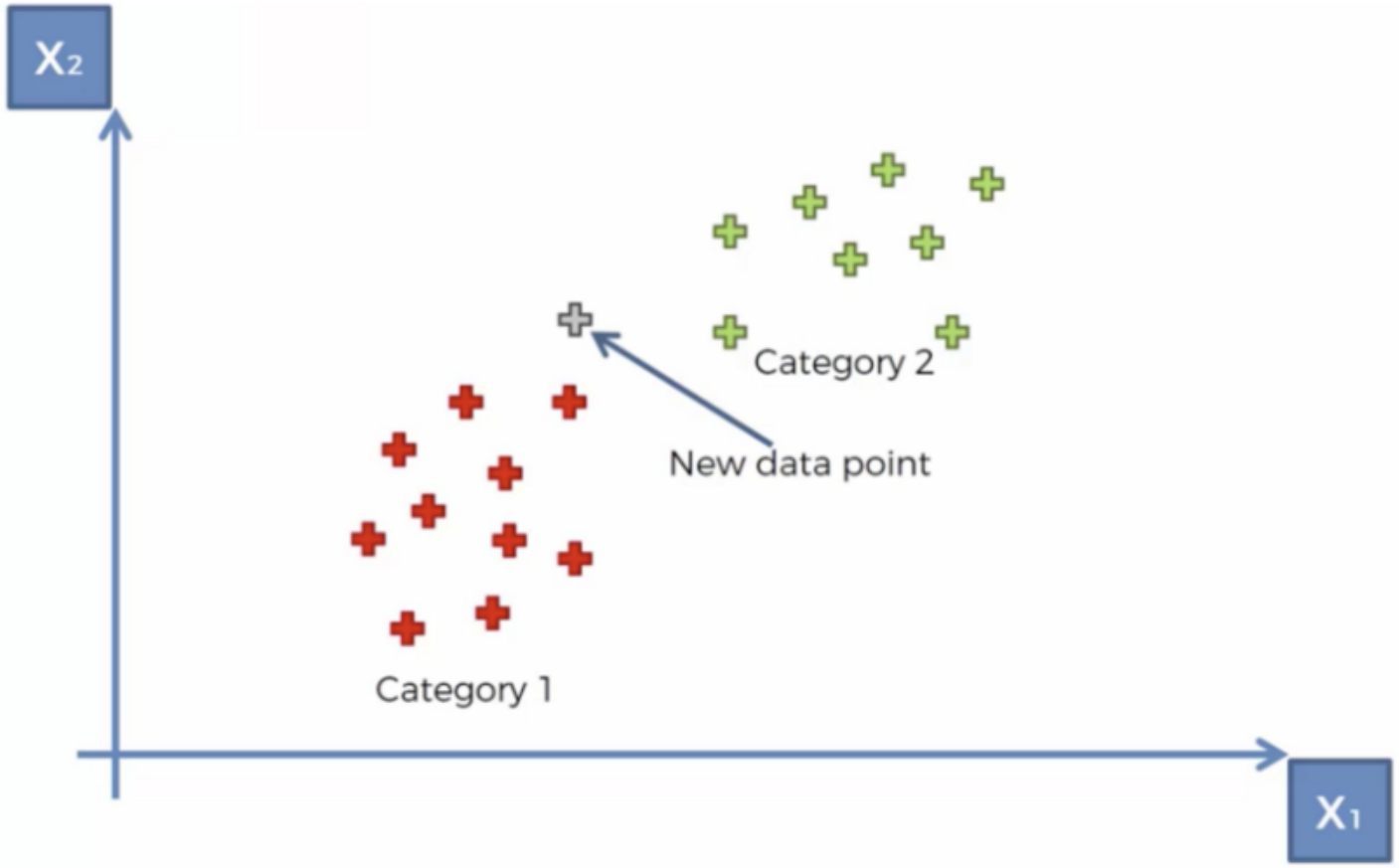


Figure: Pick a value for K (i.e. 5)

K-Nearest Neighbors - Steps

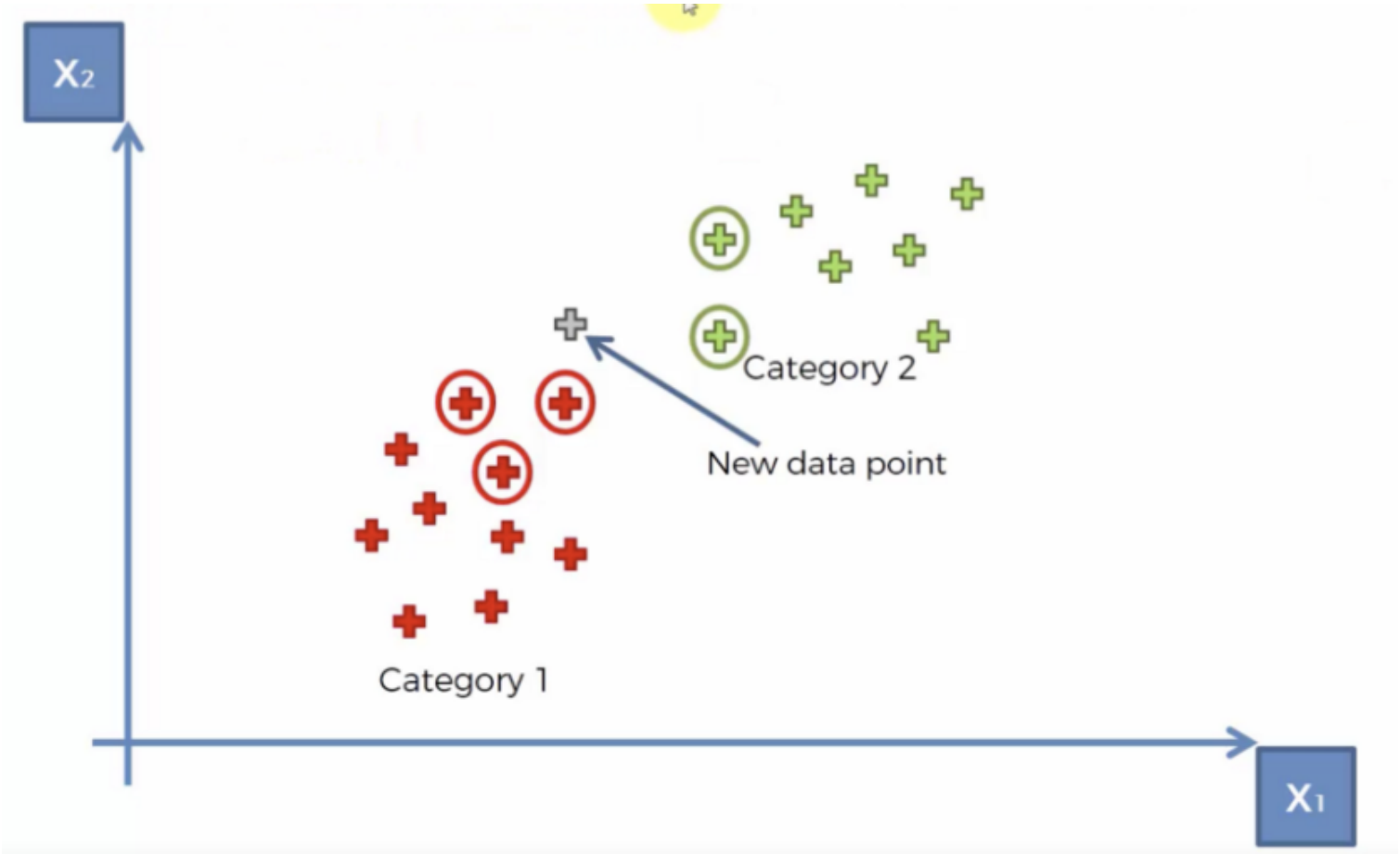
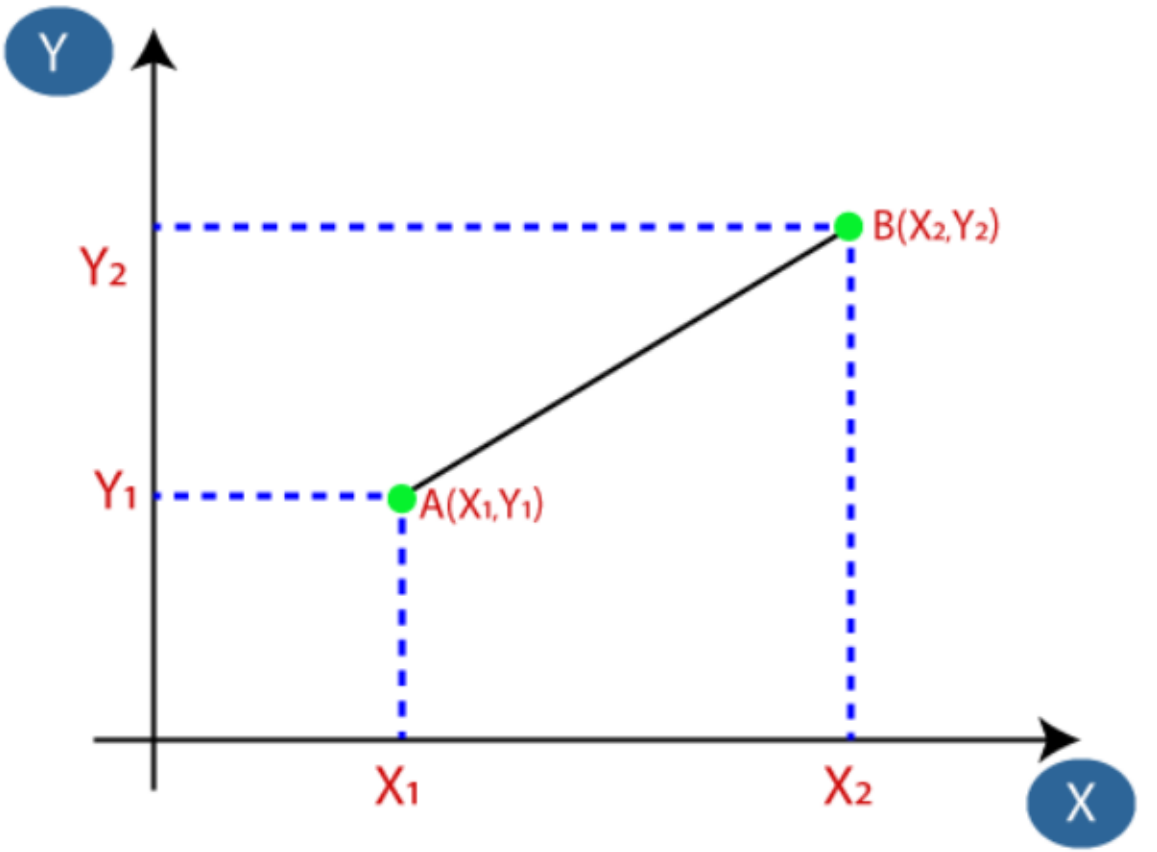


Figure: Take the K nearest neighbors of the new data point according to their Euclidean distance

K-Nearest Neighbors - Steps



Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

Figure: According to Euclidean distance

K-Nearest Neighbors - Steps

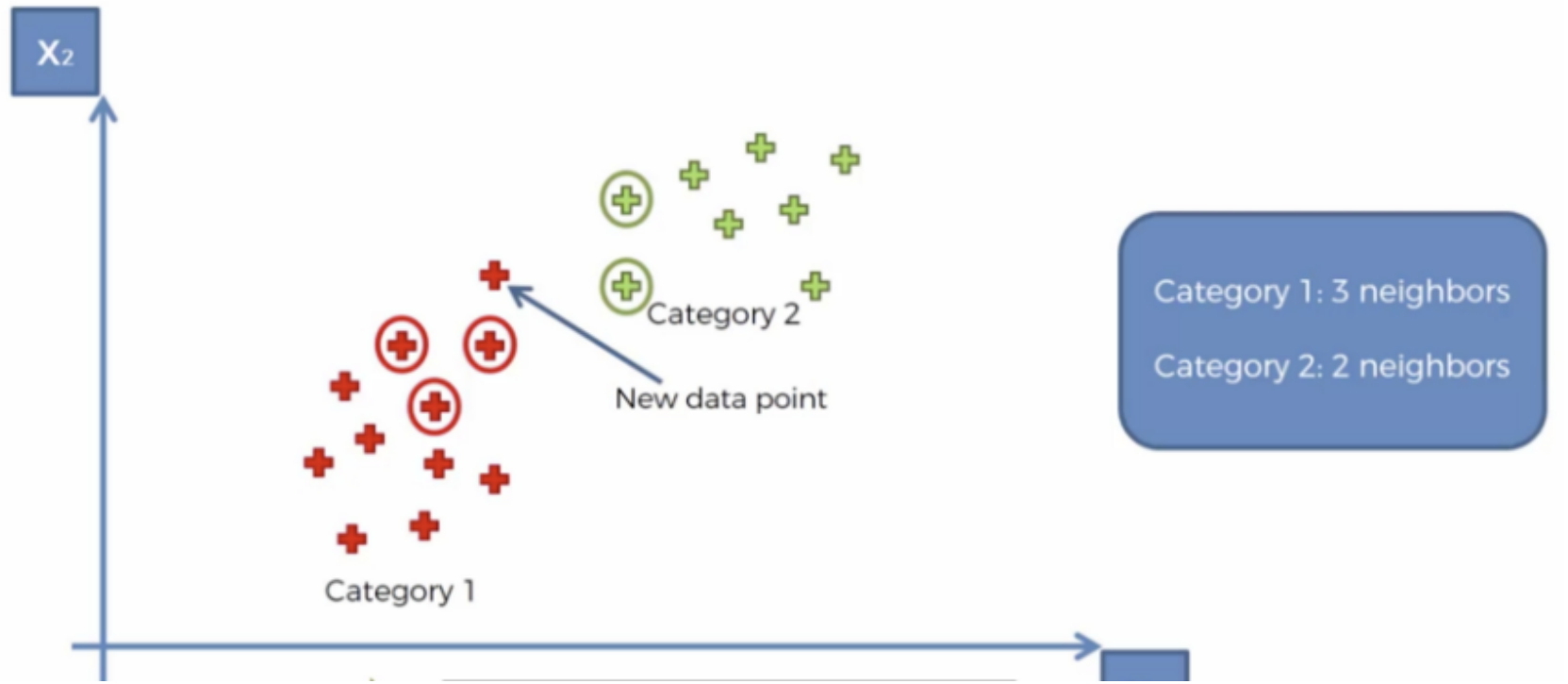


Figure: Count the number of data points in each category and assign the new data point to the category where you counted the most neighbors

