Data Science& Business Analytics

Internship

CH.VAMSHI(4511-18-733-017)

Contents

- Predictions using Supervised Machine Learning.
- Prediction using Unsupervised Machine Learning.
- Exploratory Data Analysis.
- Prediction Using Decision Tree Algorithm.

Introduction

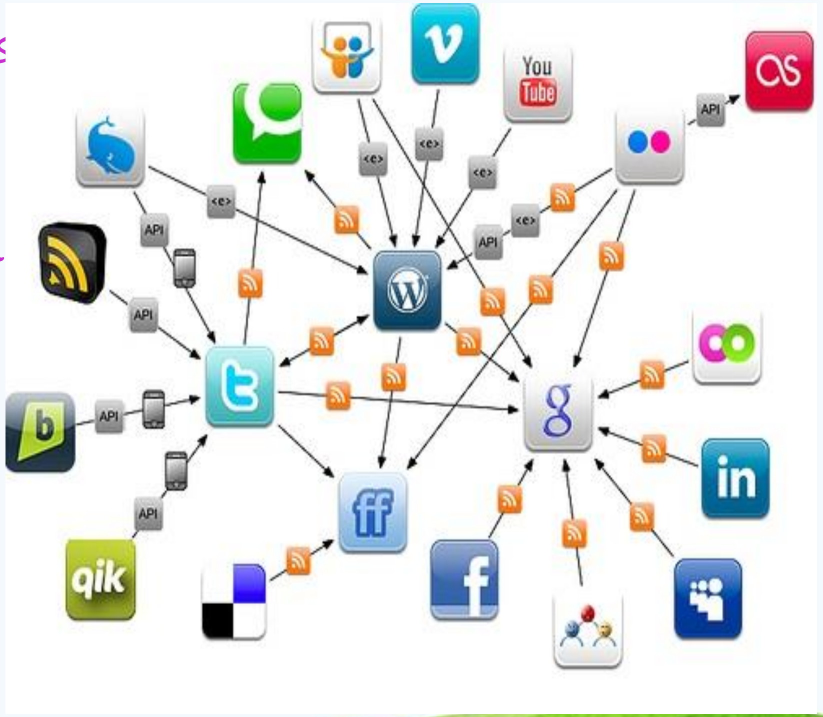
- In the past,we used to have data in structured format but now days the volume of the data is increasing, so there is massive amount of data is there but it is unstructured format.

- To deal with that unstructured data we need data science techniques.

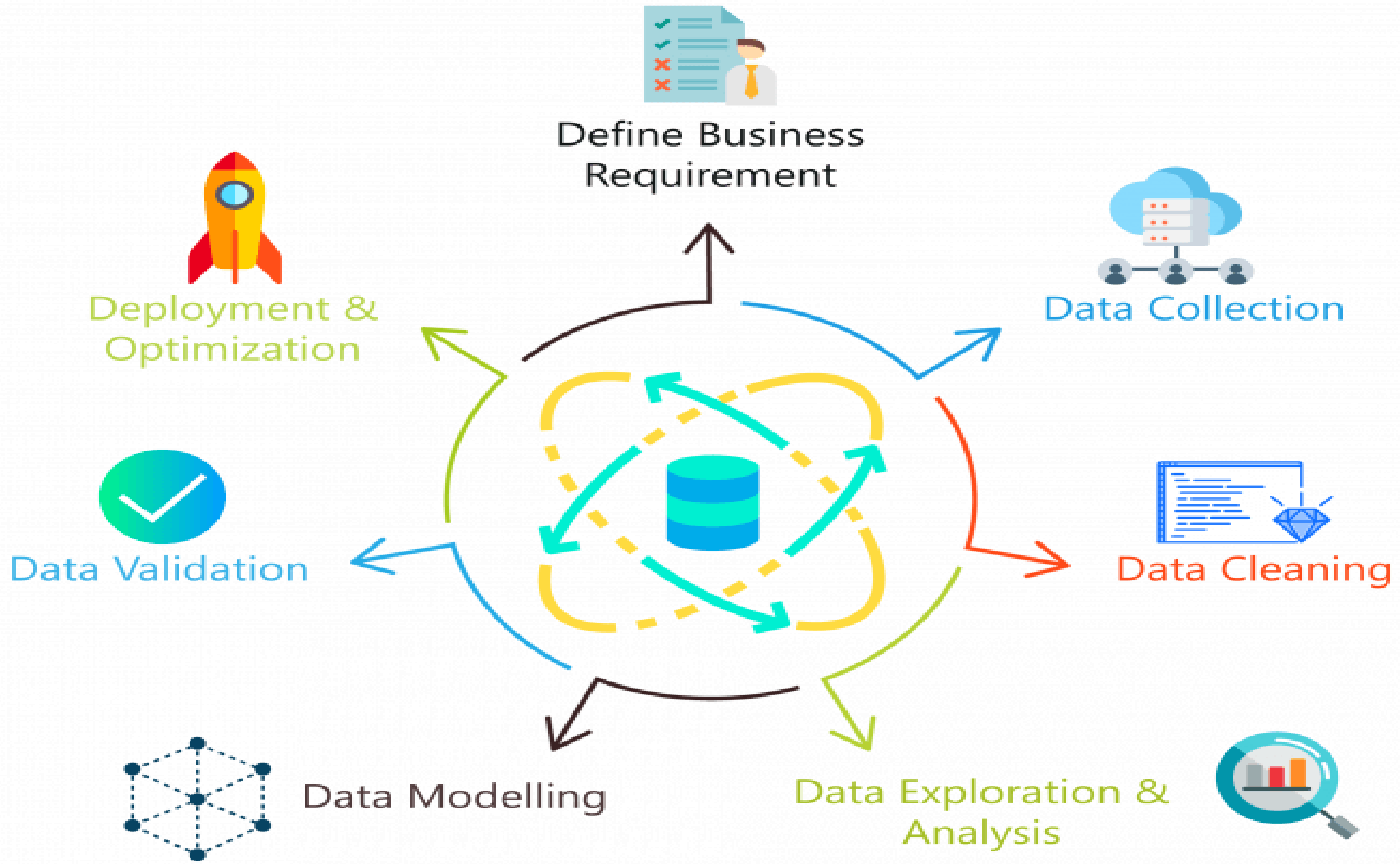
- That data can be used to get the proper insights and hidden trends from them.

- To define data science is Process of finding meaningfull information from massive amount of data.

- Data Science is mixture of various tools and algorithms in machine learning to discover hidden patterns from unstructured data.



Data Science Life Cycle



Define Business Requirement

Data Collection

Data Cleaning

Data Exploration & Analysis

Data Modelling

Data Validation

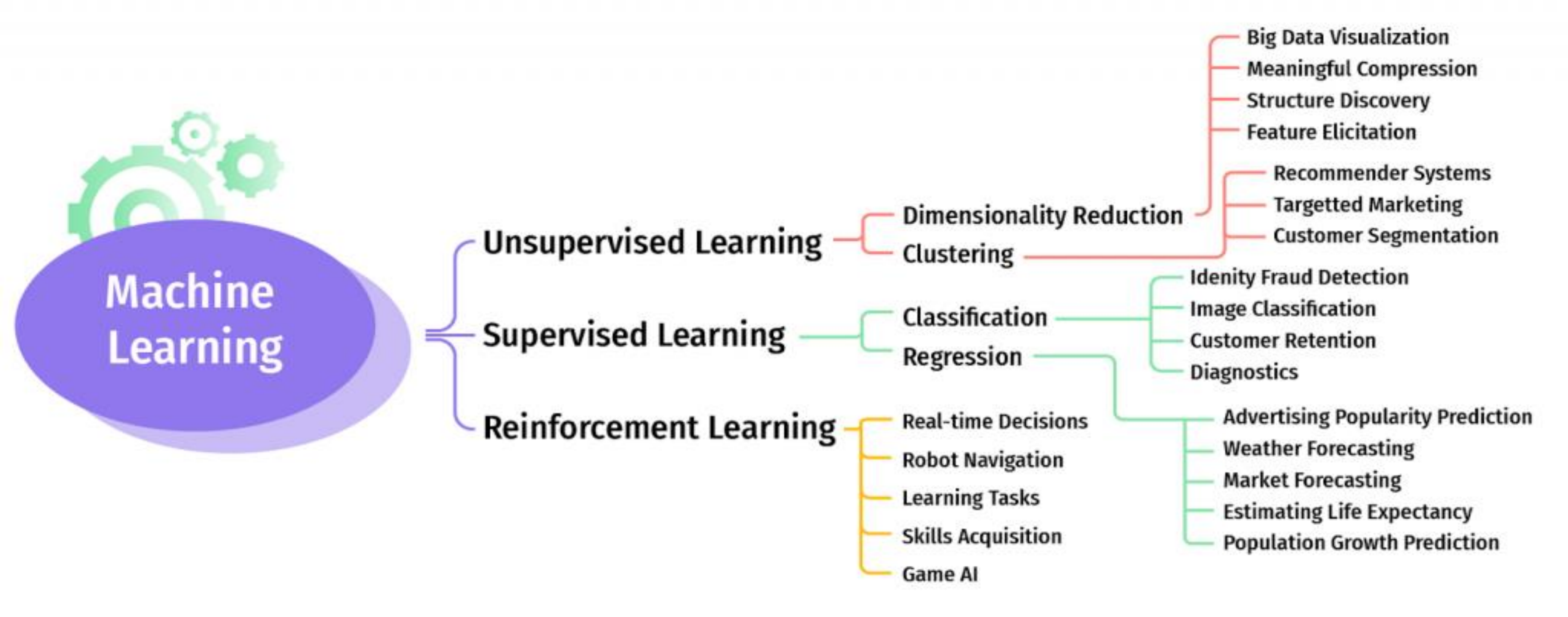
Deployment & Optimization

Machine Learning

“ It is field of study that gives Computers (or) Machines can Learn Without Being Explicity”- Author Somuel

