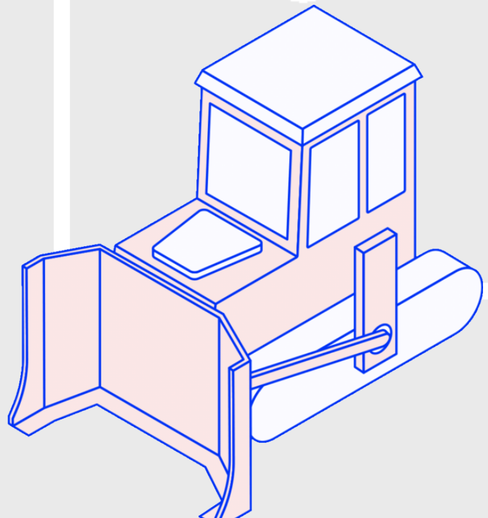
MACHINE LEARNING

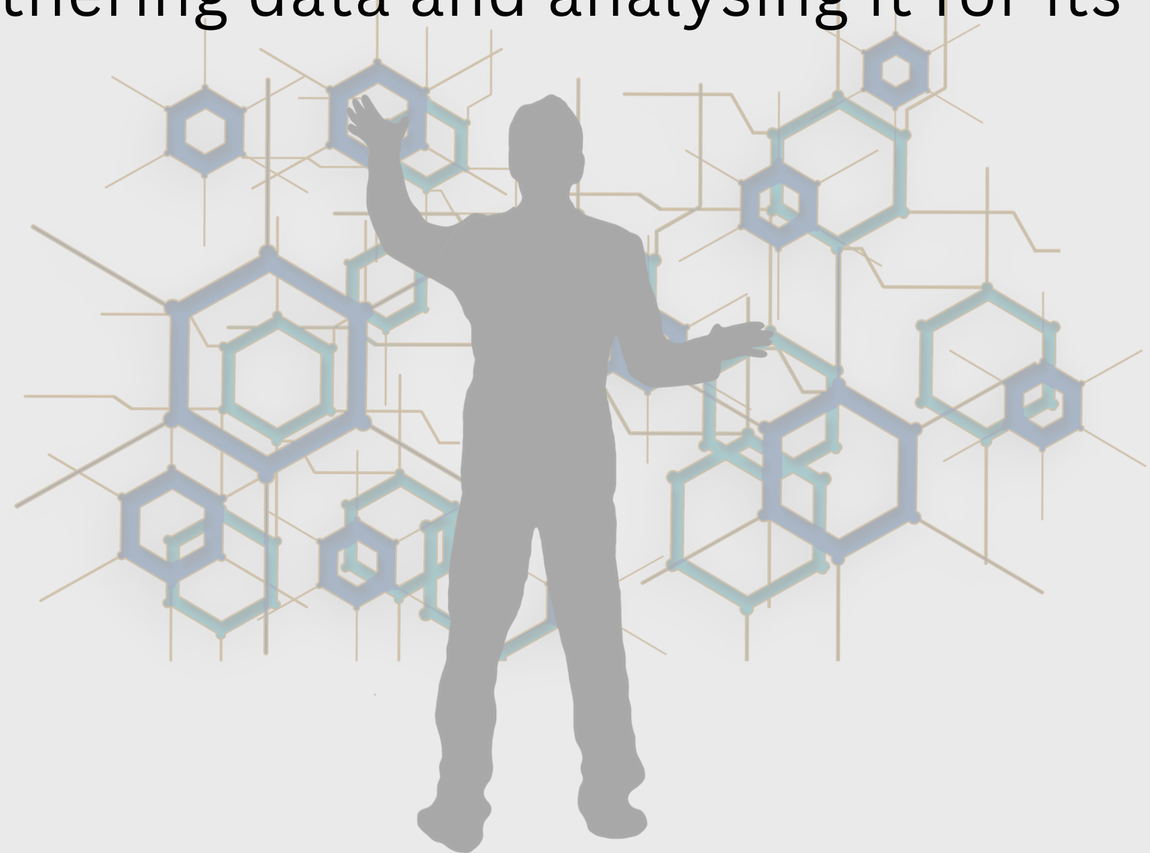
Machine Learning



1. Introduction
2. Types of ML
3. Algorithms
4. Supervised ML Algorithms
5. Unsupervised ML Algorithms
6. Reinforcement ML