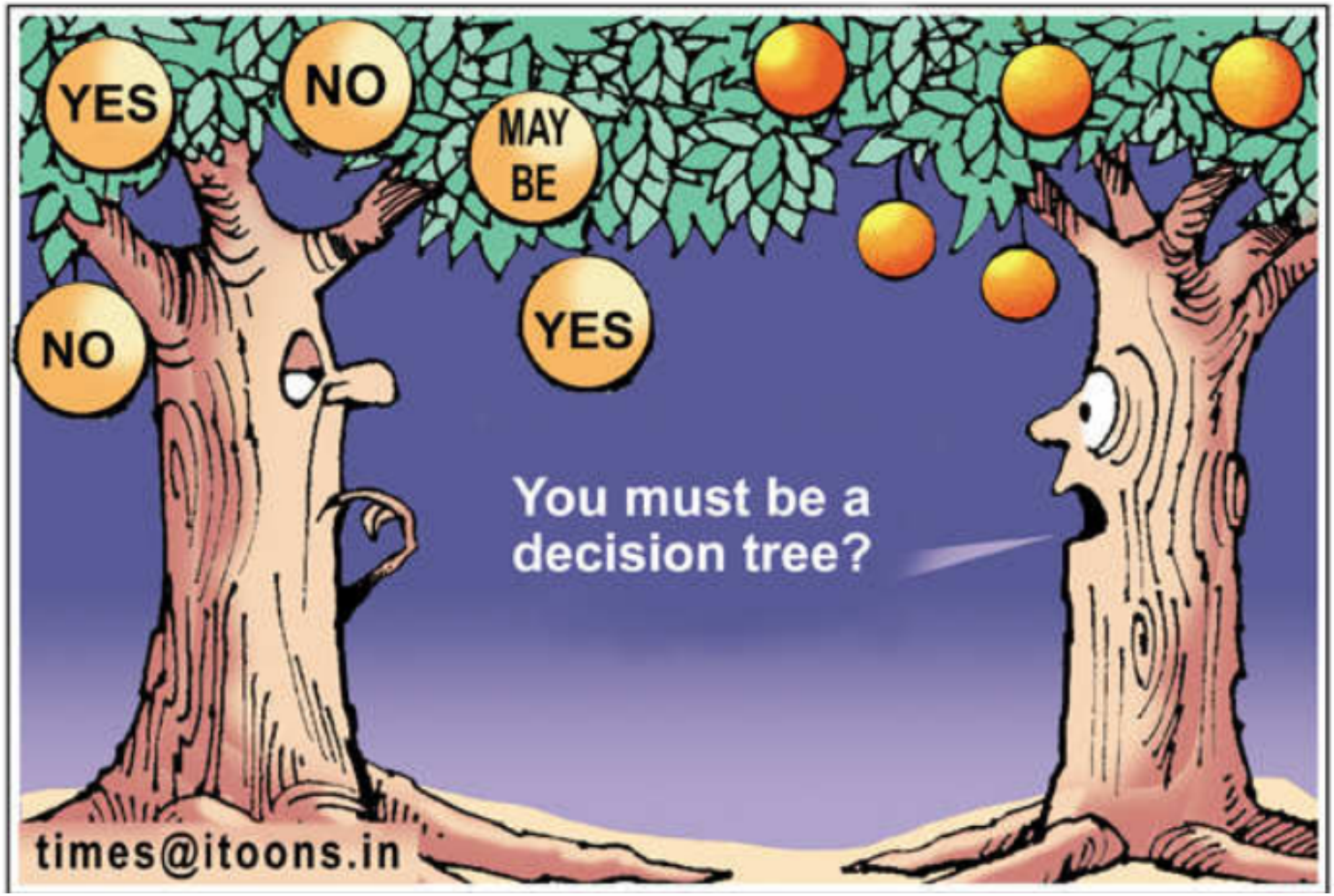
K-Nearest Neighbors - Advantage & Disadvantage

- **Advantage**
  - Simple algorithm — easy to understand
  - No assumptions about data
  - There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)
  - Since the algorithm requires no training before making predictions, new data can be added seamlessly.
- **Disadvantage**
  - Sensitive to the scale of the data since we're computing the distance to the closest K points
  - Doesn't work well with categorical features since it is difficult to find the distance between dimensions with categorical features.
  - Doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate distance in each dimension

K-Nearest Neighbors - Applications

- Economic forecasting
- Data compression
- Genetics
- Recommender System

Algorithm Explanation - Decision Tree

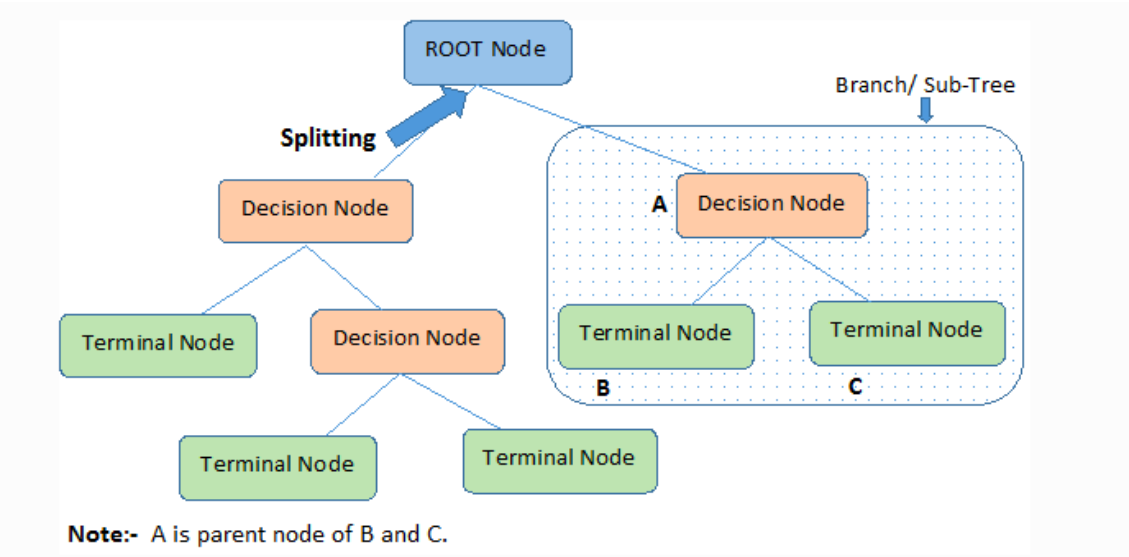


Decision Tree

- Supervised learning
- Classification (categorical or qualitative) as well as regression problems (continuous or quantitative)
- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- Ex: Toll free number of your bank or broadband connection process
- Different algorithms in DT,
  - Chi-squared Automatic Interaction Detection (CHAID)
  - Classification And Regression Tree (CART)
  - Iterative Dichotomiser 3 (ID3)
  - C4.5 (successor of ID3) (C5.0 - latest algorithm)

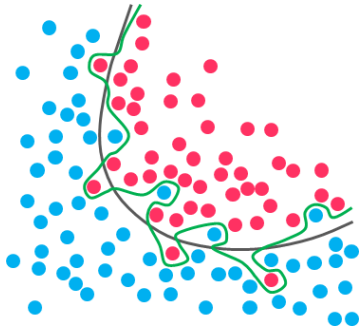
Decision Tree - Structure

- **Root node** - entire dataset, which further gets divided into two or more homogeneous sets - top most decision node
- **Leaf / Terminal node** - final output node, and the tree cannot be segregated further
- **Splitting** - dividing the decision node/root node into sub-nodes
- **Branch / Sub Tree** - formed by splitting the tree
- **Pruning**- removing the unwanted branches from the tree.
- **Parent / Child node** - root node of the tree is called the parent node, and other nodes are called the child nodes.



Decision Tree - Prune that tree

- model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data



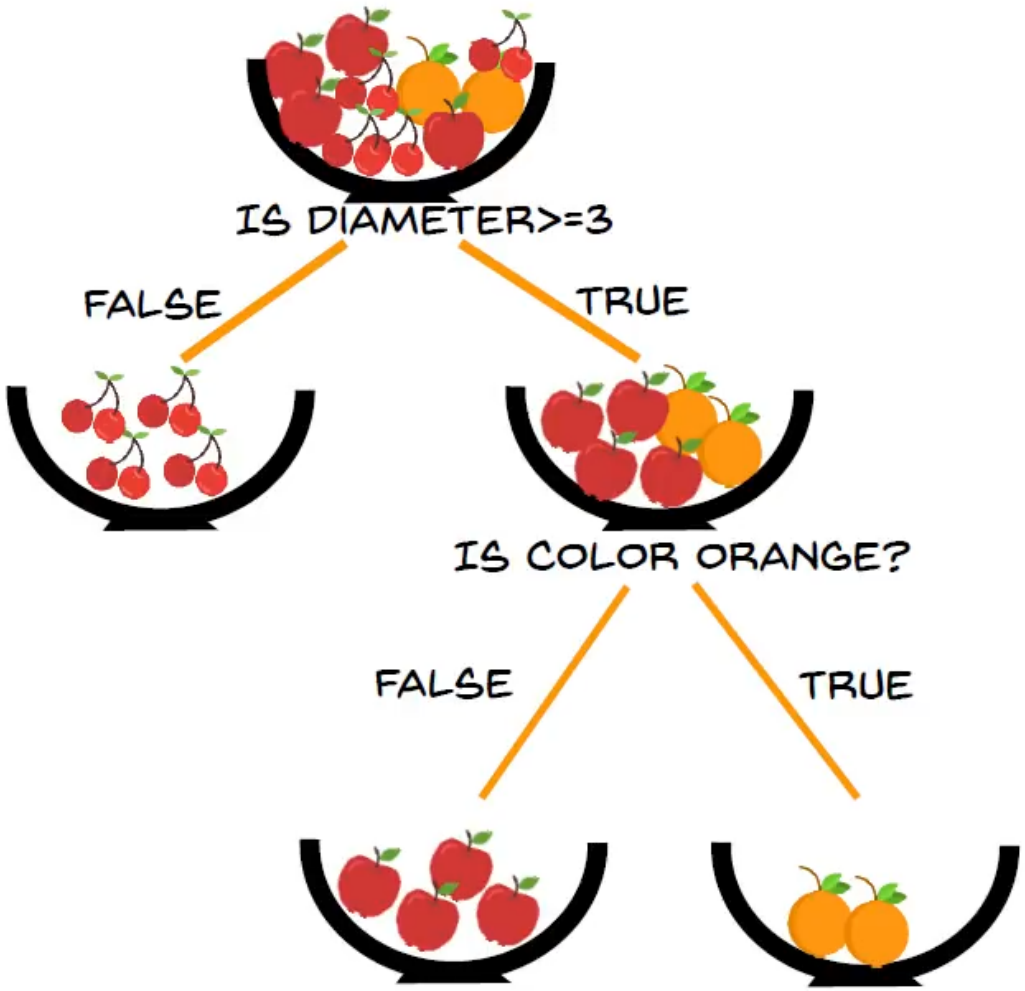
- Pruning is a technique used to deal with overfitting, that reduces the size of DTs by removing sections of the Tree that provide little predictive or classification power.
- **Pre-prune:** When you stop growing DT branches when information becomes unreliable.
- **Post-prune:** When you take a fully grown DT and then remove leaf nodes only if it results in a better model performance.

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Decision Tree - Steps

- Begin the tree with the root node, says S, which contains the complete dataset.
- Find the best attribute in the dataset using **Attribute Selection Measure (ASM)**
- Divide the S into subsets that contains possible values for the best attributes
- Generate the decision tree node, which contains the best attribute
- Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node

Decision Tree - Steps



Decision Tree - Attribute Selection Measure

- Entropy (ID3 Algorithm)
- Information Gain
- Gini Index (CART Algorithm)

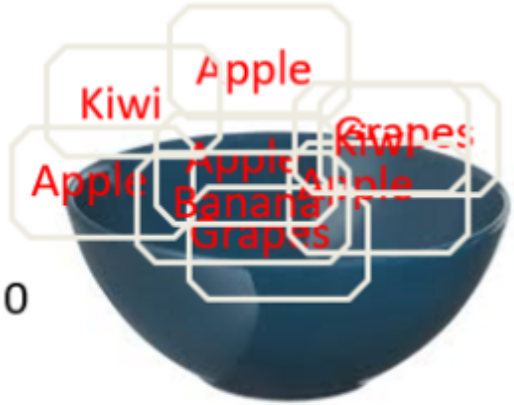
Impurity

Figure: When impurity is equal to zero

Impurity

Grapes

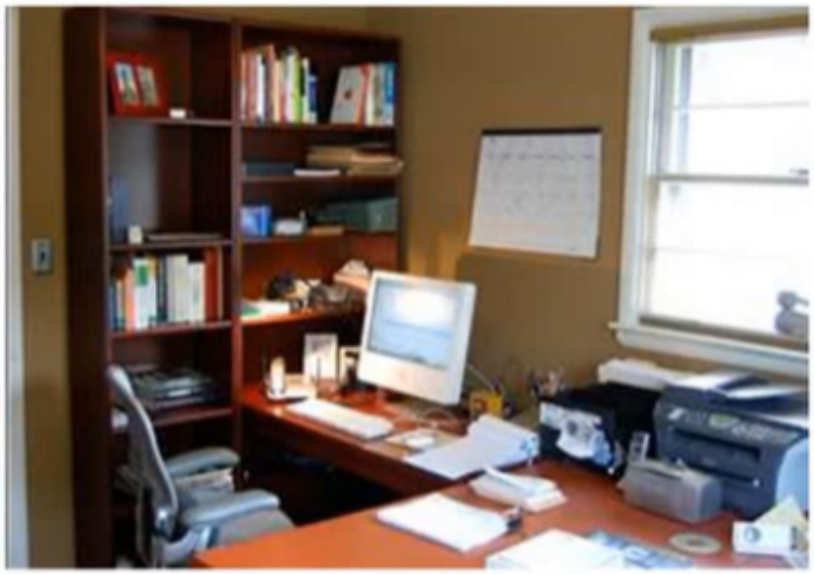
Impurity is not equal to 0



Entropy

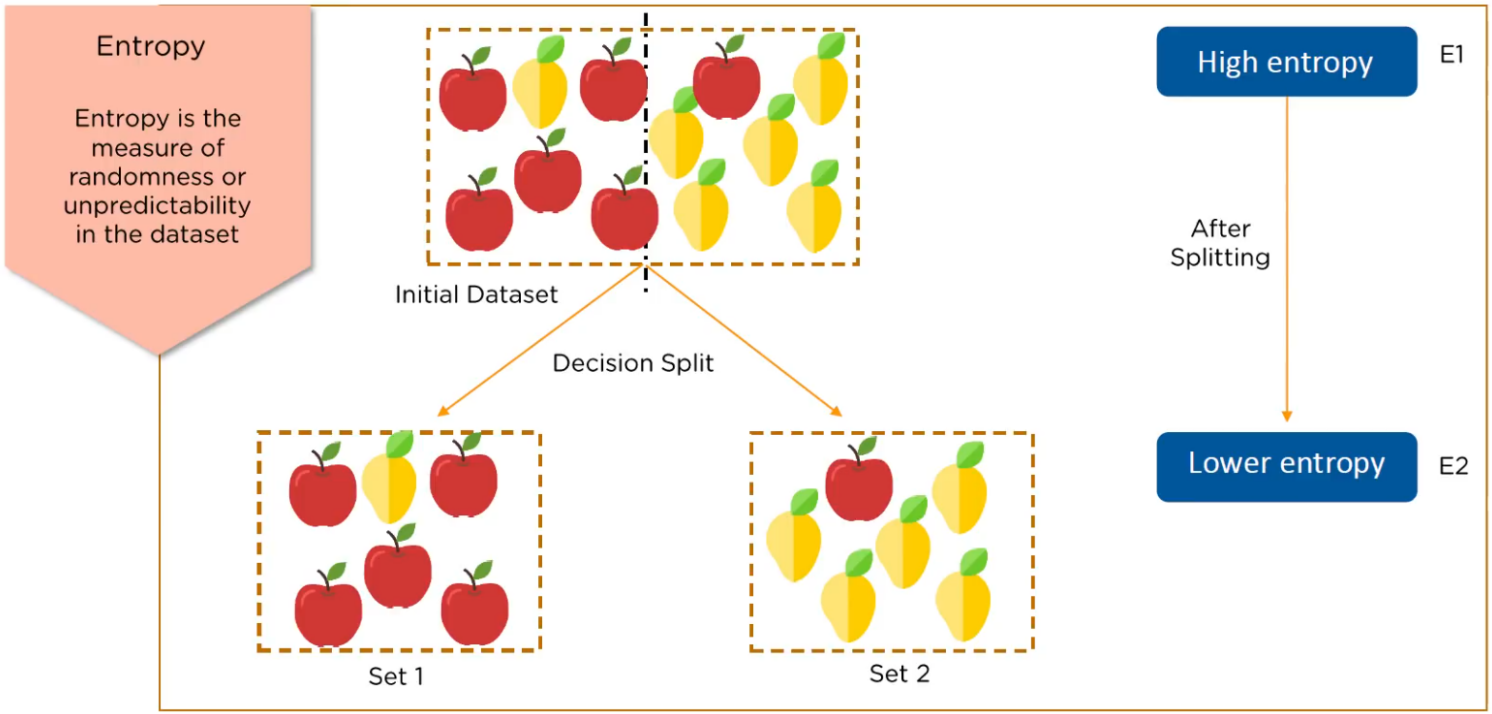
- Entropy is an indicator of how messy your data is
- Entropy is the measure of randomness or unpredictability in the dataset
- It controls how a decision tree decides to split the data
- Its value ranges from 0 (if all samples of a node belong to the same class) to 1

High Entropy (messy)