(OR)

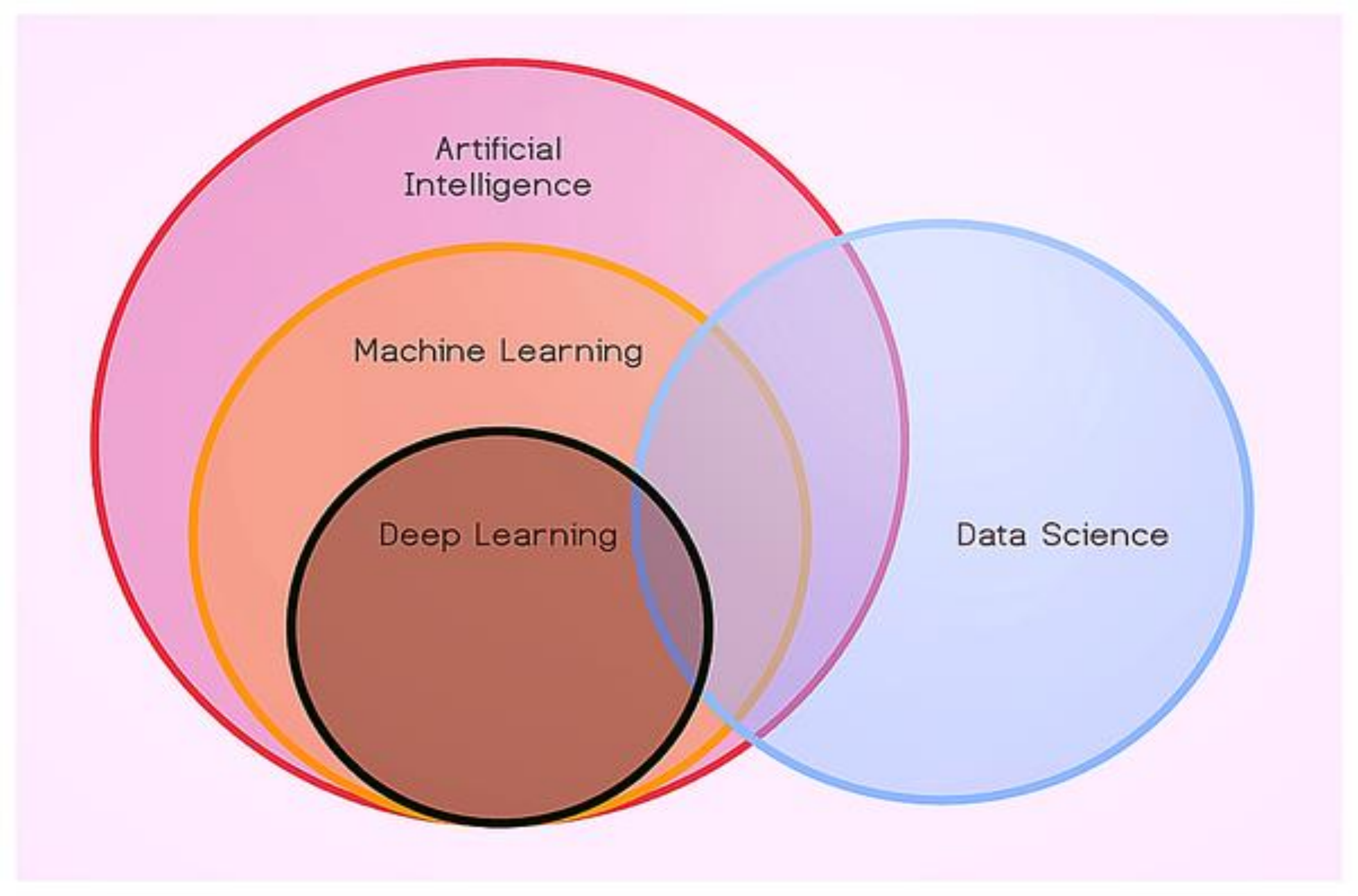
“ A Computer programm is said to learn from the Experience E with Respect to some task T, and Performance measure P,if it Performance at task in T as measured by P,improve the Experience E” – Tom Mitchell



Machine Learning

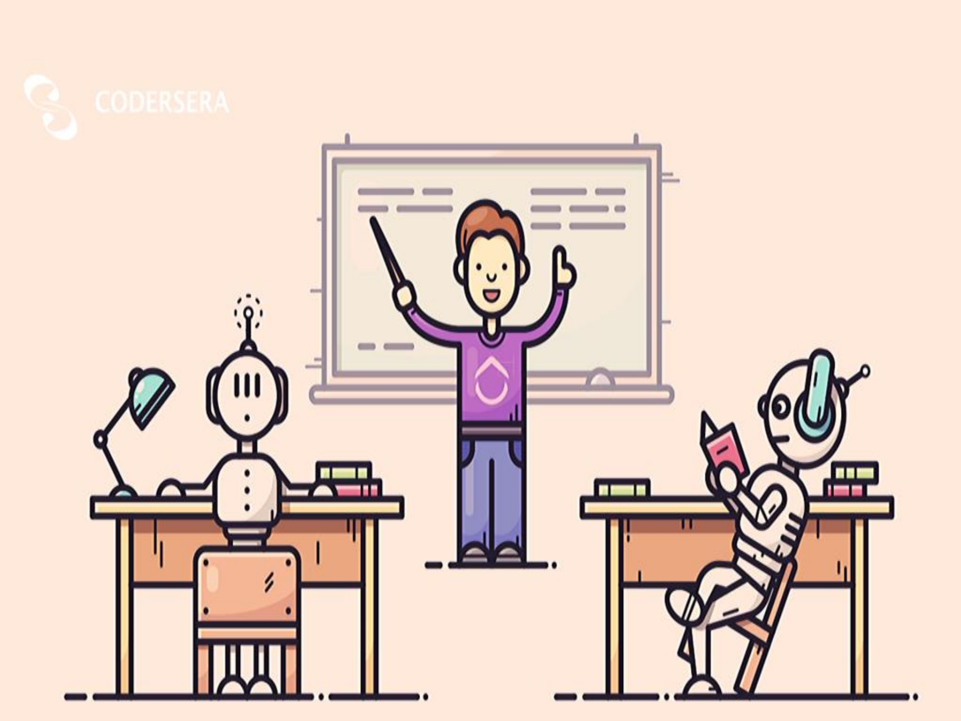
- Unsupervised Learning
  - Dimensionality Reduction
    - Big Data Visualization
    - Meaningful Compression
    - Structure Discovery
    - Feature Elicitation
  - Clustering
    - Recommender Systems
    - Targetted Marketing
    - Customer Segmentation
- Supervised Learning
  - Classification
    - Idenity Fraud Detection
    - Image Classification
    - Customer Retention
    - Diagnostics
  - Regression
    - Advertising Popularity Prediction
    - Weather Forecasting
    - Market Forecasting
    - Estimating Life Expectancy
    - Population Growth Prediction
- Reinforcement Learning
  - Real-time Decisions
  - Robot Navigation
  - Learning Tasks
  - Skills Acquisition
  - Game AI