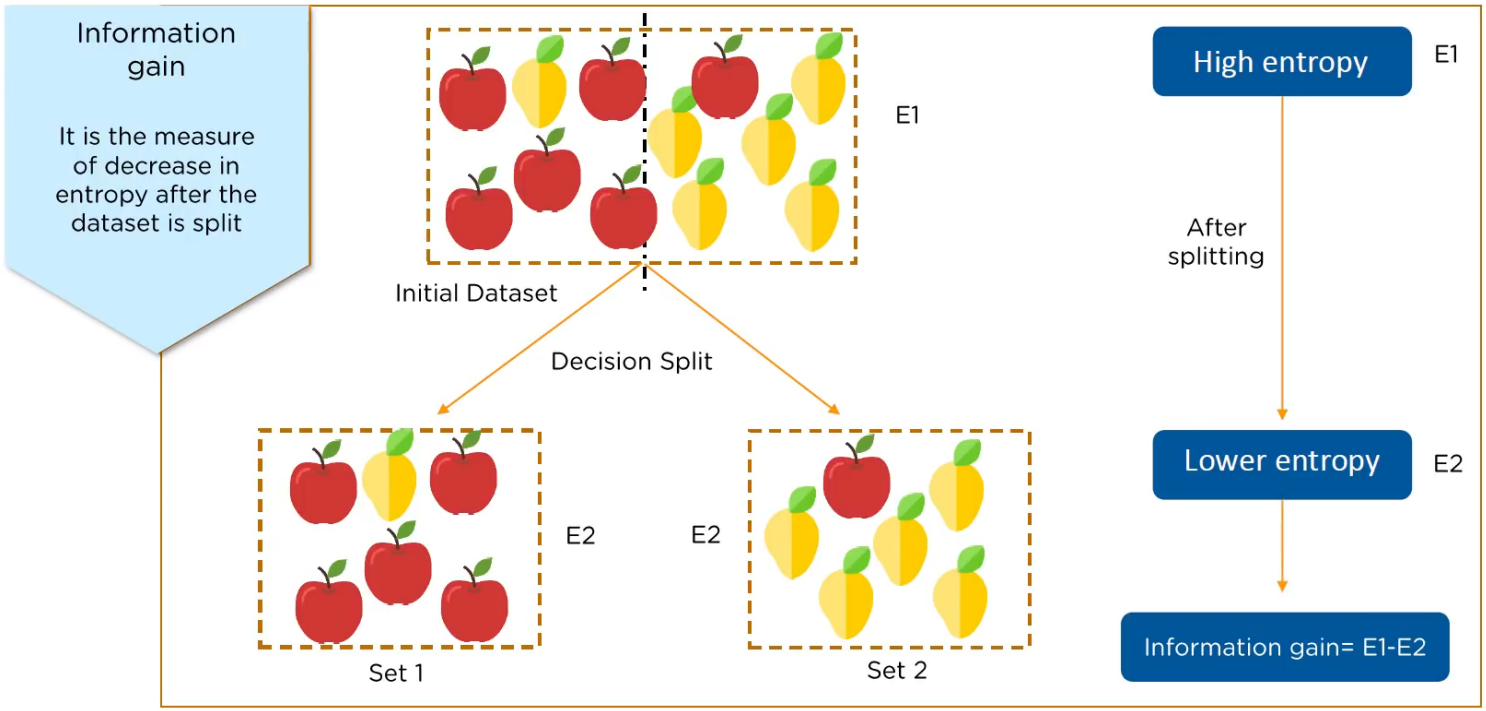
Low Entropy (Clean)

Entropy


Entropy
Entropy is the measure of randomness or unpredictability in the dataset
Initial Dataset
Decision Split
Set 1
Set 2
High entropy
E1
After Splitting
Lower entropy
E2

Information Gain

- Measure of decrease in entropy after the dataset is split
- An attribute with the highest Information gain will be tested/split first



Information Gain








PROBLEM STATEMENT

TO TEACH YOUR BABY TO PICK WINTER FAMILY VACATION PHOTO USING DECISION TREE

95

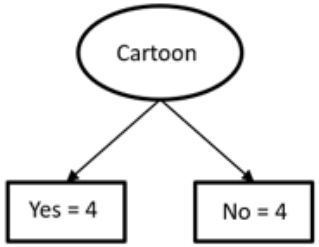
Information Gain

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

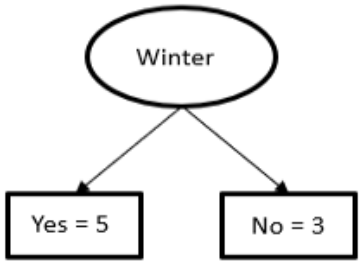
Information Gain

- Total of 8 photos. Winter family photo — 1 (Yes), Now winter family photo — 7 (No). If we substitute in the above entropy formula
- -(1/8) * log2(1/8) — (7/8) * log2(7/8)
- Entropy = 0.543
- Attributes = cartoon, winter and >1

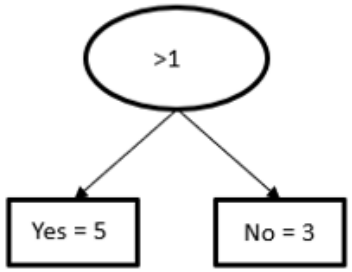
Information Gain



Information Gain(winter family photo, cartoon)
= 0.543 – (4/8 * E([0+,4-]) + 4/8 * E([1+, 3-]))
= 0.138

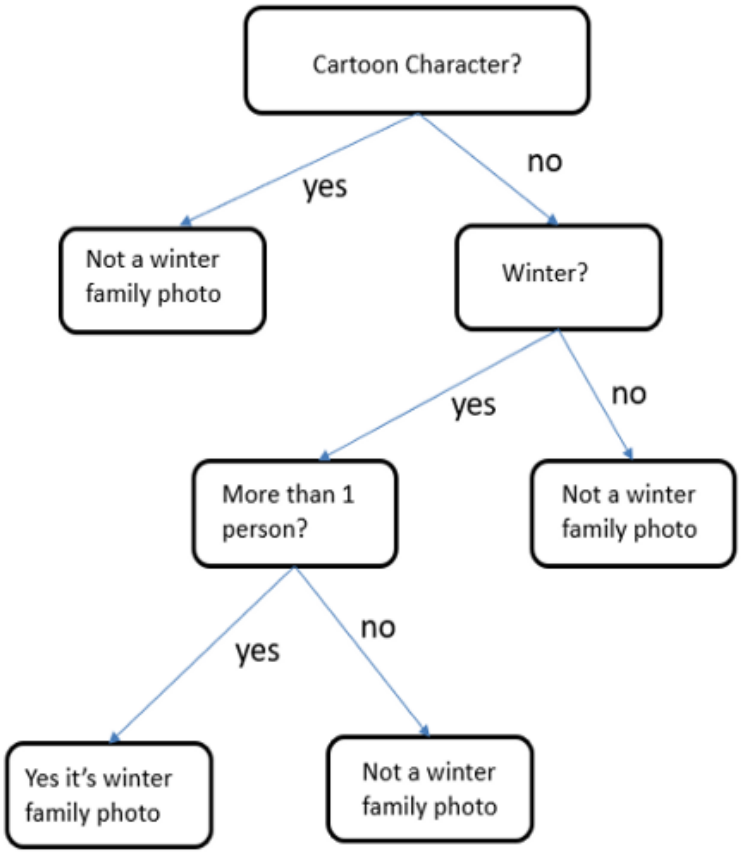


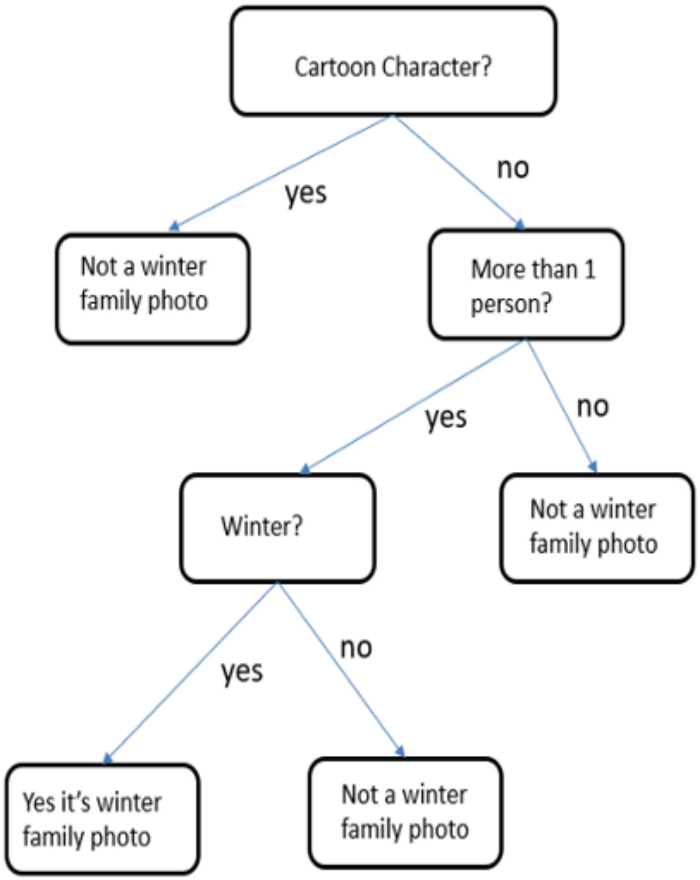
Information Gain(winter family photo, winter)
= 0.543 – (5/8 * E([1+,4-]) + 3/8 * E([0+, 3-]))
= 0.093



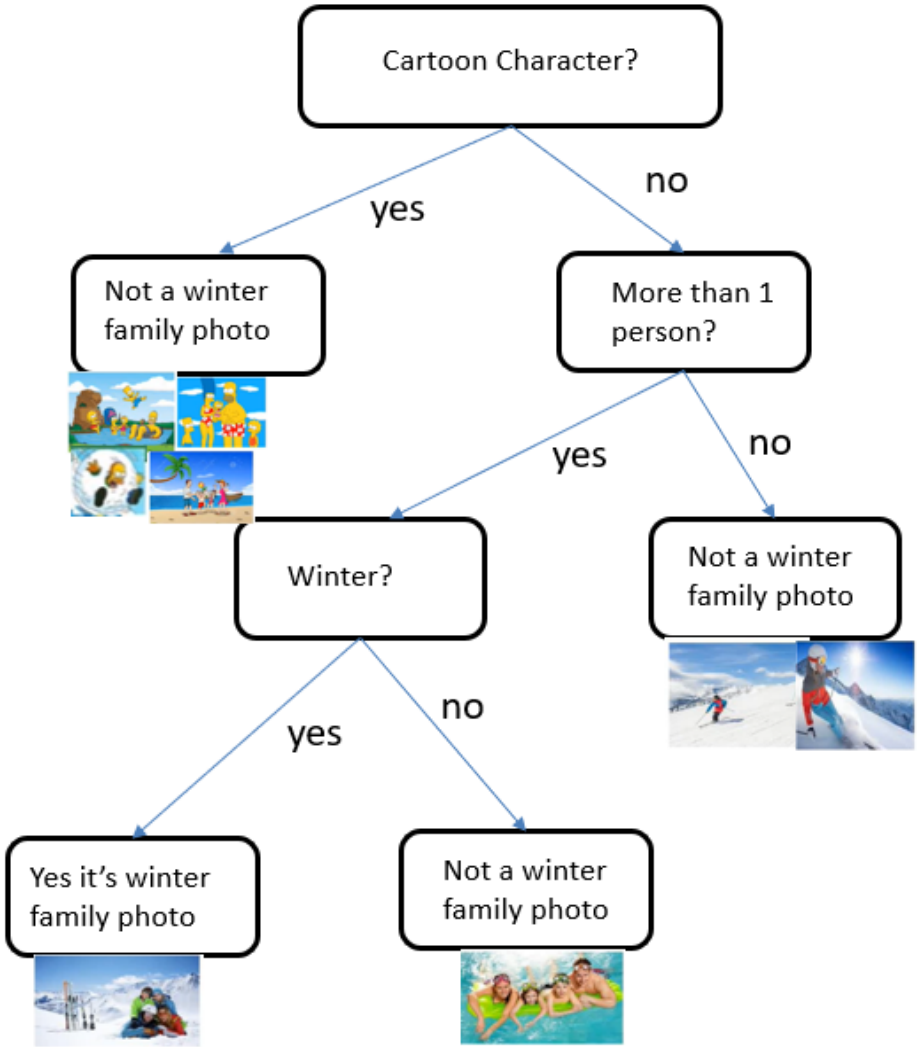
Information Gain(winter family photo, >1)
= 0.543 – (5/8 * E([1+,4-]) + 3/8 * E([0+, 3-]))
= 0.093

Information Gain





Information Gain



Decision Tree - Advantage

- Decision trees are easy to visualize and interpret
- It can easily capture non — linear patterns
- It can handle both numerical and categorical data
- Little effort required for data preparation. (example, no need to normalize the data)

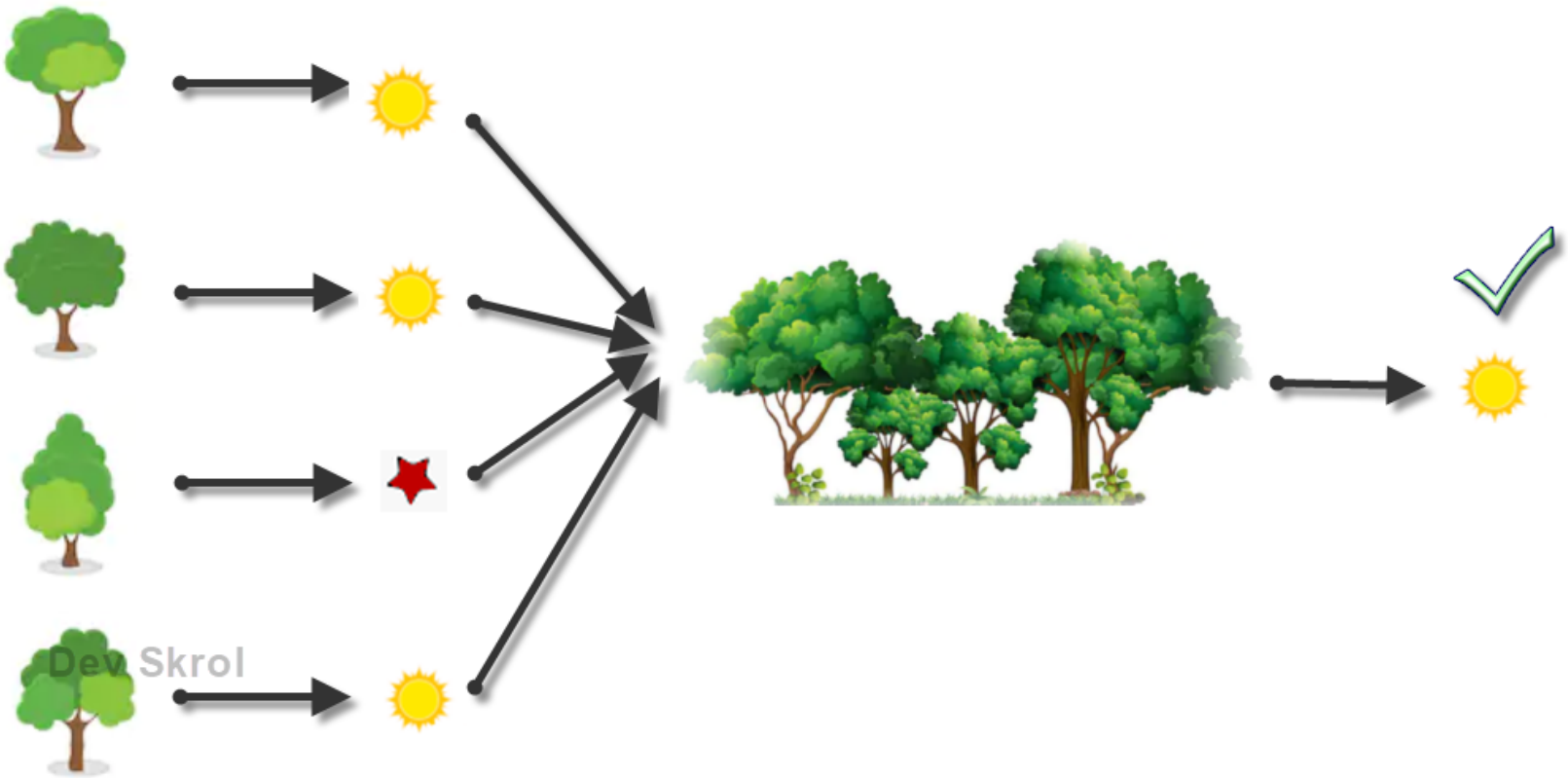
Decision Tree - Disadvantage

- Overfitting is one of the most practical difficulties for decision tree models
- Decision trees are biased with imbalance dataset, so it is recommended that balance out the dataset before creating the decision tree
- It is unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
- Low accuracy for continuous variables: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories.

Decision Tree - Applications

- Sophia robot uses DT when interaction with humans
- Chat-bots (Health insurance) , Google (Onward company), amazon
- Selecting a flight to travel
- Choosing a best friend
- DTs and satellite imagery are also used in agriculture to classify different crop types and identify their phenological stages.
- Investment Solutions
- Sentiment Analysers (text)
- Financial fraud detection

Algorithm Explanation - Random Forest



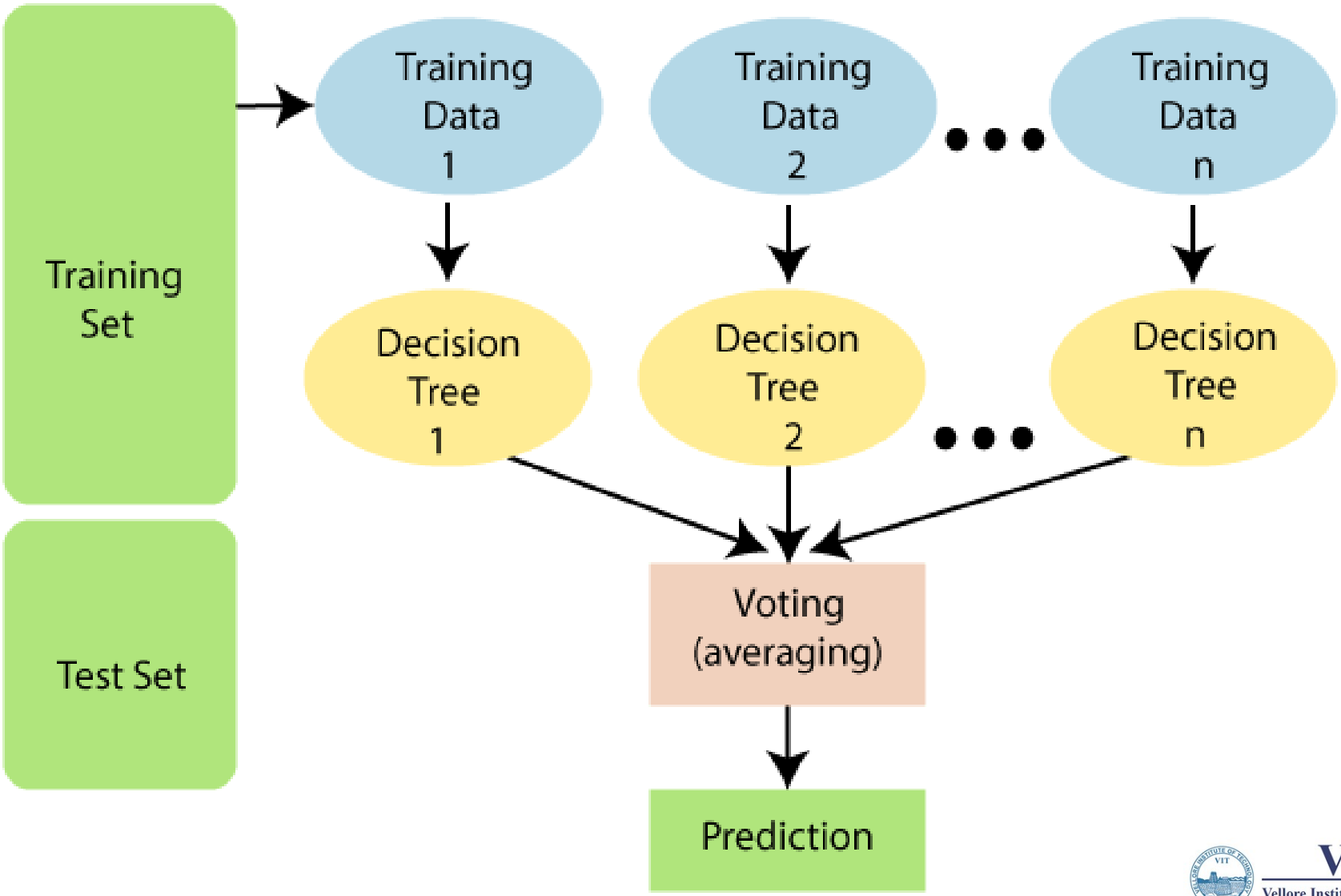
Random Forest

- Random Decision Forest - Supervised learning
- High accuracy - runs efficiently on large database
- Based on the concept of ensemble learning (Bagging - Bootstrap Aggregation), which is a process of combining multiple classifiers
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting

Random Forest

- The decision of the majority of the trees is chosen by the random forest as the final decision
- RF (Classification) - Majority Vote
- RF (Regression) - Continuous data - Mean or Median
- DT - Low Bias (Training Error - low - as goes deep) — High Variance (Testing error - more) $->$ Low Variance (Acc) (Random Forest - Expert (multiple times))
- Ex: Asking opinion for picnic to friends and deciding the spot

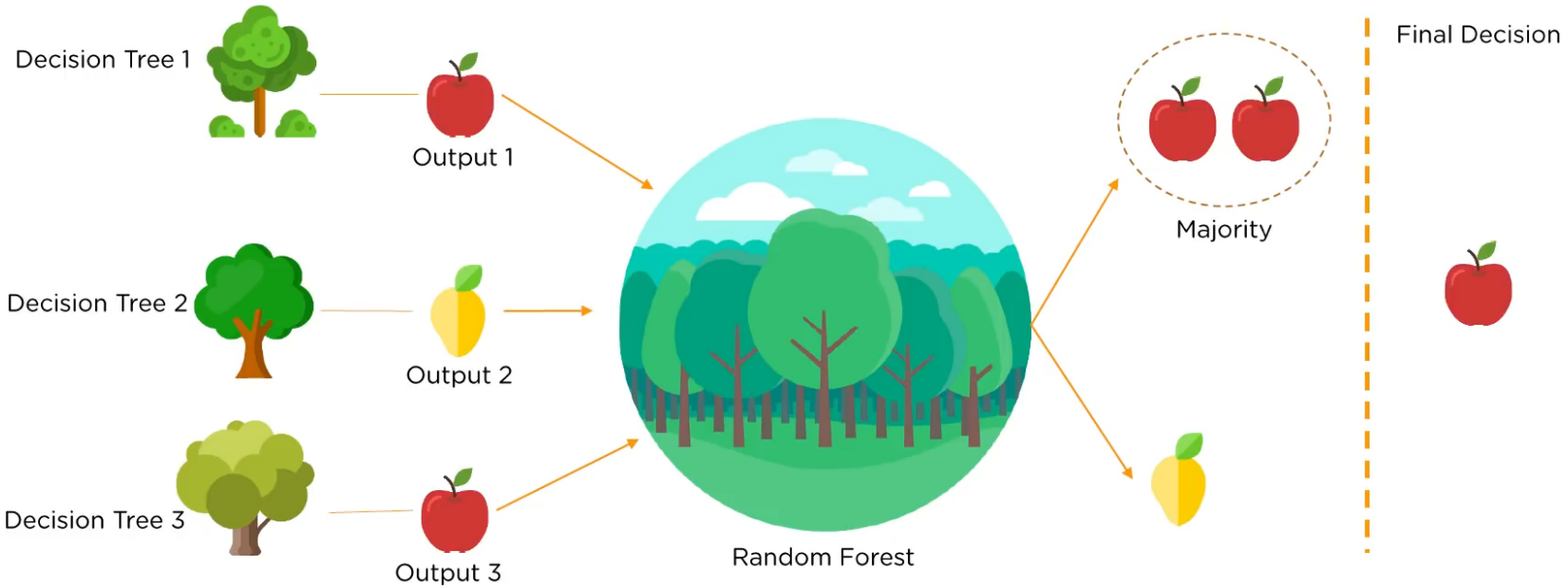
Random Forest - How it works?



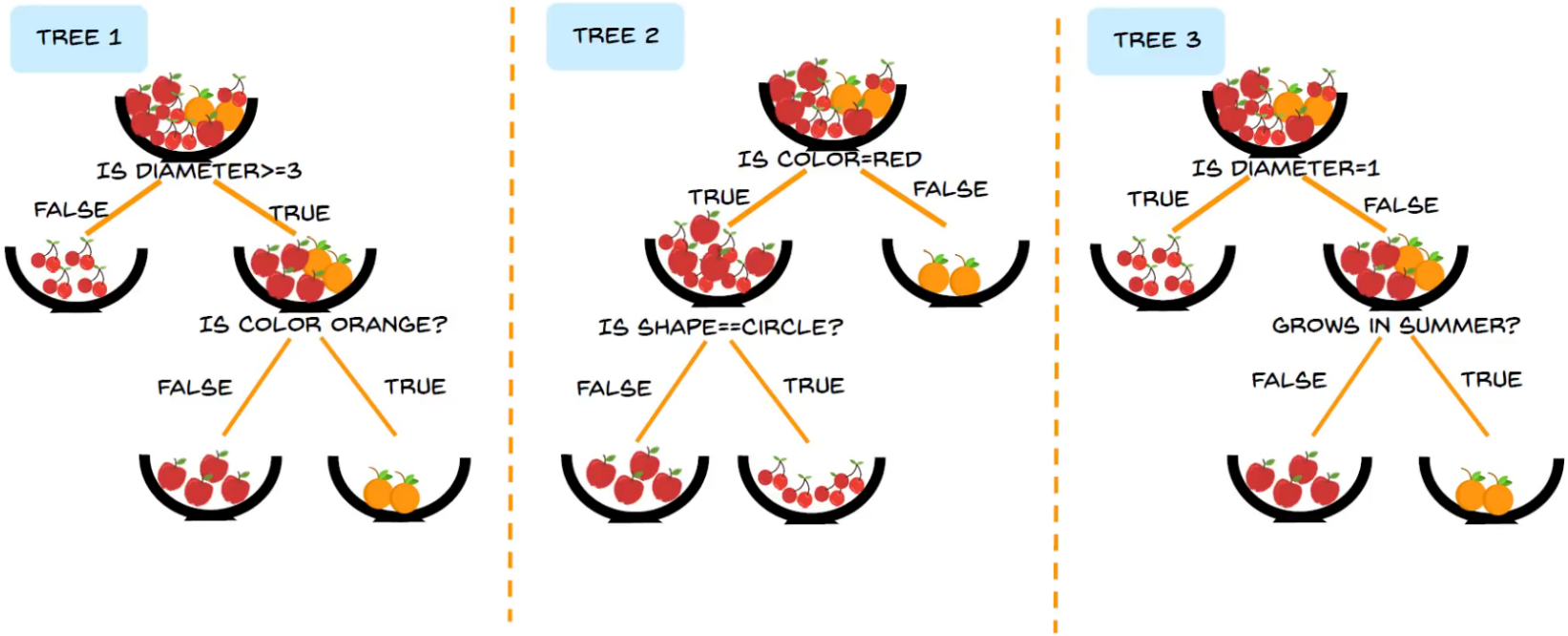
Random Forest - Steps

- Select random K data points from the training set.
- Build the decision trees associated with the selected data points (Subsets)
- Choose the number N for decision trees that you want to build
- Repeat Step 1 2
- For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes

Random Forest - How it works?



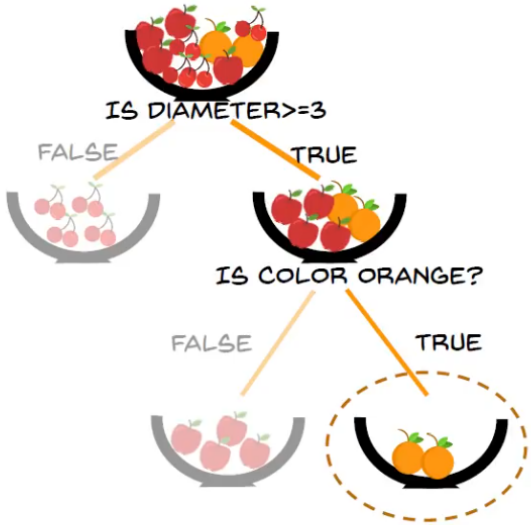
Random Forest - How it works?



Random Forest - How it works? - Missed values

TREE 1 CLASSIFIES IT AS AN ORANGE

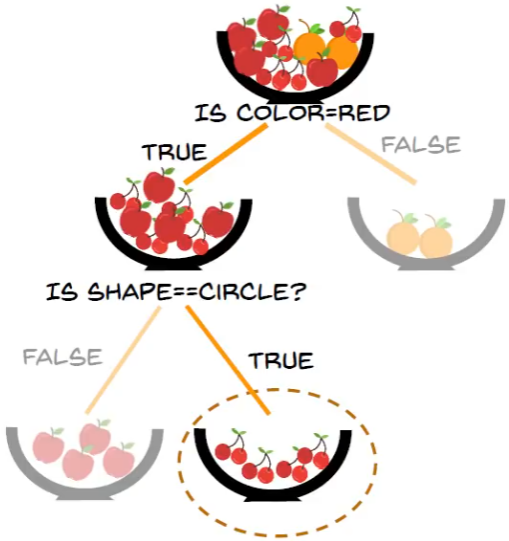
DIAMETER = 3
COLOUR = ORANGE
GROWS IN SUMMER = YES
SHAPE = CIRCLE



Random Forest - How it works? - Missed values

TREE 2 CLASSIFIES IT AS CHERRIES

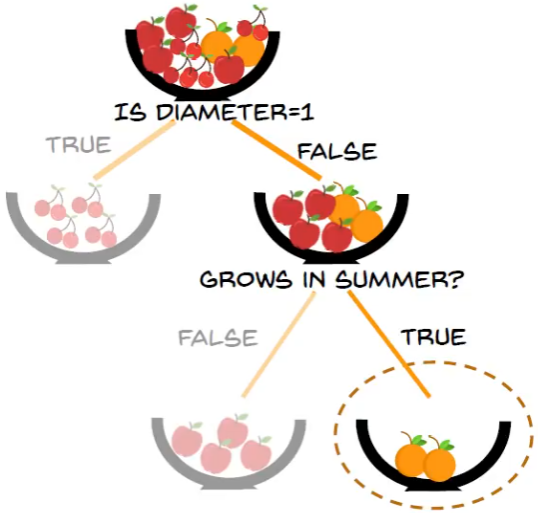
DIAMETER = 3
COLOUR = ORANGE
GROWS IN SUMMER = YES
SHAPE = CIRCLE



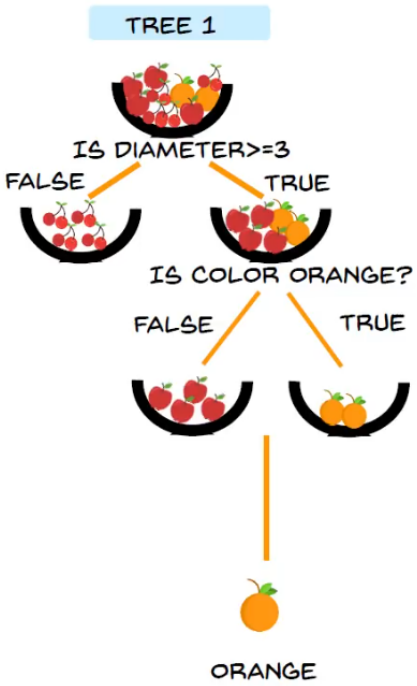
Random Forest - How it works? - Missed values

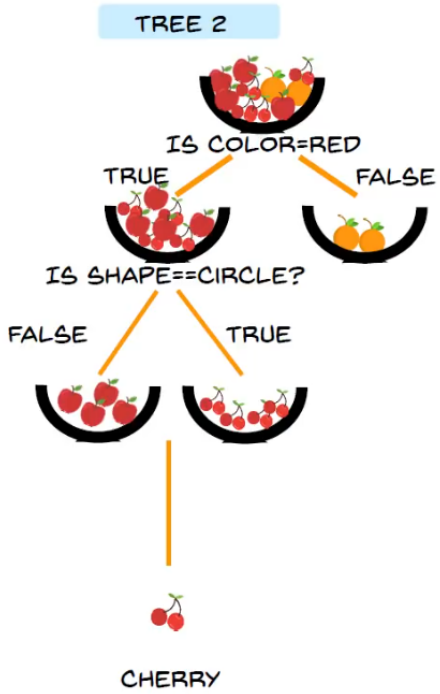
TREE 3 CLASSIFIES IT AS ORANGE

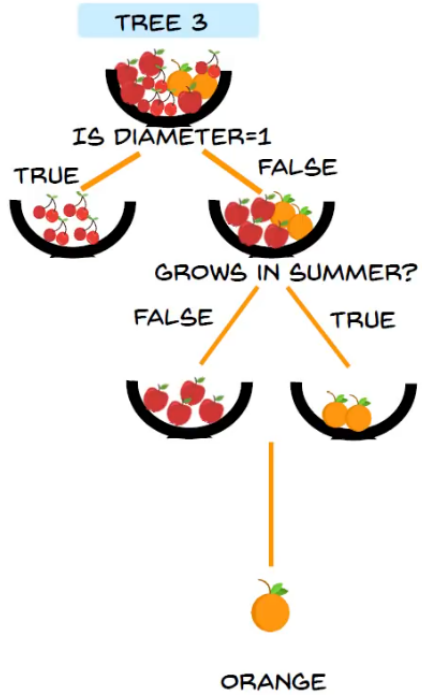
DIAMETER = 3
COLOUR = ORANGE
GROWS IN SUMMER = YES
SHAPE = CIRCLE