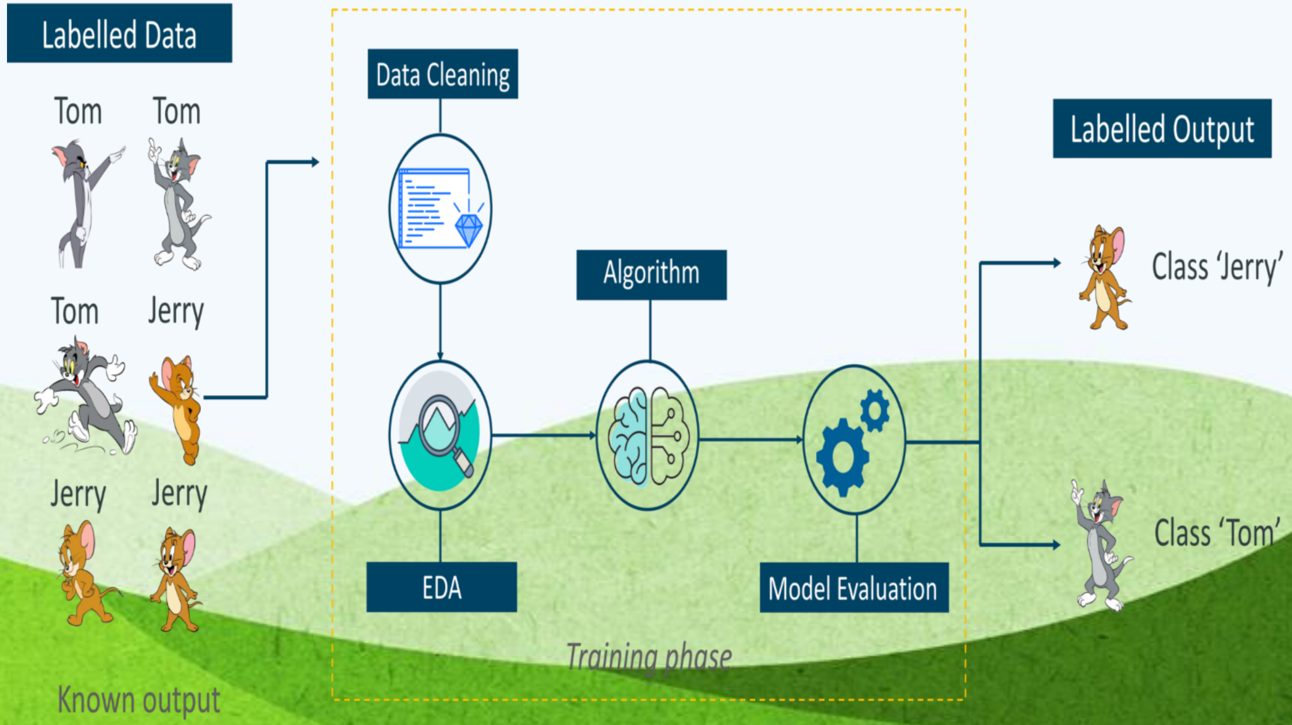
Relationship B/w Data Science and ML



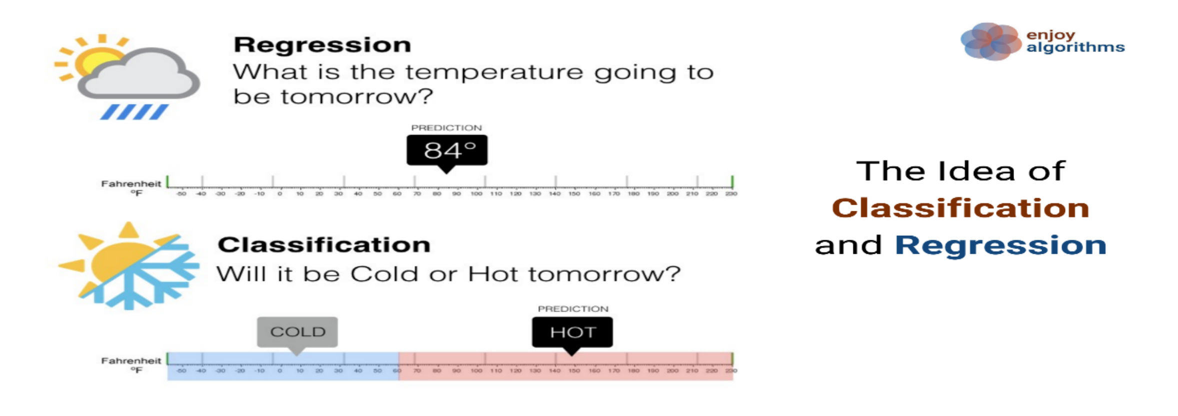
1.Prediction Using Supervised Machine Learning

- Supervised Machine learning is one of the type of Machine Learning.
- Supervised Machine Learning: when we are training a machine with Labeled data.After that we are giving new data, that machine can be predict by using training labeled data and Produce output.
- Supervised Machine Learning Algorithms gives always correct answers.





Types of Supervised Machine Learning:


Regression
What is the temperature going to be tomorrow?
PREDICTION
84°
Fahrenheit °F
Classification
Will it be Cold or Hot tomorrow?
COLD
PREDICTION
HOT
Fahrenheit °F
enjoy algorithms
The Idea of Classification and Regression

There are Two types of supervised ML

1.Regression

2.Classifiaction

Regression:

- ➢ It is used to find the Relationship between dependent variable and independent variable.
- ➢ It is always gives the output in the form of Numerical Values (or) Continues Values.

Types of Regression:

1.Simple Linear Regression.

2.Multi Variable Regression.

Regression Algorithms:

1.Linear Regression Algorithm.

2.Ploynomial Regression Algorithm.

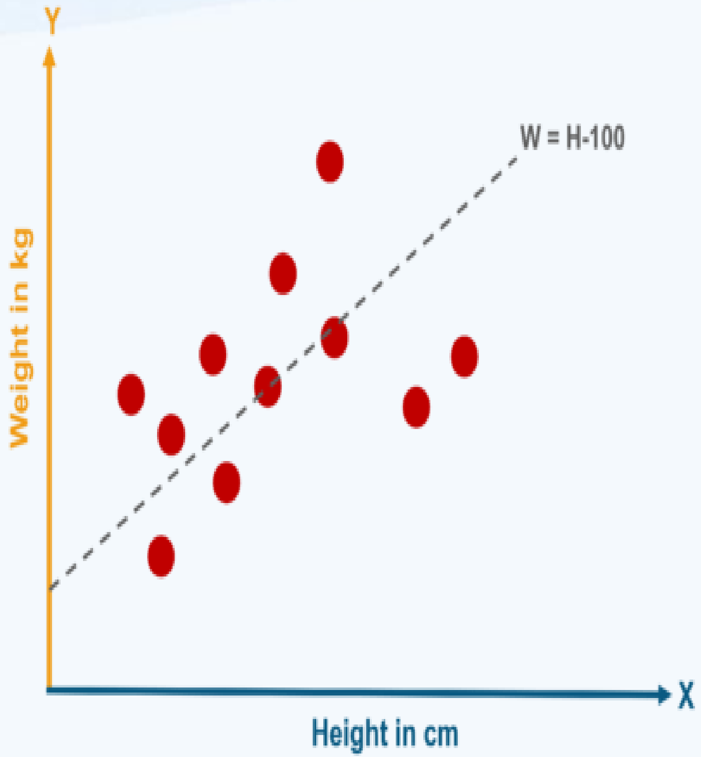
3.Lasso Regression Algorithm.

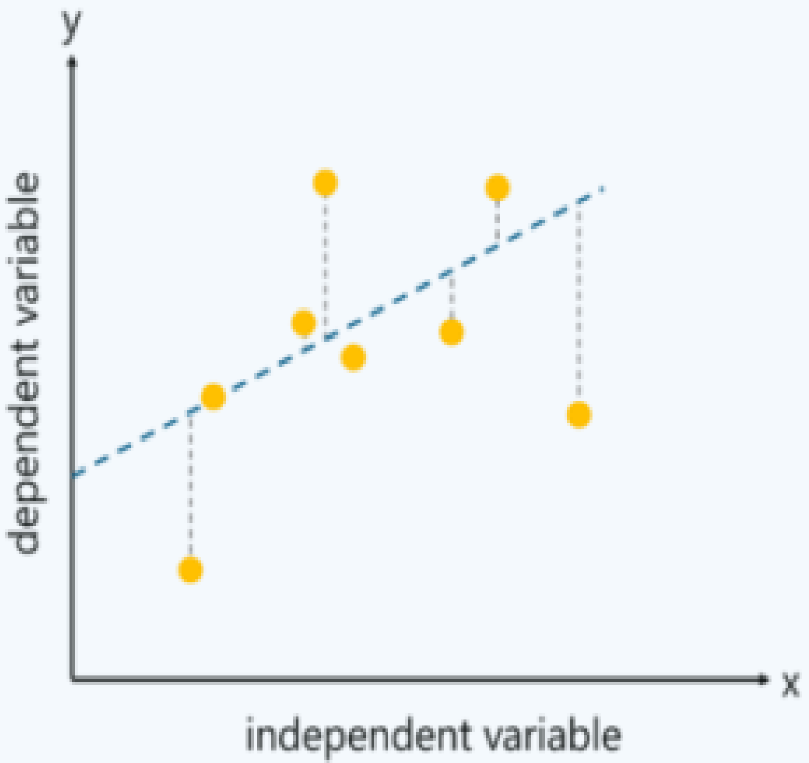
4.Ordinal Regression Algorithm.

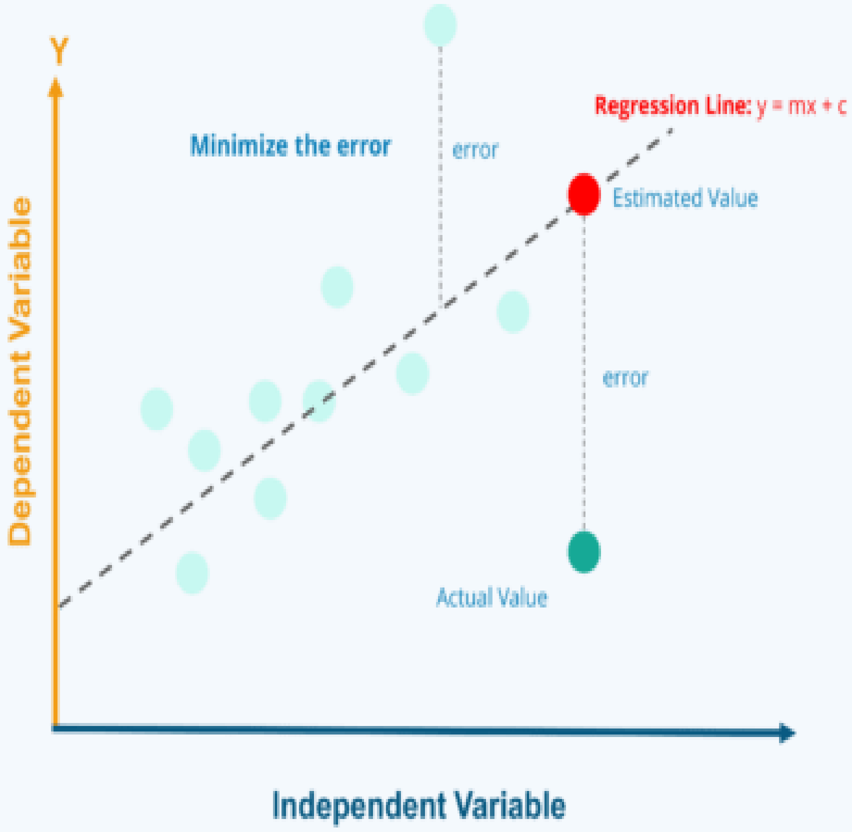
5.Poission Regression Algorithm.

6.Bayesian Regression Algorithm.

Regression


Y
Weight in kg
W = H-100
X
Height in cm


y
dependent variable
x
independent variable


Y
Dependent Variable
Minimize the error
error
Regression Line: y = mx + c
Estimated Value
error
Actual Value
Independent Variable

Classification:

- It is the Process of categerizing data into different classes.
- The Predicted output is always in the form of categorical (or) descreat data.
- There are two types of classification.

  1.Binary classification.

  2.Multi classification.

Binary classification:

- If the Given labelled data can be categerized into two classes is called Binary classification.

Muticlass classification:

- If the given labelled data can be categerized into more than two classes is called Multiclass classification.

Classification Algorithms:

1.Navie Bayes.

2.Support Vector Machine.

3.Decision Trees.