Random Forest - How it works? - Missed values

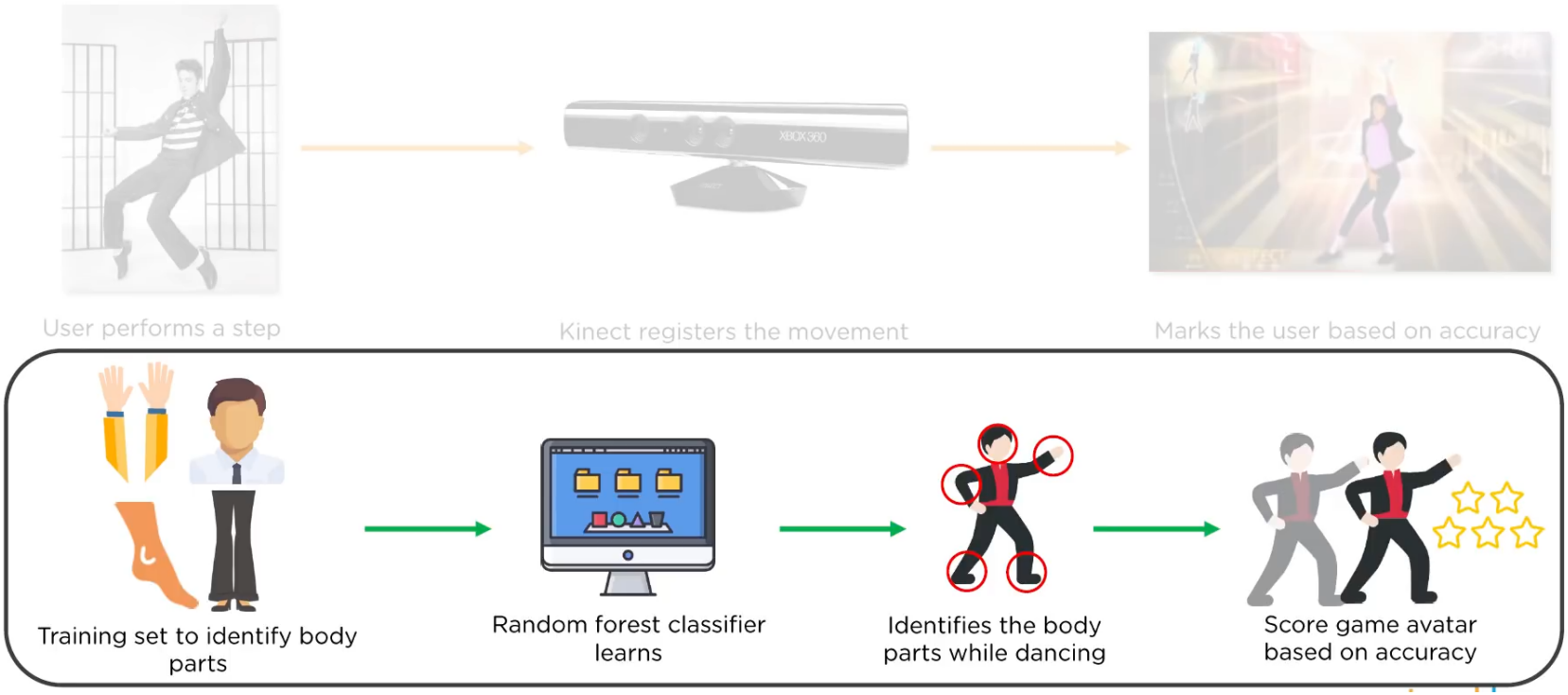






Random Forest - How it works? - Kinect

- How it works?! Demo - Xbox



Random Forest - Important Hyperparameter

- Increasing the predictive power
  - nestimators
  - maxfeatures
  - minsampleleaf
- Increasing the model's speed
  - njobs
  - randomstate
  - oobscore (leave-one-out-cross-validation method)

Random Forest - Advantage Disadvantage

- **Advantage**
  - Random Forest is capable of performing both Classification and Regression tasks
  - Relative importance it assigns to the input features
  - It is capable of handling large datasets with high dimensionality
  - It enhances the accuracy of the model and prevents the overfitting issue (Many trees)
- **Disadvantage**
  - Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.
  - Large number of trees can make the algorithm too slow and ineffective for real-time predictions

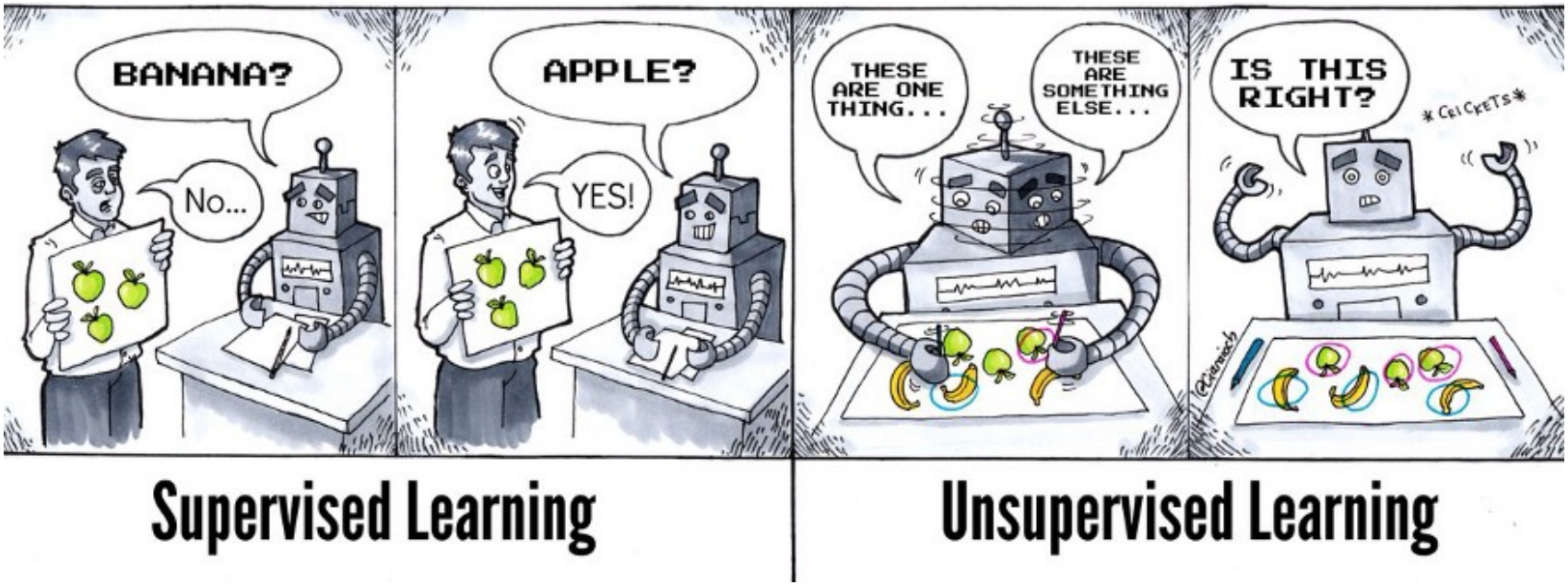
Random Forest - Application

- Kinect - Microsoft
- **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
- **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
- **Land Use:** We can identify the areas of similar land use by this algorithm.
- **Marketing:** Marketing trends can be identified using this algorithm.

Random Forest Vs Decision Tree

- The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the processes of finding the root node and splitting the feature nodes will run randomly
- Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model
- Random Forest can handle missing values
- Random Forest classifier can be modeled for categorical values

Unsupervised Learning



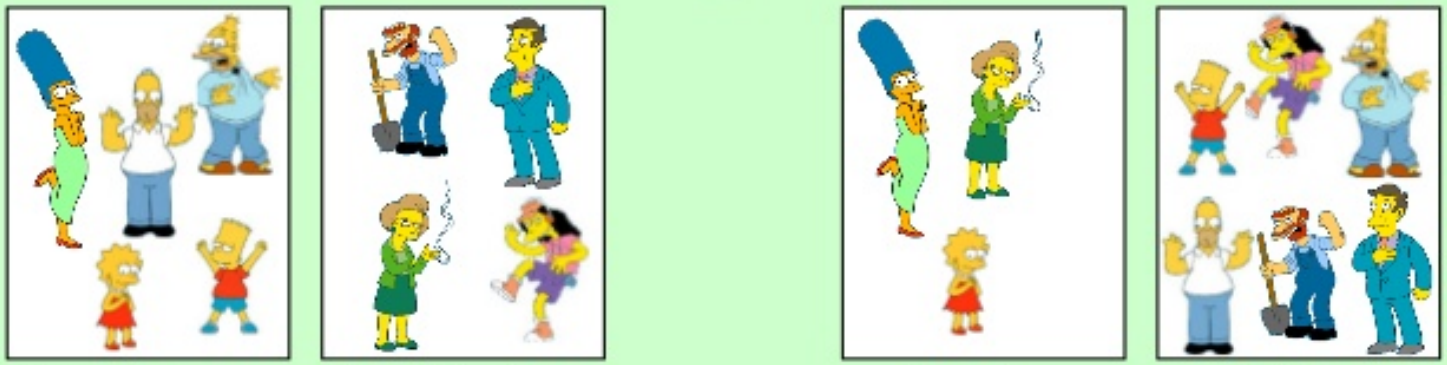
Supervised Learning Vs Unsupervised Learning

- Task Driven — **Data Driven**
- Pre-Categorized Data — **Unlabelled Data**
- Classification - Regression — **Clustering - Association - Dimentionality Reduction**

Algorithm Explanation - K Means Clustering

What is a natural grouping among these objects?

Clustering is subjective



VIT
ellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

K Means Clustering

- What is Clustering?
- How is Clustering an Unsupervised Learning Problem?
- Properties of Clusters
- Applications of Clustering in Real-World Scenarios
- Understanding the Different Evaluation Metrics for Clustering
- What is K-Means Clustering?
- Challenges with K-Means Algorithm
- K-Means ++ to choose initial cluster centroids for K-Means Clustering
- How to choose the Right Number of Clusters in K-Means?

What is Clustering?

- Clustering is the process of dividing the entire data into groups based on the patterns in the data
- Bank - Customers - Credit card offers

Different Algorithms - Clustering

- K-Means Clustering
- Hierarchical Clustering (Divisive and Agglomerative types)
- Fuzzy C Means Algorithm
- Mean-Shift Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Gaussian Mixed Models (GMM) with Expectation-Maximization Clustering

How is Clustering an Unsupervised Learning Problem?

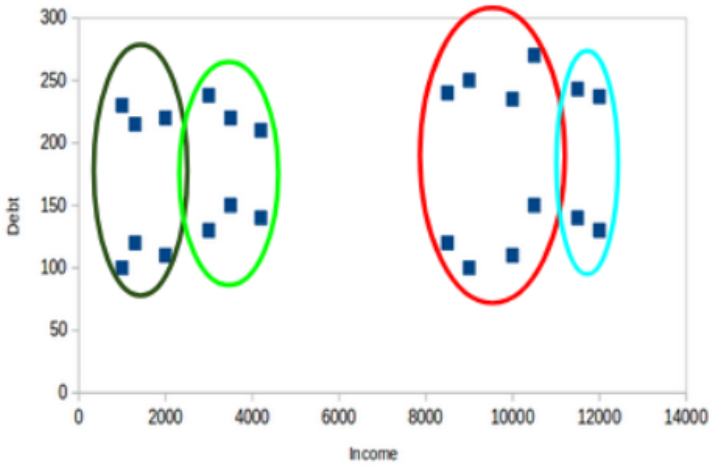
- We do not have a target to predict.
- We look at the data and then try to club similar observations and form different groups.

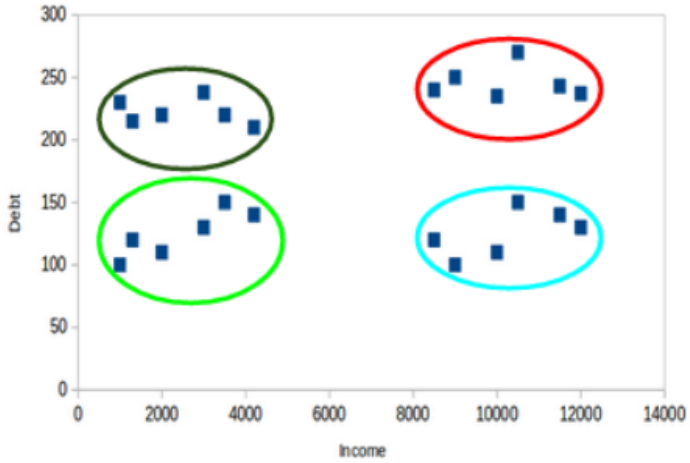
Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
Medium	Tier 1	Supermarket Type1	3735.1380
Medium	Tier 3	Supermarket Type2	443.4228
Medium	Tier 1	Supermarket Type1	2097.2700
NaN	Tier 3	Grocery Store	732.3800
High	Tier 3	Supermarket Type1	994.7052

Loan_ID	Gender	Married	ApplicantIncome	LoanAmount	Loan_Status
LP001002	Male	No	5849	130.0	Y
LP001003	Male	Yes	4583	128.0	N
LP001005	Male	Yes	3000	66.0	Y
LP001006	Male	Yes	2583	120.0	Y
LP001008	Male	No	6000	141.0	Y

Properties of Clusters

- **P1** : All the data points in a cluster should be similar to each other.
- **P2** : The data points from different clusters should be as different as possible.



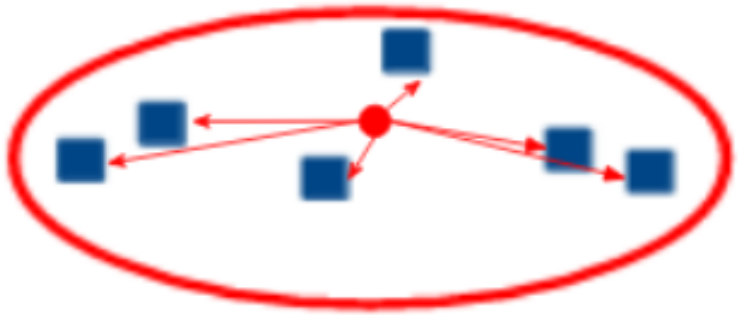


Understanding the Different Evaluation Metrics for Clustering

- Inertia
- Dunn Index

Metrics - Inertia

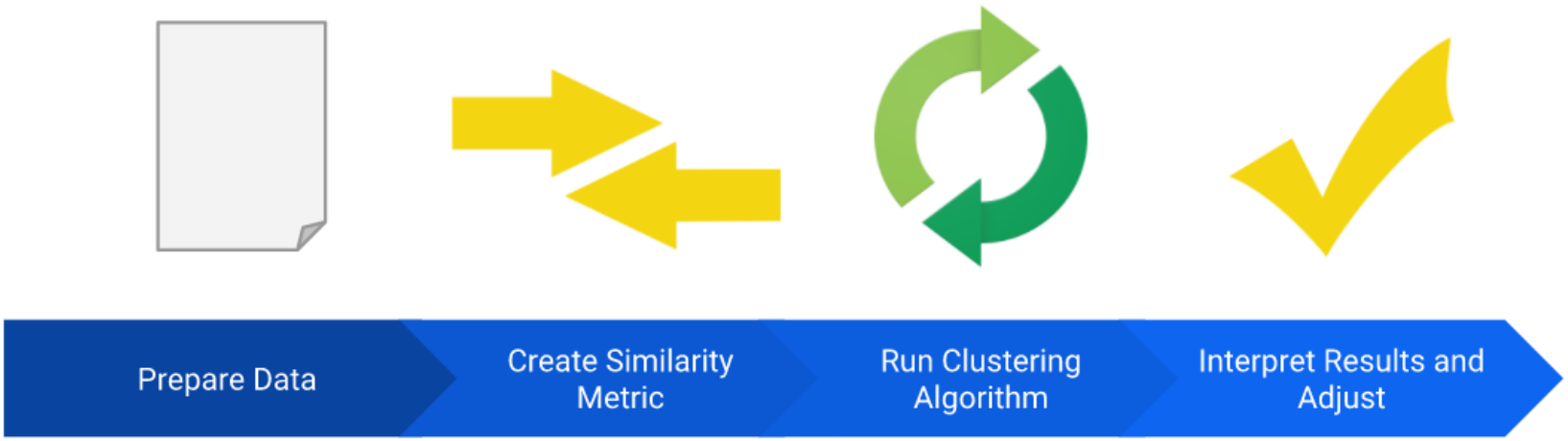
- Calculates the sum of distances of all the points within a cluster from the centroid of that cluster.
- Intra-cluster distance - Inertia tries to minimize
- *Keeping this in mind, we can say that the lesser the inertia value, the better our clusters are.*



Metrics - Dunn Index

- If the distance between the centroid of a cluster and the points in that cluster is small, it means that the points are closer to each other **(Property - 1 / Inertia)**
- Different clusters should be as different from each other as possible **(Property - 2 / ?)**
- More the value of the Dunn index, the better will be the clusters.
- $Dunn - Index = \frac{min(Inter-cluster-distance)}{max(Intra-cluster-distance)}$

So, How Clustering works in general?

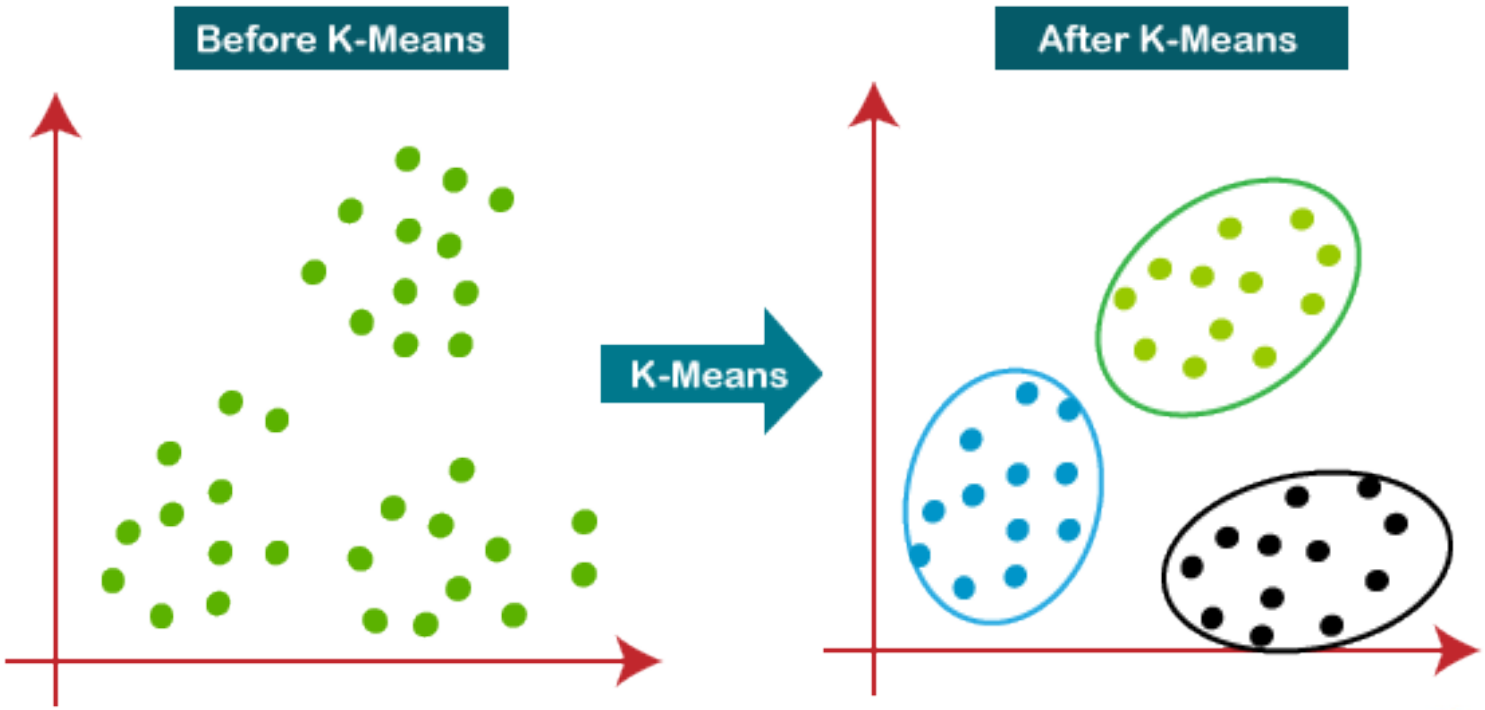


Applications of Clustering

- Netflix has used clustering in implementing movie recommendations for its users.
- Various job search portals use clustering to divide job posting requirements into organized groups which becomes easier for a job-seeker to apply and target for a suitable job
- Satellite imagery can be segmented to find suitable and arable lands for agriculture
- Document clustering is effectively being used in preventing the spread of fake news on Social Media
- Music segmentation - Genre
- Used in image segmentation in bioinformatics where clustering algorithms have proven their worth in detecting cancerous cells from various medical imagery – eliminating the prevalent human errors and other bias

K Means Clustering

- Centroid-based algorithm, or a Distance-based algorithm,
- The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid



So, How K Means Clustering works?

Figure: Consider the data for clustering

So, How K Means Clustering works?

- S1: Choose the number of clusters k
- S2: Select k random points from the data as centroids

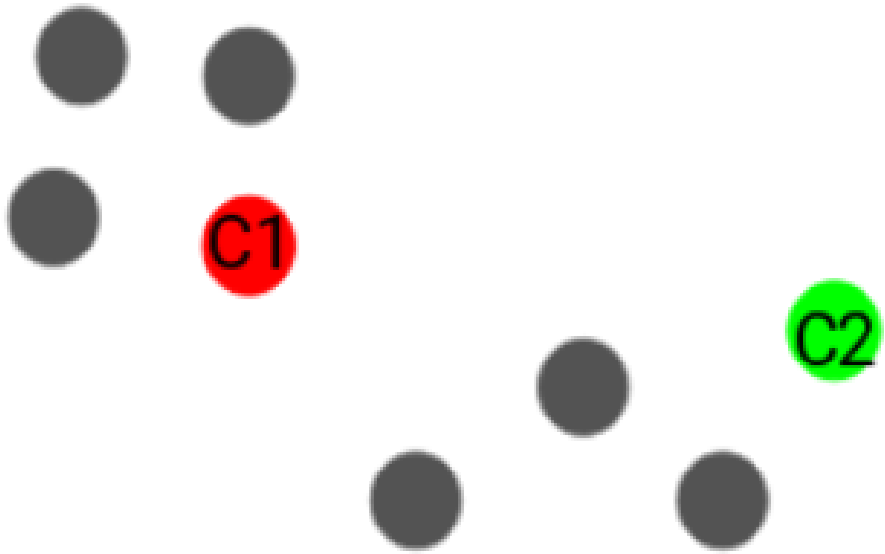


Figure: Red and Green circles represent the centroid for these clusters

So, How K Means Clustering works?

- S3: Assign all the points to the closest cluster centroid

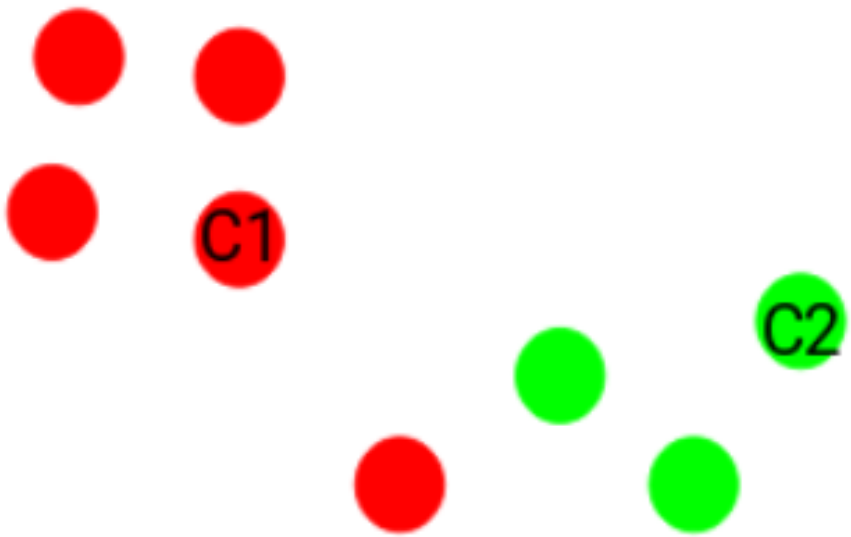


Figure: Points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster

So, How K Means Clustering works?

- S4: Recompute the centroids of newly formed clusters

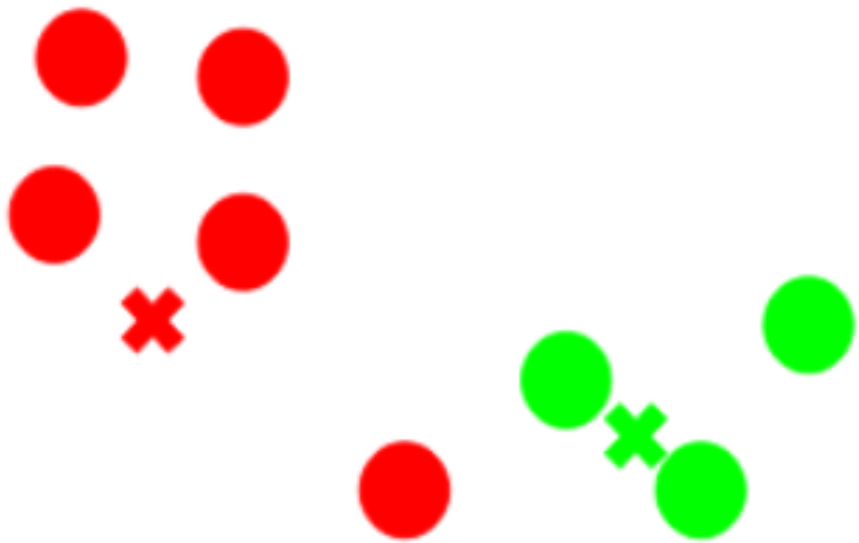


Figure: Red and Green crosses are the new centroids.

So, How K Means Clustering works?

- S5: Repeat steps S3 and S4

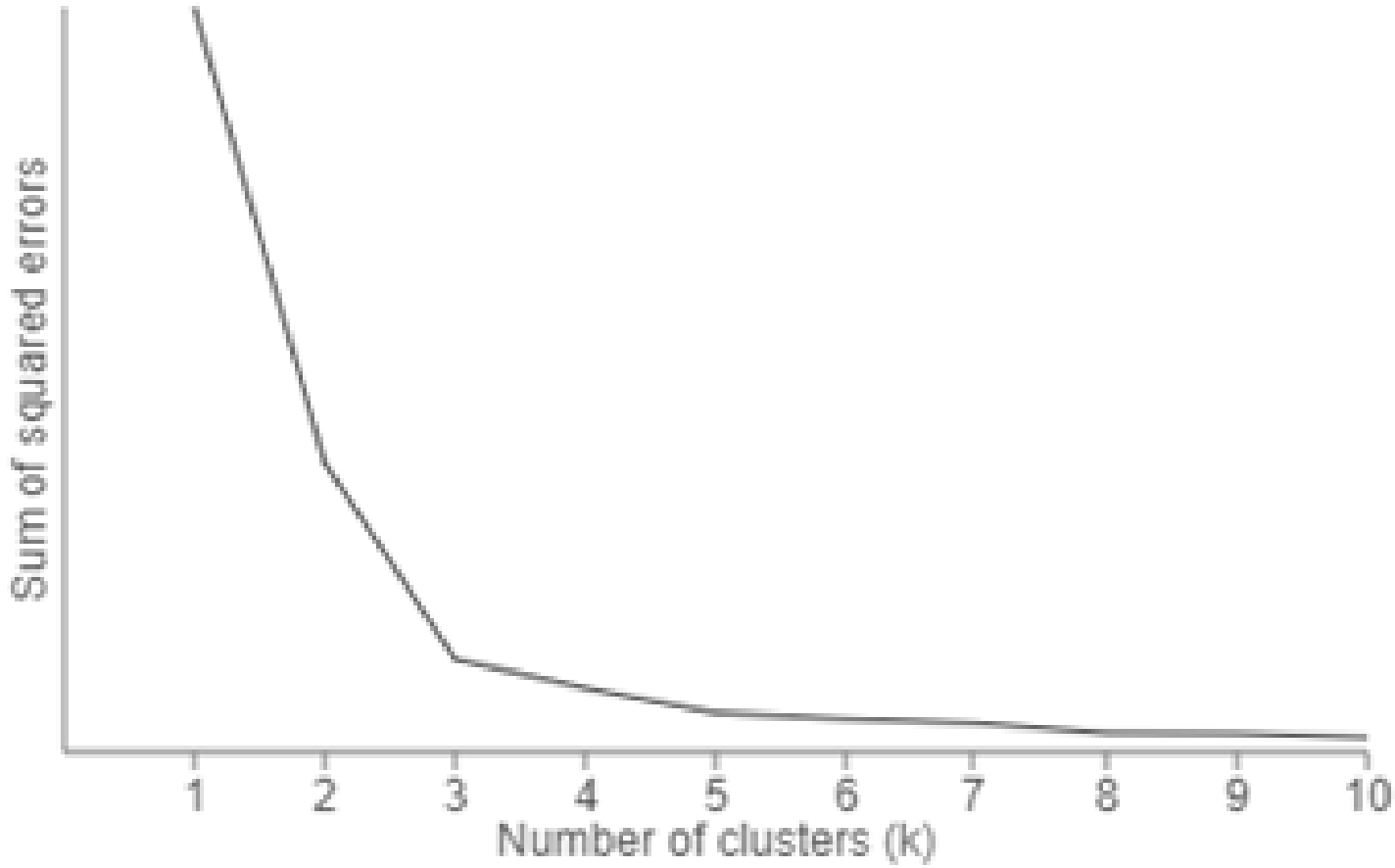
Figure: The step of computing the centroid and assigning all the points to the cluster based on their distance from the centroid is a single iteration

Stopping criteria for K means clustering

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

Decide value of **K** value - Elbow Method

- K Vs Percentage of Variance Explained



VIT
ellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Advantage & Disadvantage of K Means Clustering

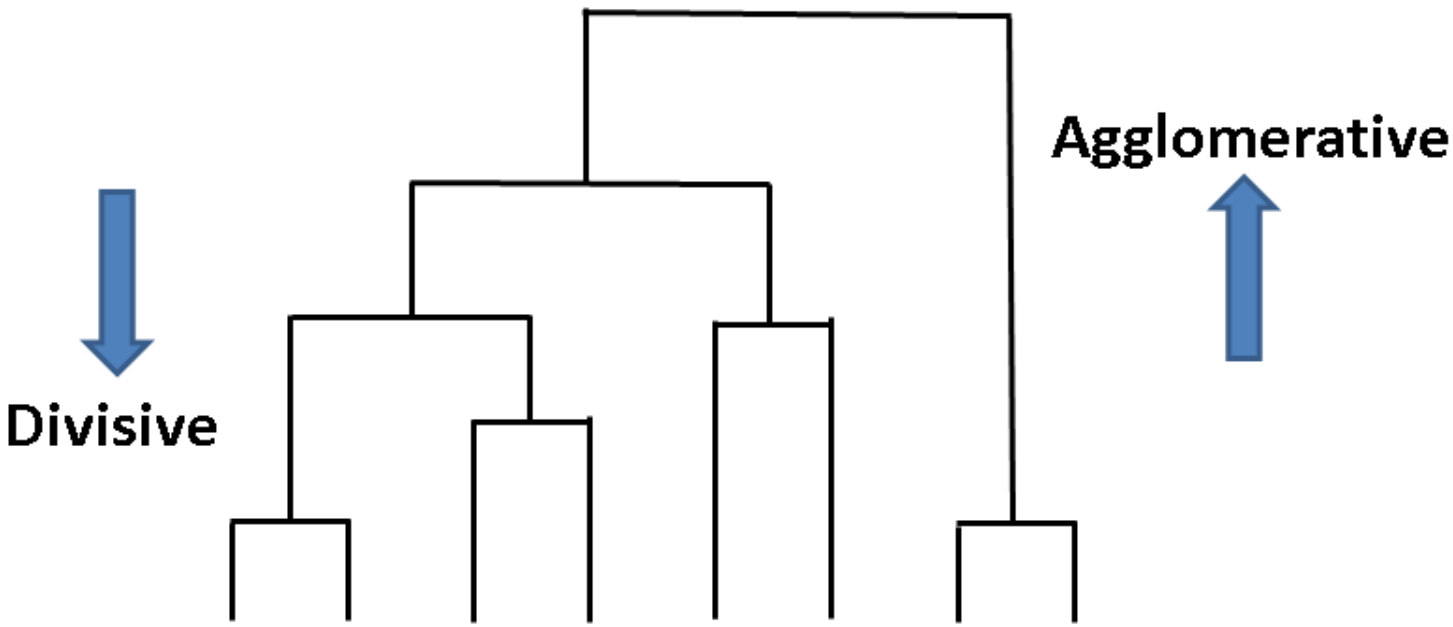
- **Advantage**
  - Relatively simple to implement
  - Scales to large data sets
  - Easily adapts to new examples
- **Disadvantage**
  - Choosing k manually.
  - Being dependent on initial values
  - Clustering outliers

Application of K Means Clustering

- Academic performance
- Cyber-profiling criminals
- Search engines
- Rideshare data analysis
- Identifying crime localities

Hierarchical Clustering

- Connectivity-Based Clustering
- Based on the direction of progress, i.e., whether it is the top-down or bottom-up flow of creating clusters
- Divisive Approach and the Agglomerative Approach



VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Hierarchical Clustering Vs K Means Clustering

- We have to decide the number of clusters at the beginning of the algorithm. Ideally, we would not know how many clusters should we have, in the beginning of the algorithm and hence it is a challenge with K-means.
- It takes away the problem of having to pre-define the number of clusters.