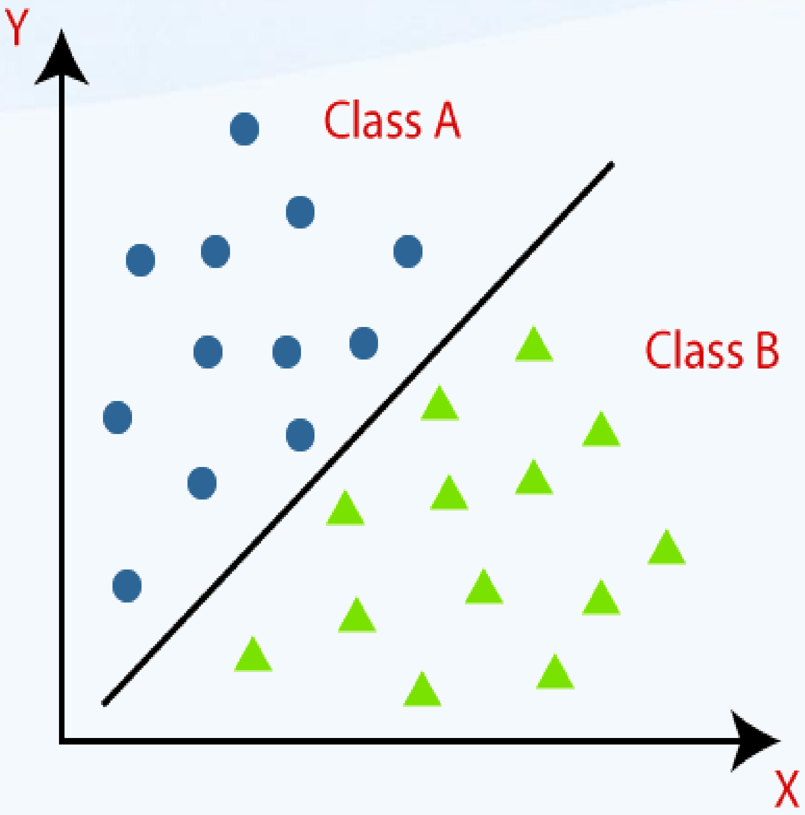
4.Random Forest.

5.K-NN.

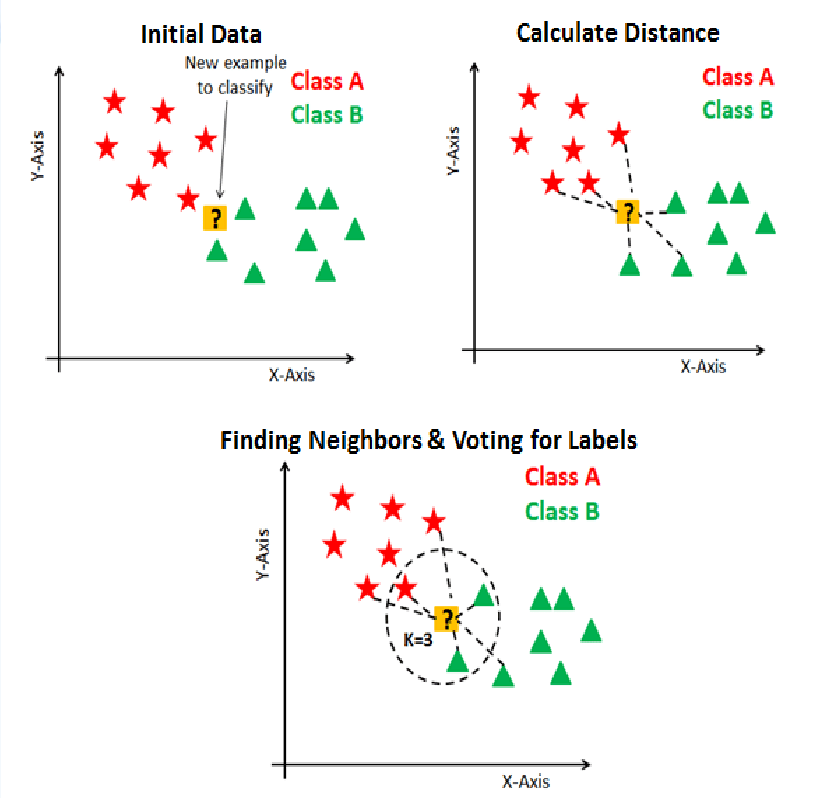
6.Logistic Regression.

7.Neural Networks.

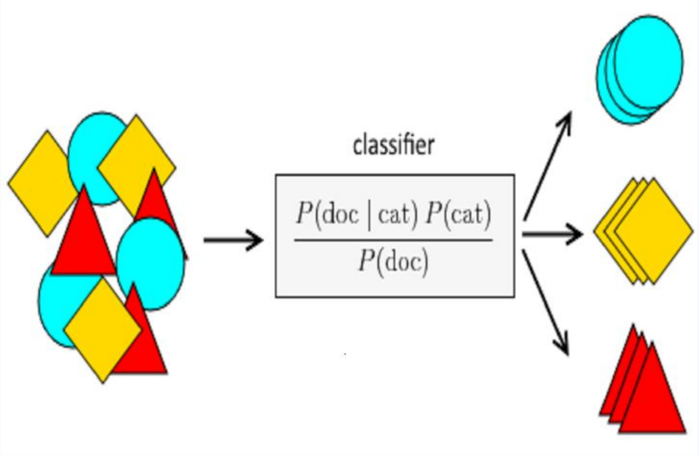
Classification:



SVM



KNN

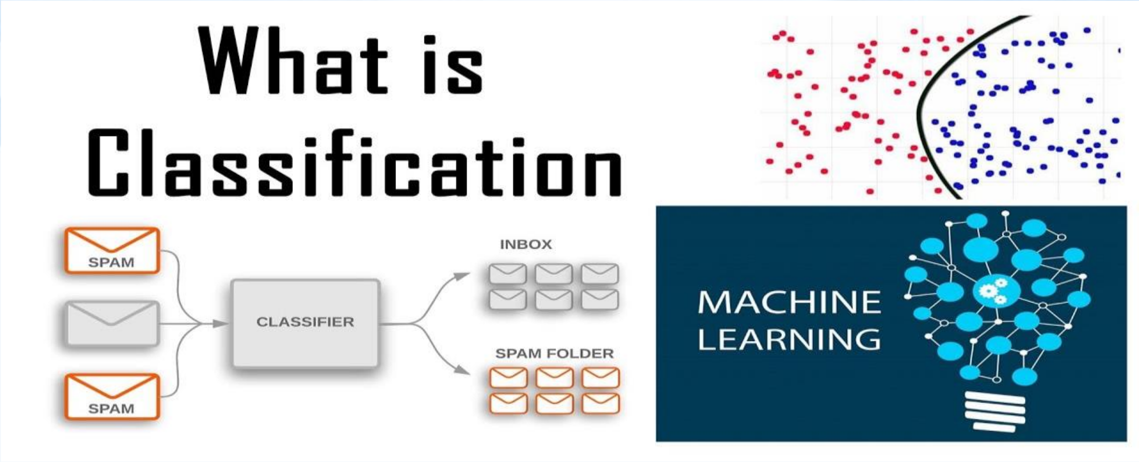


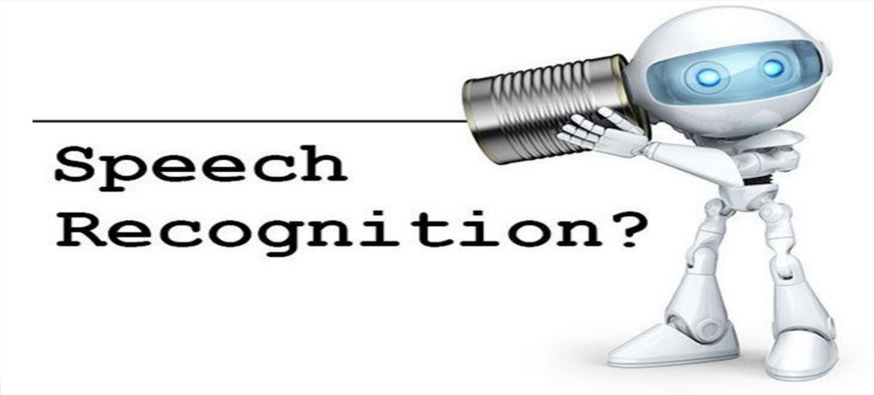
$$\frac{P(\text{doc} \mid \text{cat})\, P(\text{cat})}{P(\text{doc})}$$

NAIVE BAYES

Applications of Supervised ML :