So, How Hierarchical Clustering works?

Figure: We want to cluster them into groups

So, How Hierarchical Clustering works?

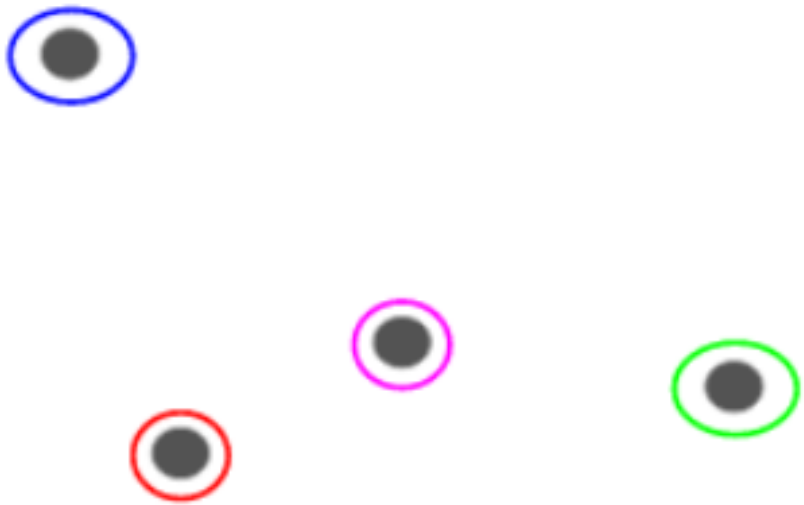


Figure: We can assign each of these points to a separate cluster

So, How Hierarchical Clustering works?

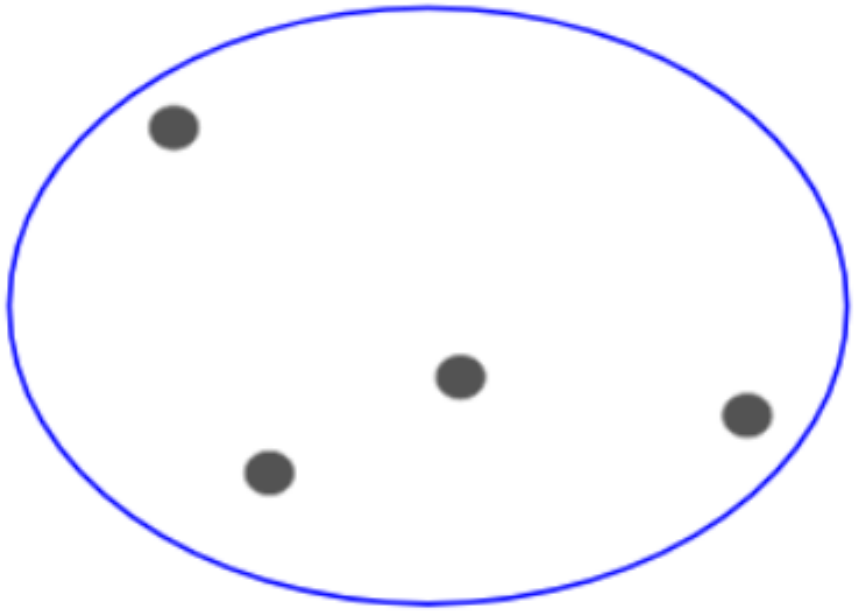


Figure: Based on the similarity of these clusters, we can combine the most similar clusters together and repeat this process until only a single cluster is left

So lets take some example, How Hierarchical Clustering works?

Figure: Suppose a teacher wants to divide her students into different groups

So lets take some example, How Hierarchical Clustering works?

Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

Figure: Lets take the sample of marks

So lets take some example, How Hierarchical Clustering works?

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Figure: Creating Proximity Matrix - diagonal elements of this matrix will always be 0 - Euclidean distance

So lets take some example, How Hierarchical Clustering works?

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Figure: we will look at the smallest distance in the proximity matrix and merge the points with the smallest distance

So lets take some example, How Hierarchical Clustering works?

Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

Figure: Smallest distance is 3 and hence we will merge point 1 and 2

So lets take some example, How Hierarchical Clustering works?

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

Figure: we will again calculate the proximity matrix for these clusters

So lets take some example, How Hierarchical Clustering works?

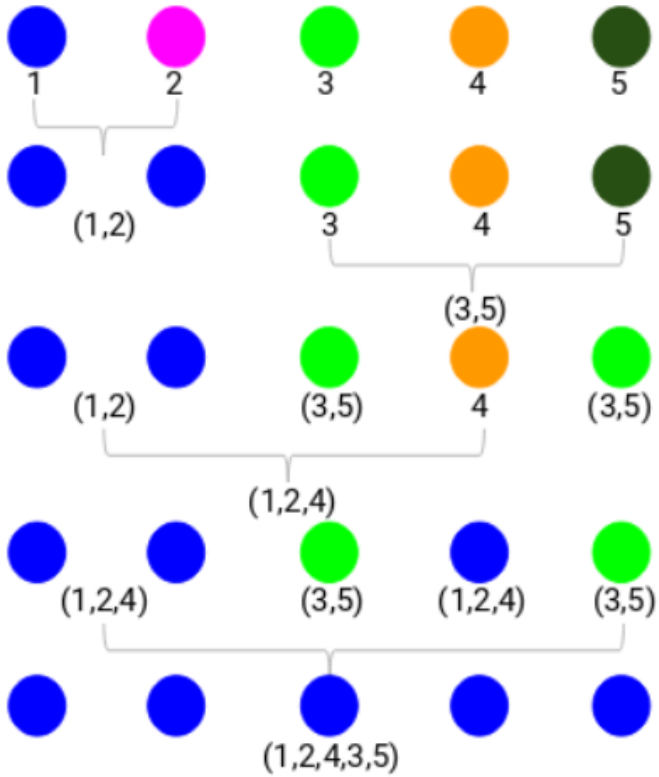


Figure: We started with 5 clusters and finally have a single cluster. This is how agglomerative hierarchical clustering works

How we Choose the Number of Clusters in Hierarchical Clustering?

- Dendrogram - tree-like diagram that records the sequences of merges or splits

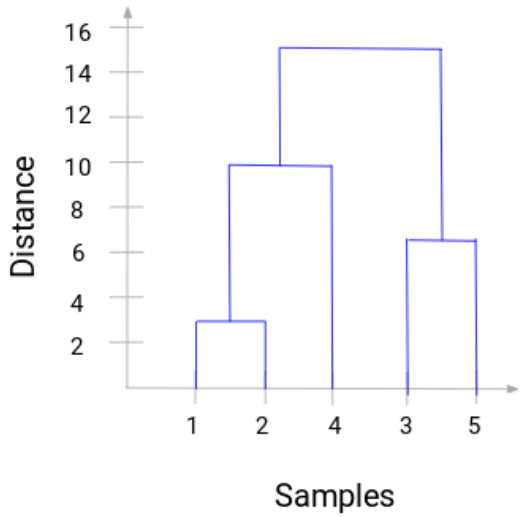


Figure: Whenever two clusters are merged, we will join them in this dendrogram and the height of the join will be the distance between these points. More the distance of the vertical lines in the dendrogram, more the distance between those clusters

How we Choose the Number of Clusters in Hierarchical Clustering?

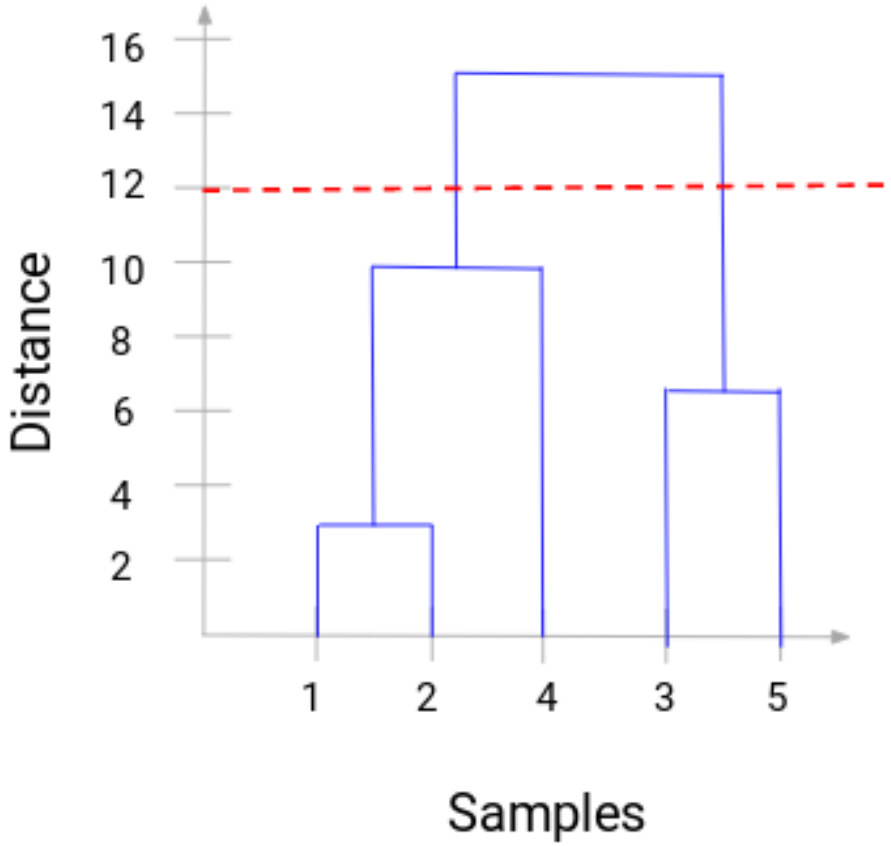


Figure: The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold.

So, How Agglomerative Hierarchical Clustering works?

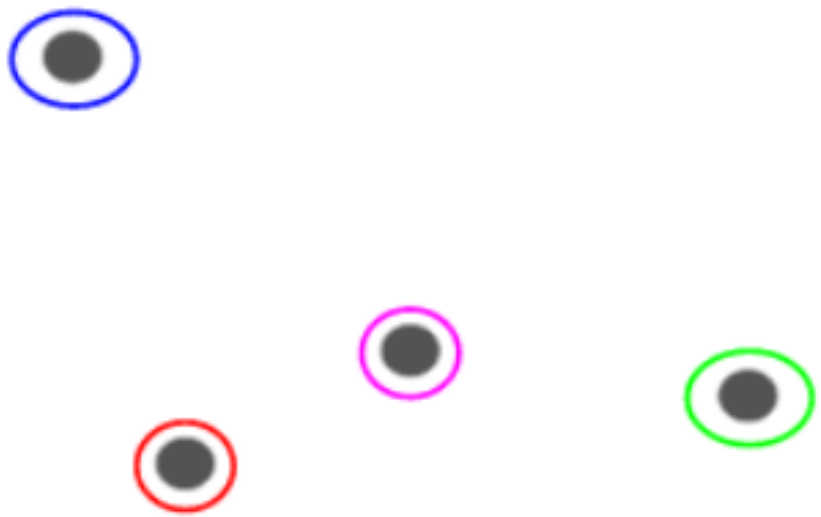


Figure: We will assign each of these points to a cluster and hence will have 4 clusters in the beginning

So, How Agglomerative Hierarchical Clustering works?

- Additive hierarchical clustering

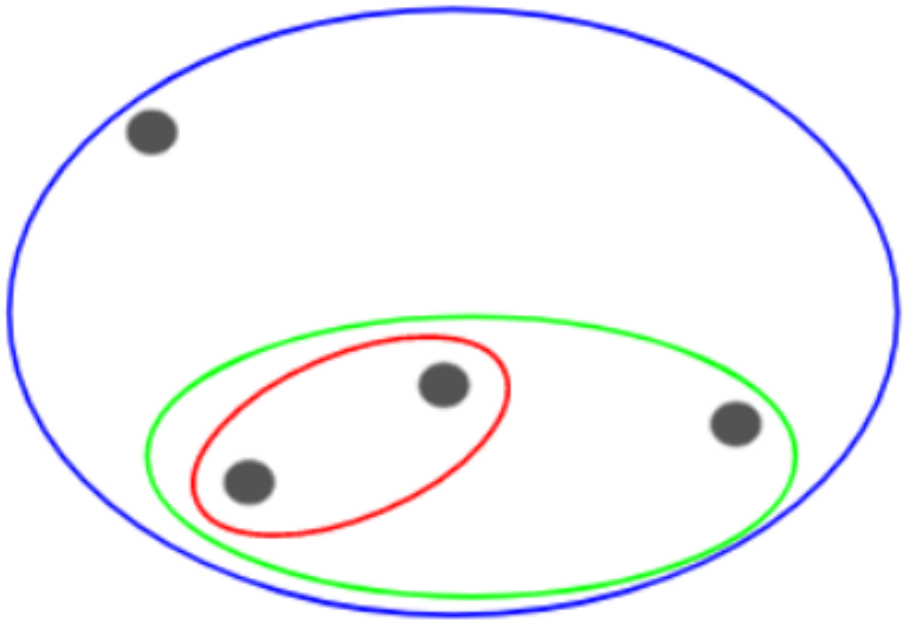


Figure: At each iteration, we merge the closest pair of clusters and repeat this step until only a single cluster is left.

So, How Divisive Hierarchical Clustering works?

- Opposite way

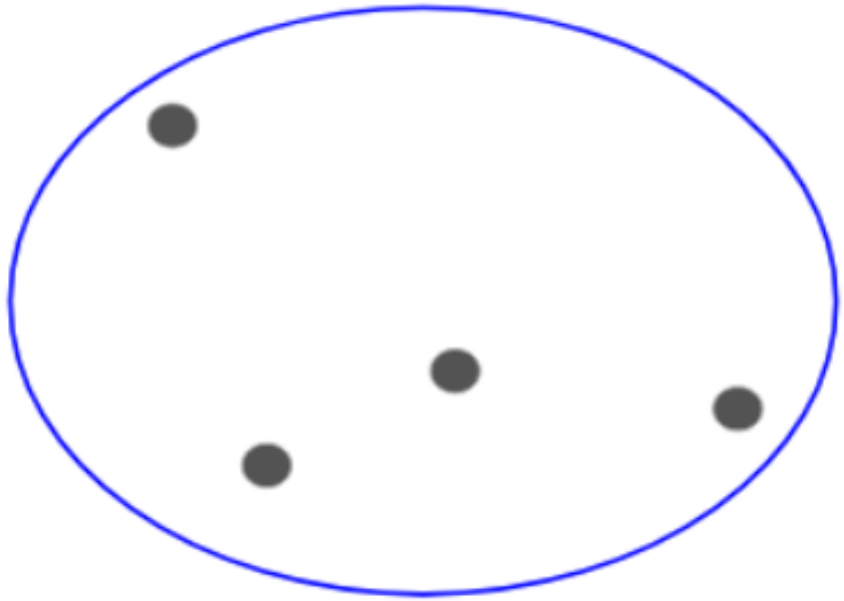


Figure: All these points will belong to the same cluster at the beginning

So, How Divisive Hierarchical Clustering works?