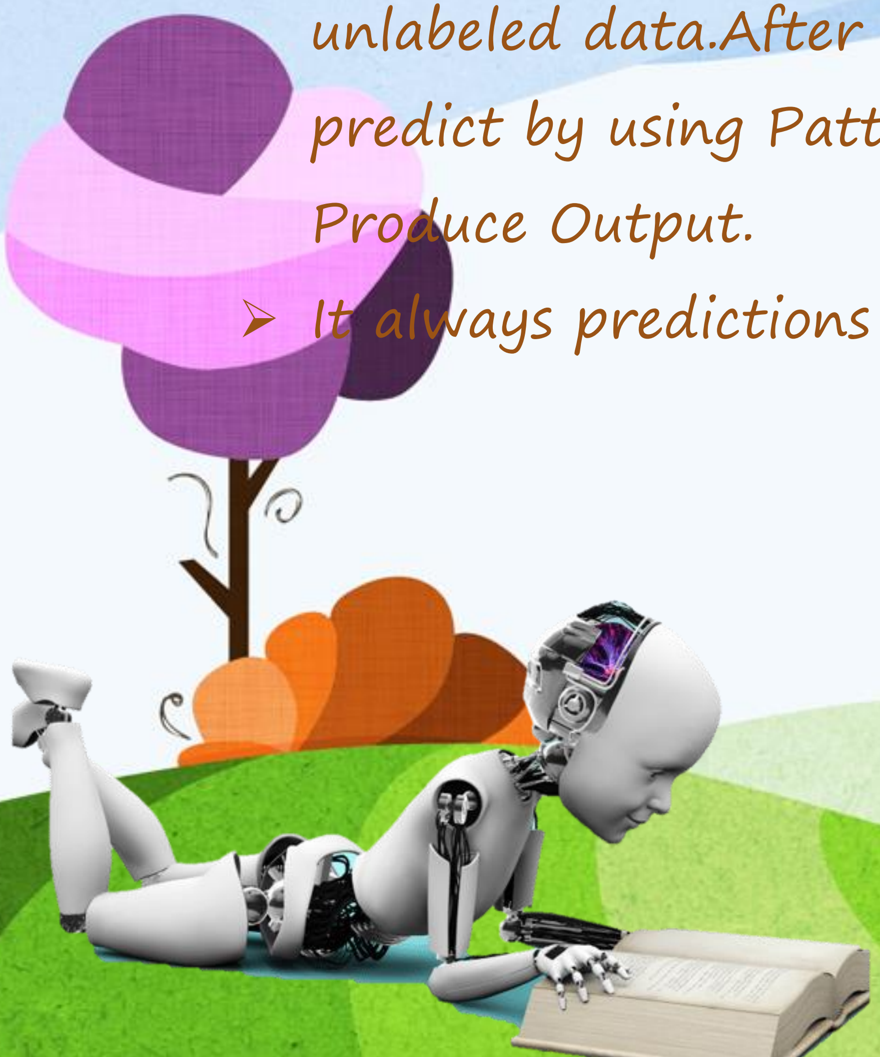
- ➢ Spam Detection.
- ➢ Speech Recognization.
- ➢ Finger print sensor.
- ➢ Text categorization.
- ➢ Weather Forecasting.
- ➢ Image classification.
- ➢ House Price Prediction.

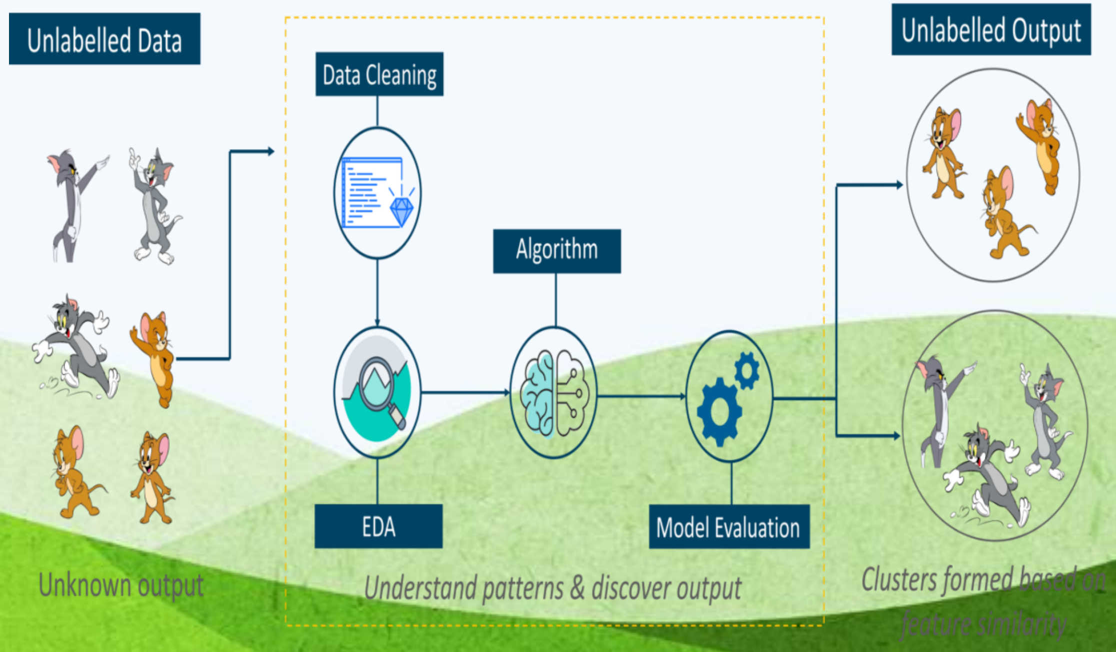




2.Prediction Using Unsupervised Machine Learning

- Unsupervised is one of the type of Machine Leaning Type.
- Unsupervised Machine Learning:when we are training a machine with unlabeled data.After that we are giving testing data,that machine can predict by using Pattern based ,insight characteristics of data and Produce Output.
- It always predictions are not correct,it some times predict wrong also.





Types of Unsupervised Machine Leaning:

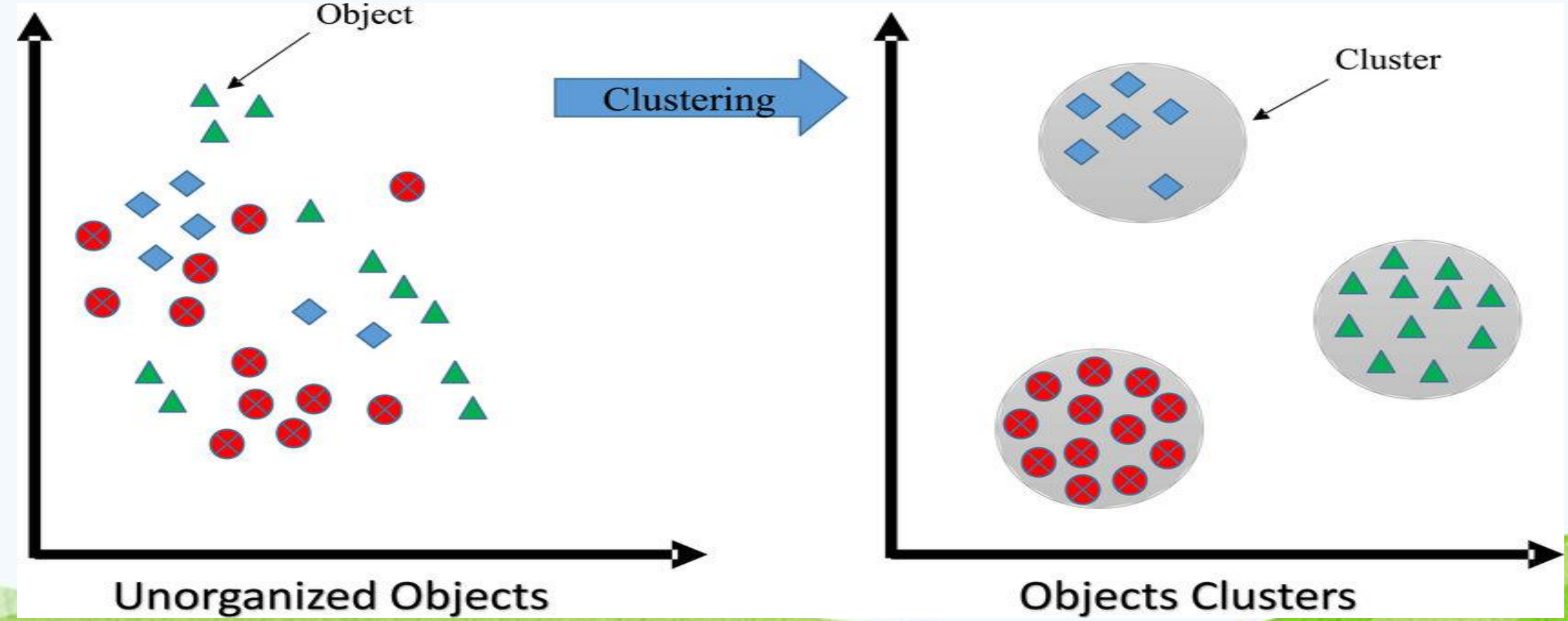
There are Main types of Unsupervised ML:

1.Clustering.

2.Dimensionality Reduction.   3. Association.

Clustering:

- A Group of objects that are similar together basis on their characteristics, Objects in one cluster is dissimilar to in that other cluster data points.



- There are three types of clustering is there.

1.Partail clustering.

i.Hard clustering.

ii.Soft clustering.

2.Hierarchial clustering.

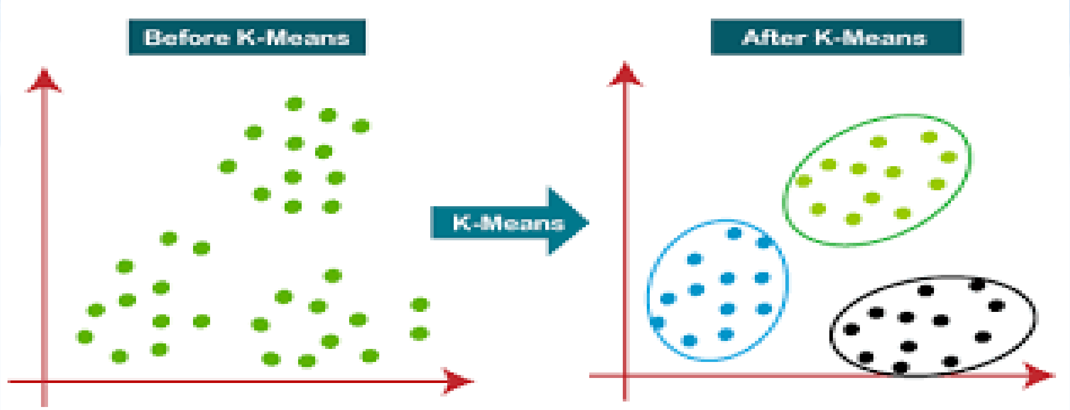
i.Buttom to Top Hierachial clustering.

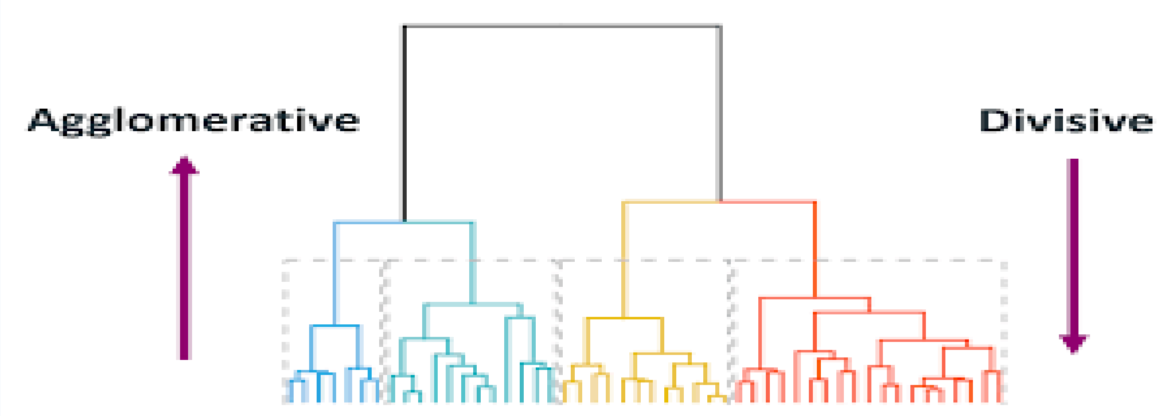
ii.Top to Buttom Hierachial clustering.

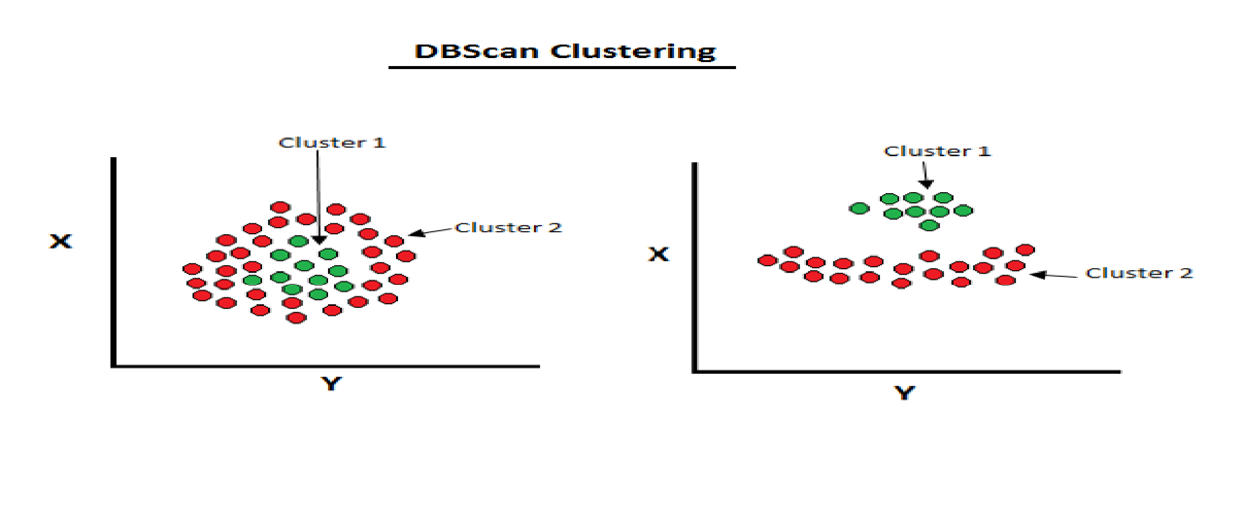
3.Density Based clustering.

Clustering Algorithms:

1.K-means clutering.
2.Fuzzy c-means clustering.
3.Agglomerative clustering.
4.Divise clustering.
5.DBSCAN clustering.







Dimensionality Reduction:

- ➢ The higher the number of features, the harder it gets to visualize the training set and then work on it.
- ➢ Sometimes, most of these features are correlated, and hence redundant.
- ➢ This is where dimensionality reduction algorithms come into play.
- ➢ Dimensionality reduction is the process of reducing some unwanted Features and that features not effect on the final Output.

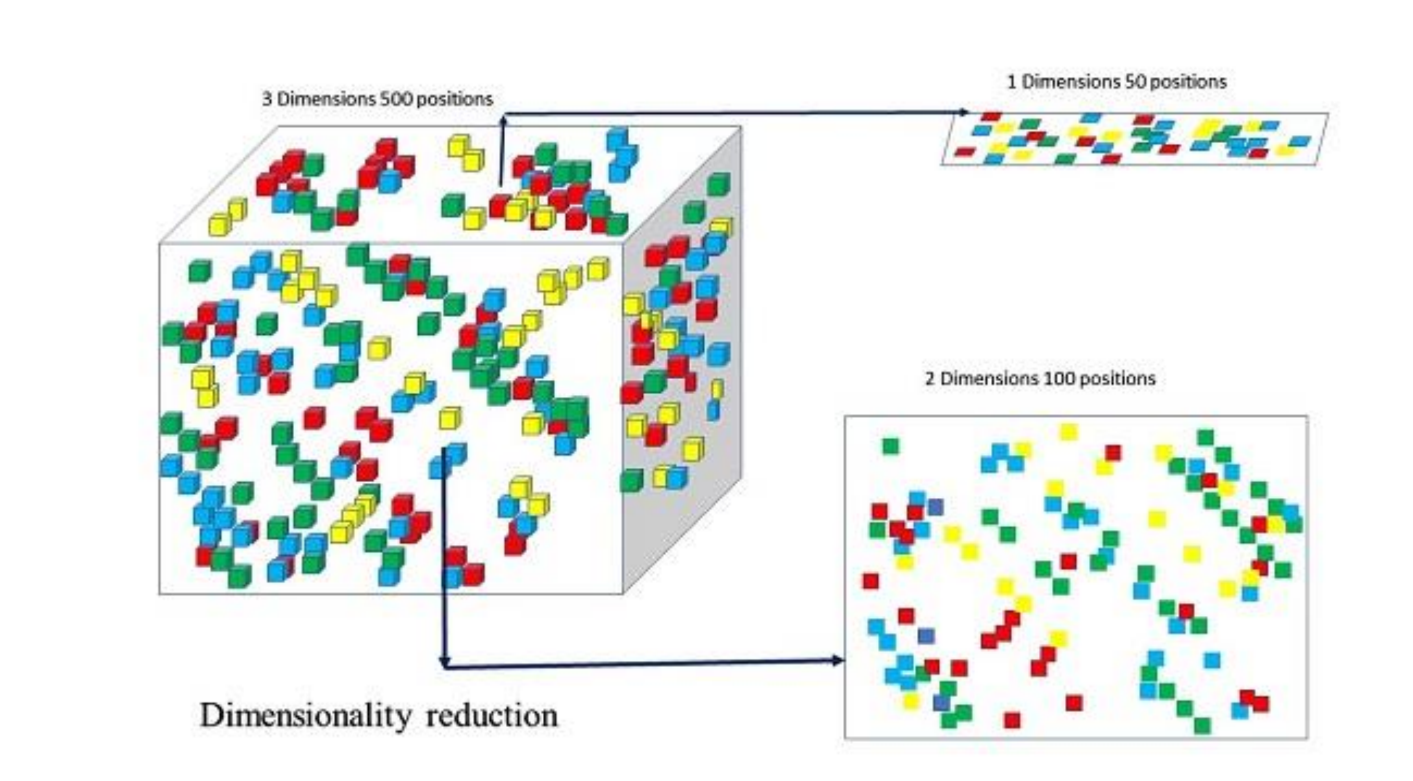
❑ Feature Selection.

❑ Feature Extraction.

➢ Dimensionality Reduction Algorithms:

1.Principle of Component Analysis.

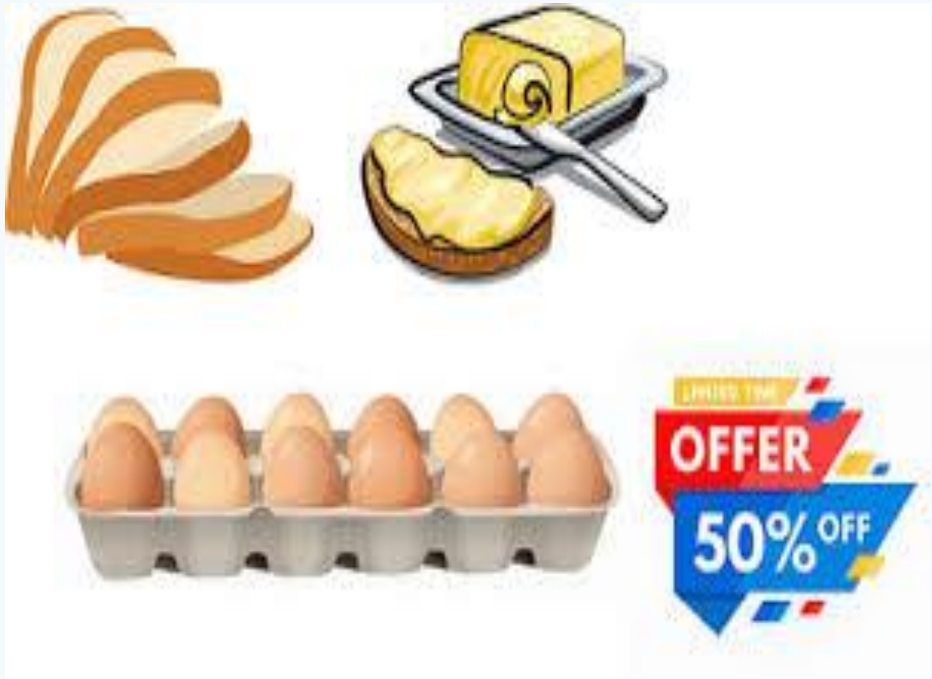
2.Linear Descriminent Analysis.