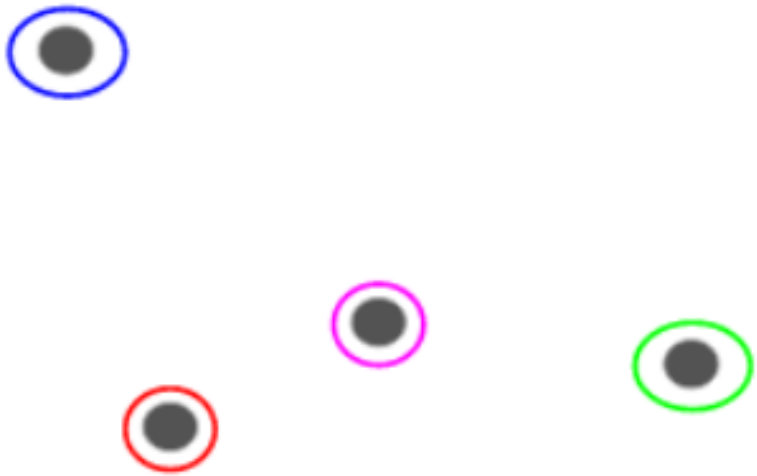
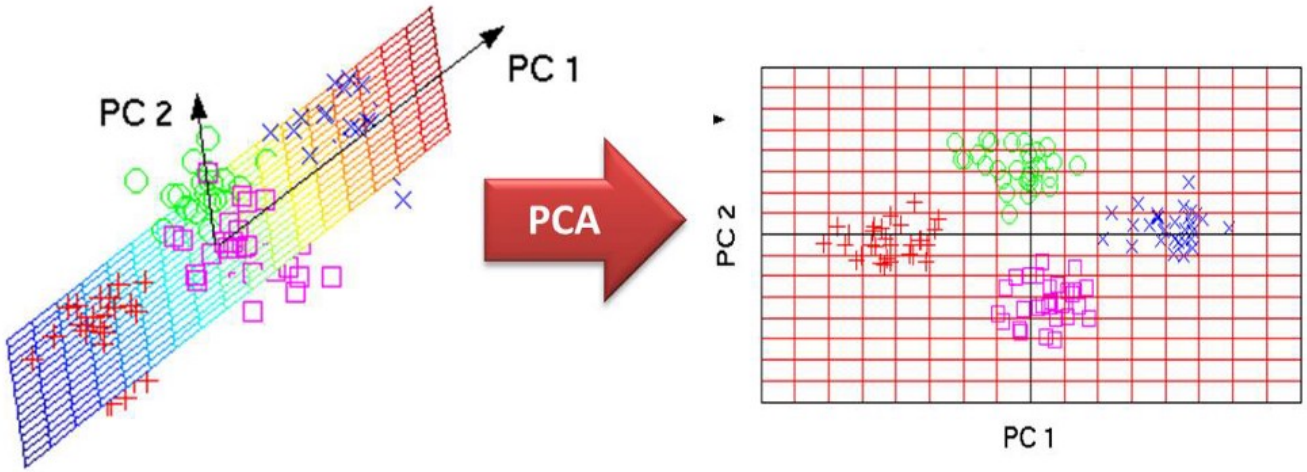


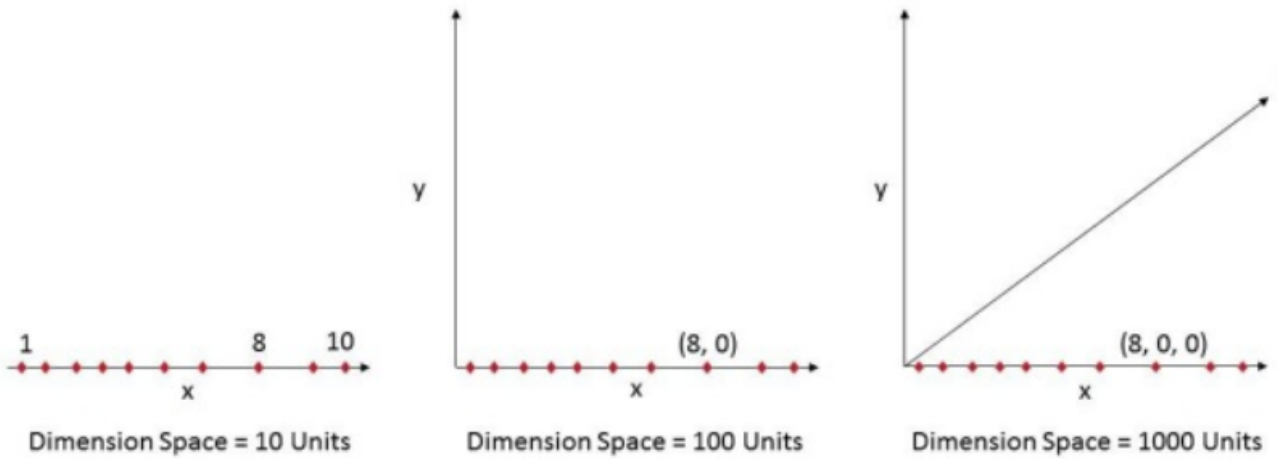
Figure: At each iteration, we split the farthest point in the cluster and repeat this process until each cluster only contains a single point

Dimensionality Reduction & Principal Component Analysis



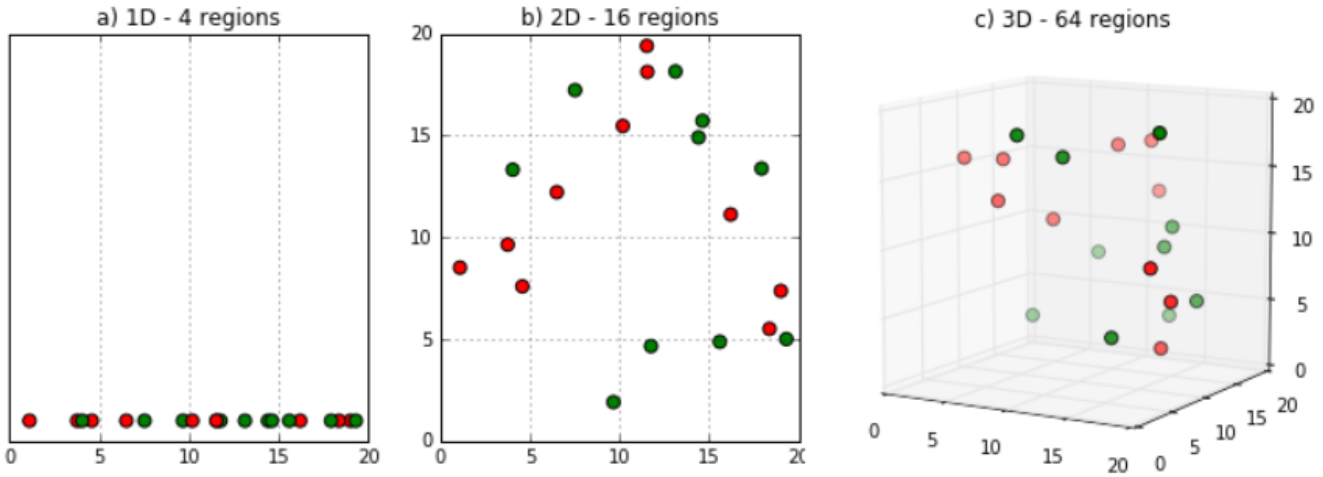
Curse of Dimensionality

- As the number of features or dimensions grows, the amount of data we need to generalize accurately grows exponentially



Curse of Dimensionality

- High resolution images 1280 720 = 921,600 pixels i.e. 921600 dimensions
- Ex: Exam writing - answers

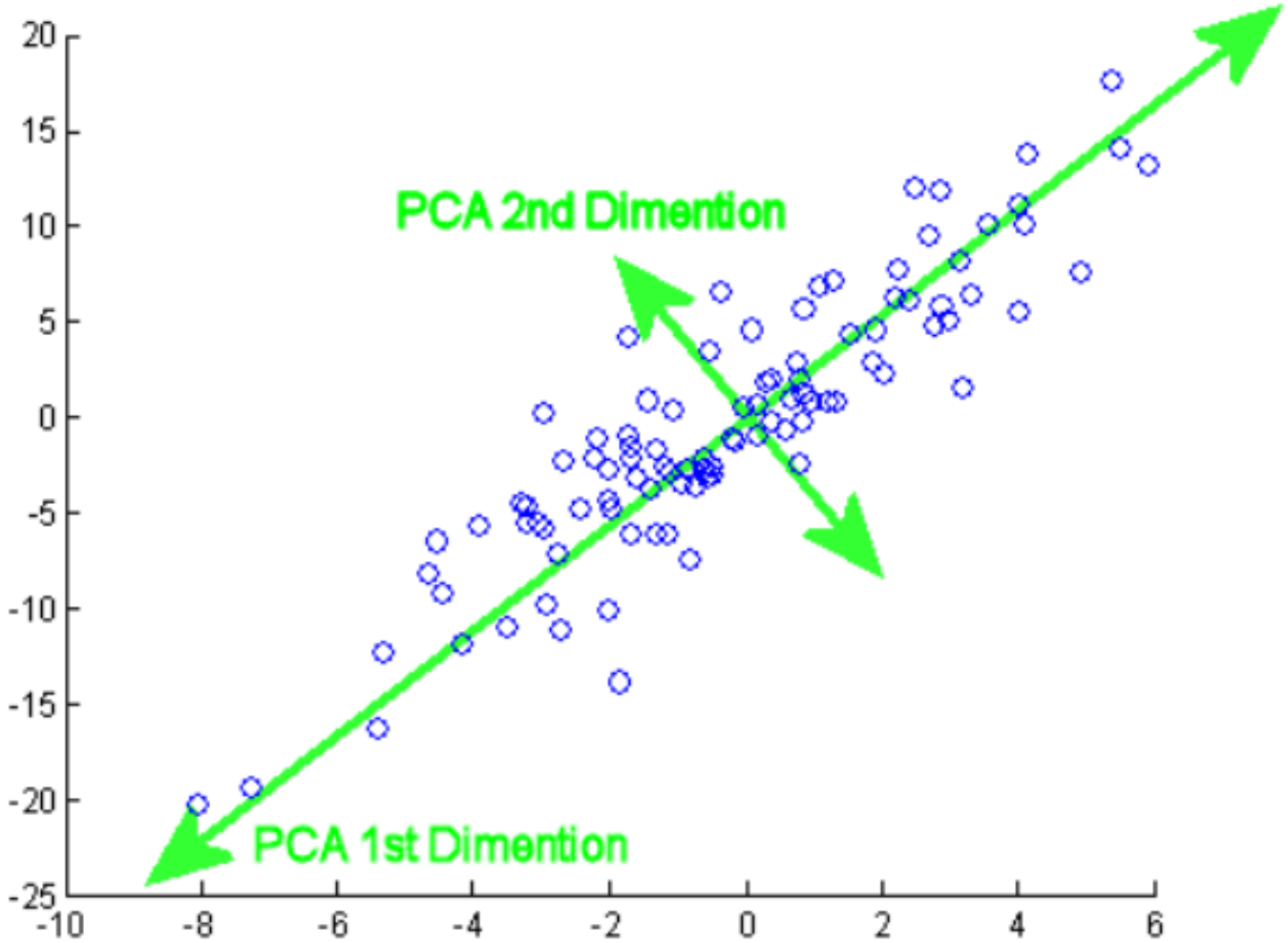


- Two options to reduce dimensionality
  - Feature elimination : remove some features directly
  - Feature extraction : keep the important fraction of all the features

Principal Component Analysis

- Unsupervised learning
- Dimentionality Reduction technique
- We have many features with high multicollinearity
- We have too many features that cause the algorithm to run very slowly
- identifies important relationships in data
- transforms the existing data based on these relationships
- quantifies the importance of these relationships so we can keep the most important relationships and drop the others
- Data in higher dimension can be explained by few principal components
- Principal components are basically vectors that are linearly uncorrelated and have a variance with in data

Principal Component Analysis



PCA - steps

- Take the n-dimensional dataset
- Standardize the dataset
- Calculate the covariance matrix for the features in the dataset.
- Calculate the eigenvalues and eigenvectors for the covariance matrix.
- Sort eigenvalues and their corresponding eigenvectors.
- Pick k eigenvalues and form a matrix of eigenvectors.
- Transform the original matrix.

PCA - Advantage

- PCA reduces the dimensionality without losing information from any features.
- Reduce storage space needed to store data
- Speed up the learning algorithm (with lower dimension)
- Help visualize data with high dimensionality (after reducing the dimension to 2 or 3)

PCA - Disadvantage / Weakness

- Tends to highly affected by outliers and missing values
- Randomized PCA, Sparse PCA...
- Standardize data - mandatory
- PCA prevents interpretation of the original features, as well as their impact because eigenvectors are not meaningful.

PCA - Applications

- Dimensionality reduction
- Finding patterns in the data
- Find essential attributes/variables(Feature selection within high-dimensional data)
- Noise filtering.

Some key jargon of Machine Learning

- Multicollinearity
- Overfitting
- Under fitting
- Bias - Variance Trade off
- Generalisation and Regularization
- Hyperparameter

Not convinced, need more clarity for ML ?

I always believe, if any concept can be conveyed through animation, it wont forget easily.
This is the link: Machine Learning @Simplilearn.

Regression - Exercise

- **Yrs of Experience Vs Salary dataset** (Simple variable)
- **Boston Housing dataset** (Multiple variable)

Classification - Exercise

- **Iris dataset** (Simple variable)
- **Breast cancer dataset** (Multiple variable)
- **Loan Prediction dataset** (Multiple variable)

Trending domains for Machine Learning

- Healthcare Diagnosis
- Education
- Speech Processing (Natural Language Processing)
- Agriculture
- Digital Marketing
- Robotics

Other important tools / technology for Machine Learning and Data Science

- MATLAB
- WEKA
- Google's Cloud AutoML
- Microsoft Azure Machine Learning
- Tableau
- Power BI

Some best course for Machine Learning across the globe

- Python (Dr. Charles Severance) & Coursera
- Springboard - Machine Learning Career Track
- Coursera - Machine Learning - Stanford University
- Udemy - Machine Learning A-Z (Python and R in Data Science)
- EdX - Data Science - Machine Learning
- Machine Learning Mastery - Machine Learning track
- Python for Everybody (youtube) - Charles Severance
- Datacamp - Machine Learning Scientist with Python - Free subscription!

Tips to improve Machine Learning coding & knowledge

- Open GitHub repository & and start coding from scratch for different dataset Github link - Details about Github - Hit me!
- Try to participate, competitions in Kaggle website for major attractions. Eg: Abhishek Thakur (Approaching (Almost) Any Machine Learning Problem) - Kaggle link
- Try to follow worlds top scientist, R & D, some reowned personalities in the field of Data science, Machine Learning and Deep Learning for their work and tips - Linkedin link

Deep Learning

- Subfield of Machine Learning
- Inspired by ANN
- BlackBox
- Tries to mimic the function of inner layers of human brain
- Some popular Deep Learning models are,
  - Multilayer Perceptron Neural Network (MLPNN)
  - Convolution Neural Network (CNN)
  - Recurrent Neural Network (RNN)
  - Long Short-Term Memory (LSTM)
  - Generative Adversarial Network (GAN)
- Deep Learning - Animation @Simplilearn

mail me: er.anandprem@gmail.com
ring me: +91 73586 79961
follow me: Linkedin
Website: tango-learning

Learning gives Creativity, Creativity leads to Thinking, Thinking provides Knowledge, and Knowledge makes you Great - Dr APJ Abdul Kalam