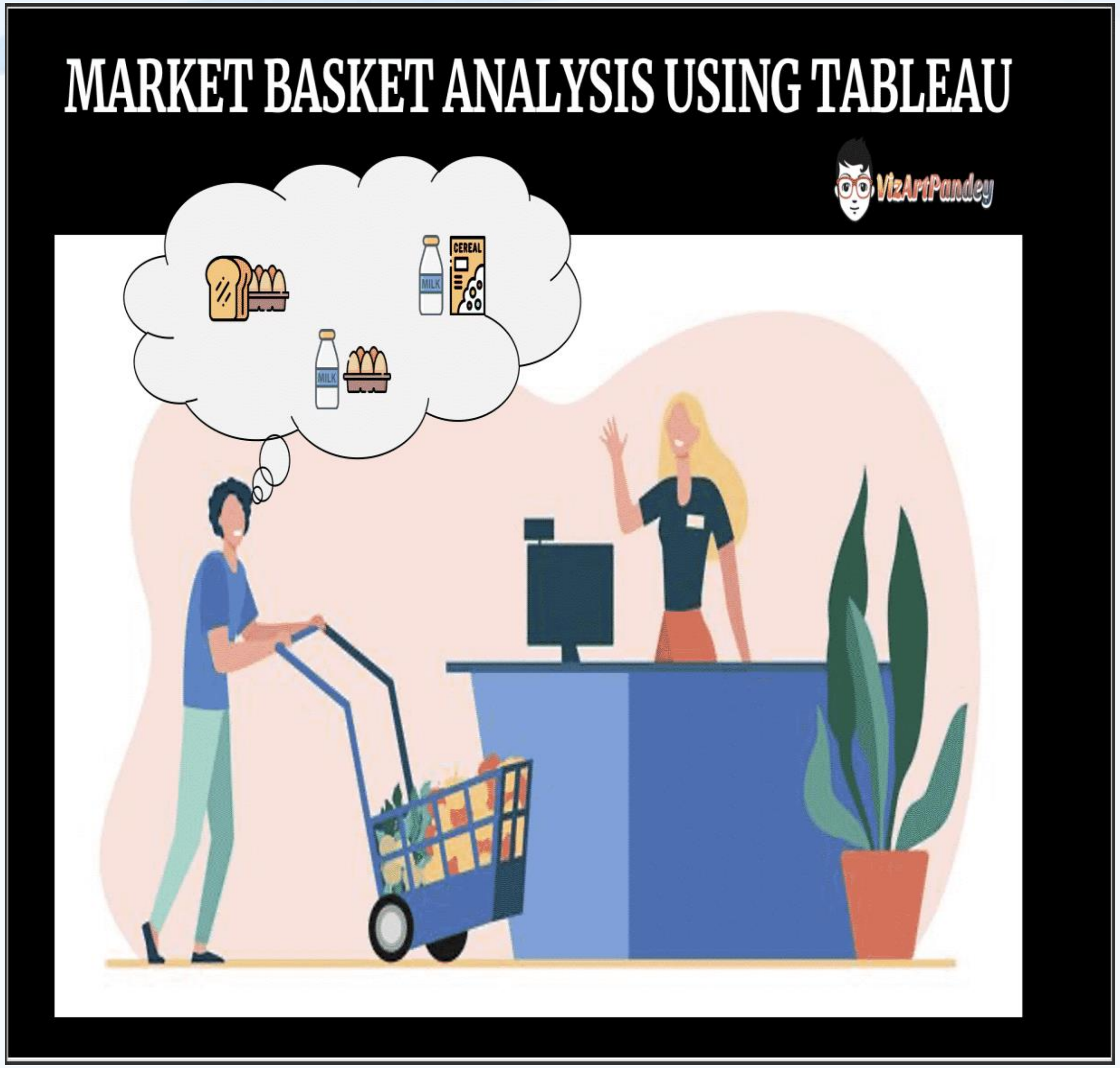

3 Dimensions 500 positions
1 Dimensions 50 positions
2 Dimensions 100 positions
Dimensionality reduction

Association:

- Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.
- It tries to find some interesting relations or associations among the variables of dataset.
- It is based on different rules to discover the interesting relations between variables in the database.
- It is used in different areas ,mainly in market basket analysis.
- We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.
- For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.

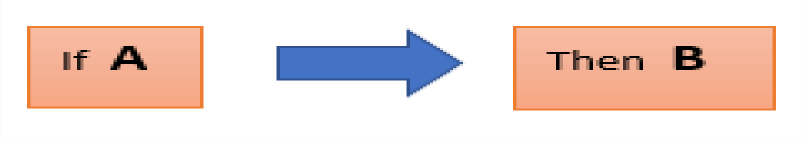
Walmart


LIMITED TIME
OFFER
50% OFF


MARKET BASKET ANALYSIS USING TABLEAU
VizArtPandey
CEREAL
MILK
MILK

- Association Rule:

Association rule learning works on the concept of If and Else Statement, if A then B.



If A

Then B

- These types of relationships where we can find out some association or relation between two items is known *as single cardinality.*
- It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly.
- So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

3.Exploratory Data Analysis

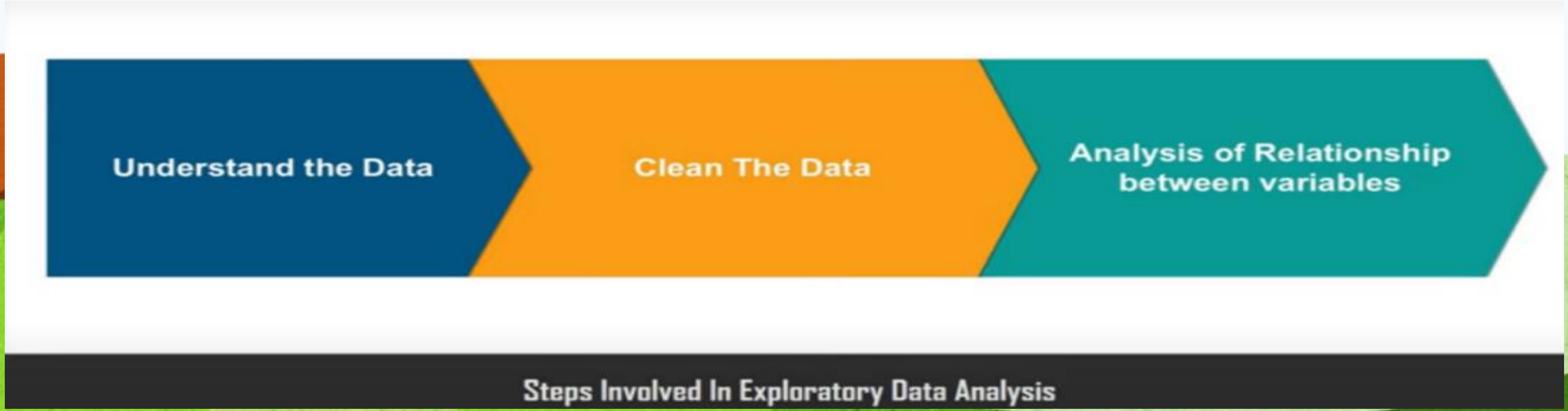
- Exploratory Data Analysis (or) EDA is nothing but a data exploration technique to understand the various aspects of the data.
- It is basically used to filter the data from redundancies.

- Steps in EDA:

  1.Understand the Data.

  2.Clean Data.

  3.Analysis of relationship between variables.



Understand the Data

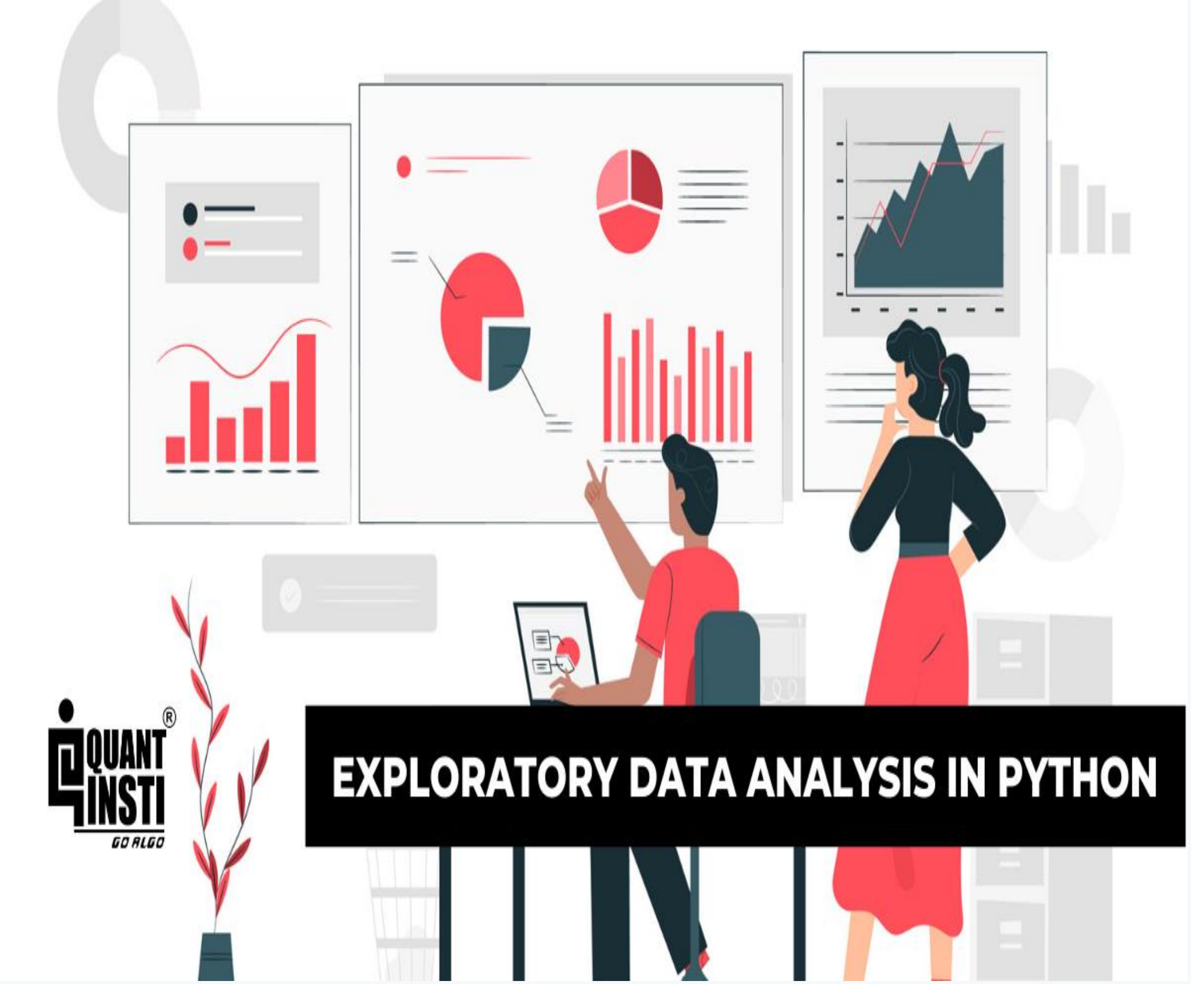
Clean The Data

Analysis of Relationship between variables

Steps Involved In Exploratory Data Analysis

EDA

analytic Steps
EDA
www.analyticssteps.com


QUANT INSTI
GO ALGO
EXPLORATORY DATA ANALYSIS IN PYTHON

There are three primary types of EDA:

- Univariate :

  This shows every observation/distribution in data on a single data variable. It can be shown with the help of various plots like Scatter Plot, Line Plot,Histrograms.

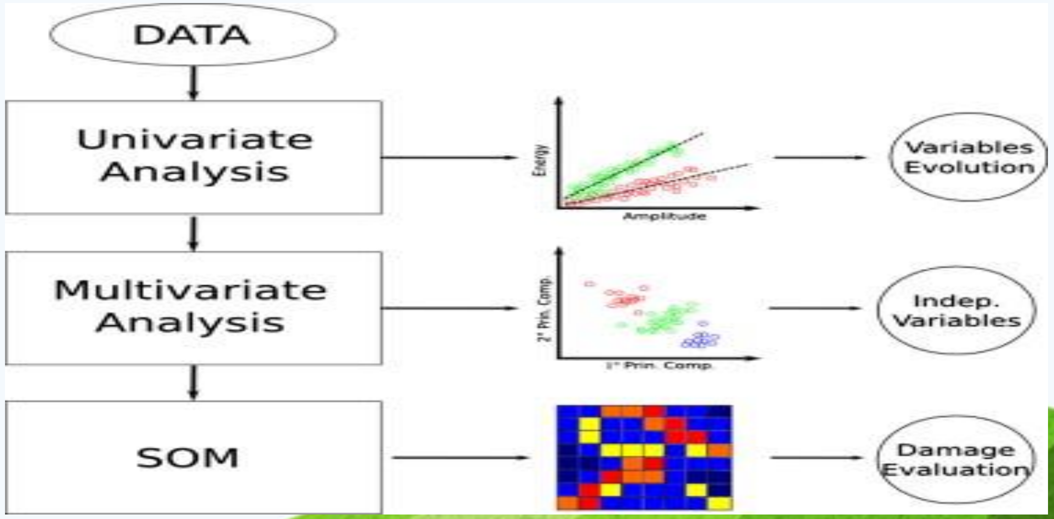
- Bivariate :

  Bivariate analysis displays are done to reveal the relationship between two data variables.

  It can also be shown with the help of Scatter plots, histograms, Heat Maps, Box Plots.

- Multivariate :

  Multivariate analysis, as the name suggests,displays are done to reveal the relationship between more than two data variables.



DATA

Univariate Analysis

Multivariate Analysis

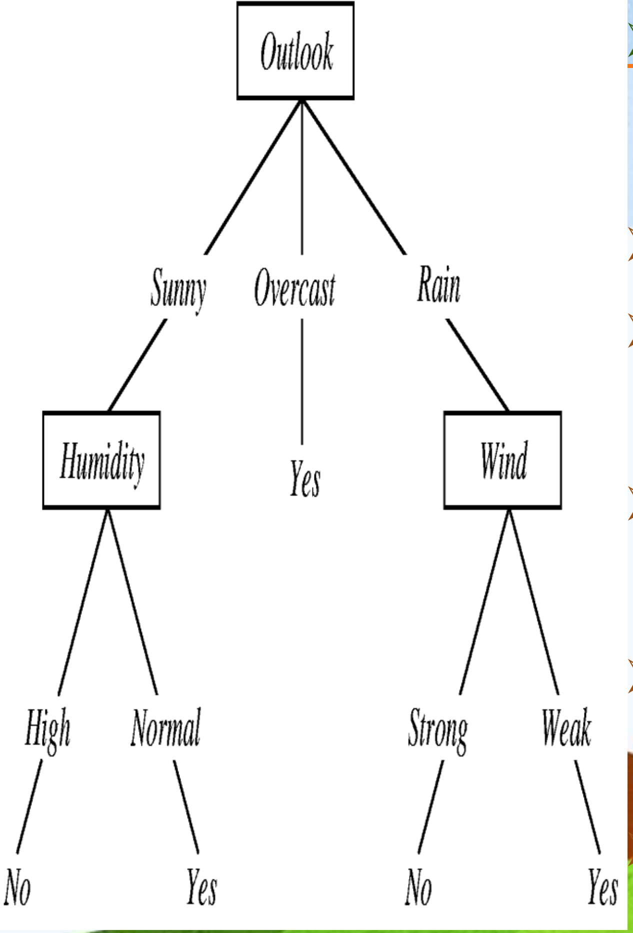
SOM

Variables Evolution

Indep. Variables

Damage Evaluation

4.Prediction Using Decision Trees



- Decision tree:It is a graphical representation of all the possible solutions to decision based on certain coditions.
- Here the root node having entire data set information.
- It is Supervised machine learning algorithm, but it can work both Regression and classification Problems.
- The leaf nodes can having final results and internal nodes can take decision making information.
- In building Decision tree the major challenge is creating Decision nodes.It can possible in two ways.we
- We want split the decision is best to split to divide the information.

  1.CART Algorithm.

  2.ID3 Algorithm.

ID3 Algorithm:

- It mainly Working based on the Information Gain.

Information Gain(IG):

- Information Gain is the decrease/increase in Entropy value when the node is split.
- An attribute should have the highest information gain to be selected for splitting.
- Based on the computed values of Entropy and Information Gain, we choose the best Attribute for split.

Entropy:

- It is measurement of degree of Randomness of information.

$$\boxed{E=-P(yes)\log 2P(yes)-P(no)\log 2P(no)}$$

- $\boxed{I.G=E-[w*E(\text{for each attribute})]}$

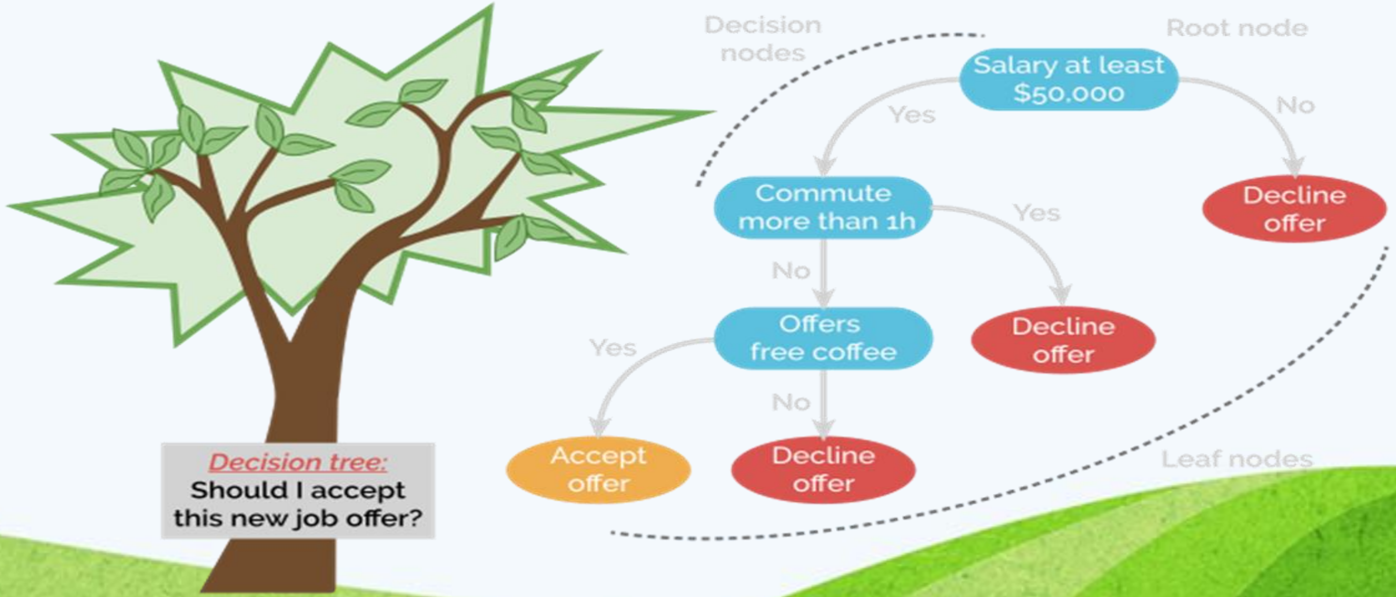
CART Algorithm:

- It is mainly Working based on Gini Index.
- CART full form is Classification & Regression Trees Algorithm.

Gini Index(GI):

- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- It means an attribute with lower Gini index should be preferred.
- The Formula for the calculation of the of the Gini Index is given below.

$$Gini = 1 - \sum_{j} p_j^2$$



Decision tree:
Should I accept this new job offer?

Thanks