Introduction

- Machine Learning(ML) is one of the branches of AI which make computer to act smartly.
- Learning in reference to ML is defined as the process of improving the knowledge of machine by gathering data and analysing it for its own utilisation.
- In today's world data is present everywhere and can be of different flavours such as text, numbers, graphs, images and in many other forms.



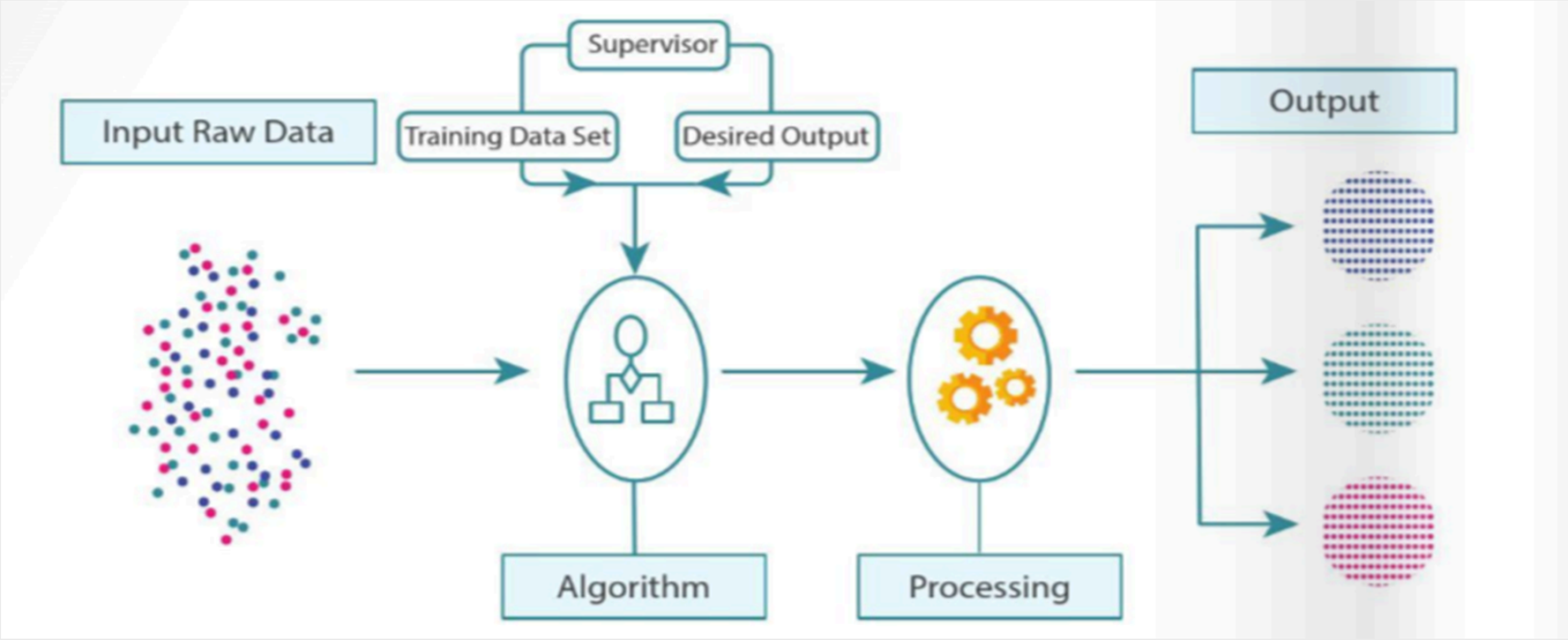
Types of ML Algorithms

1. Supervised ML Algorithms
2. Unsupervised ML Algorithms
3. Reinforcement ML Algorithms

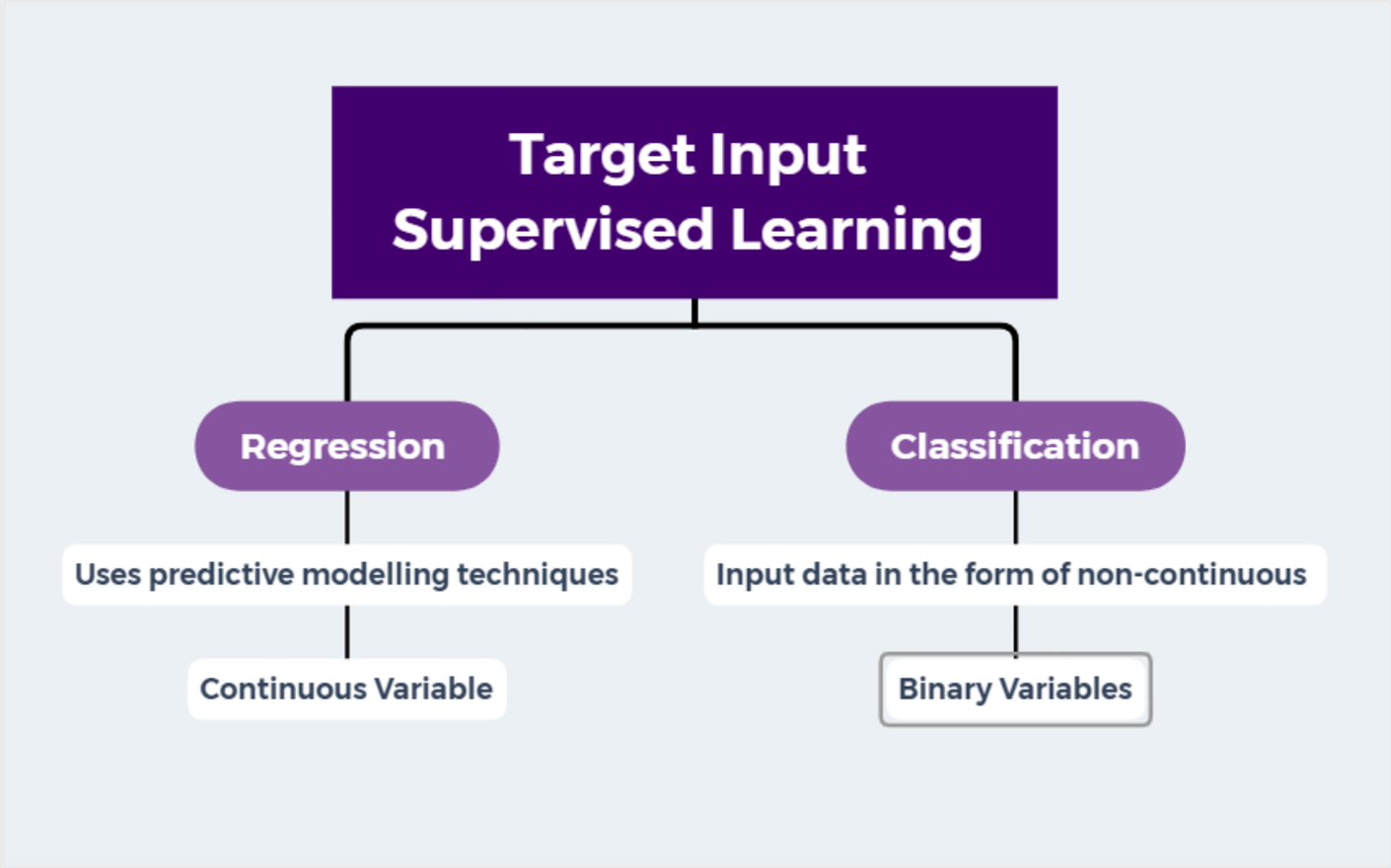
SUPERVISED MACHINE LEARNING

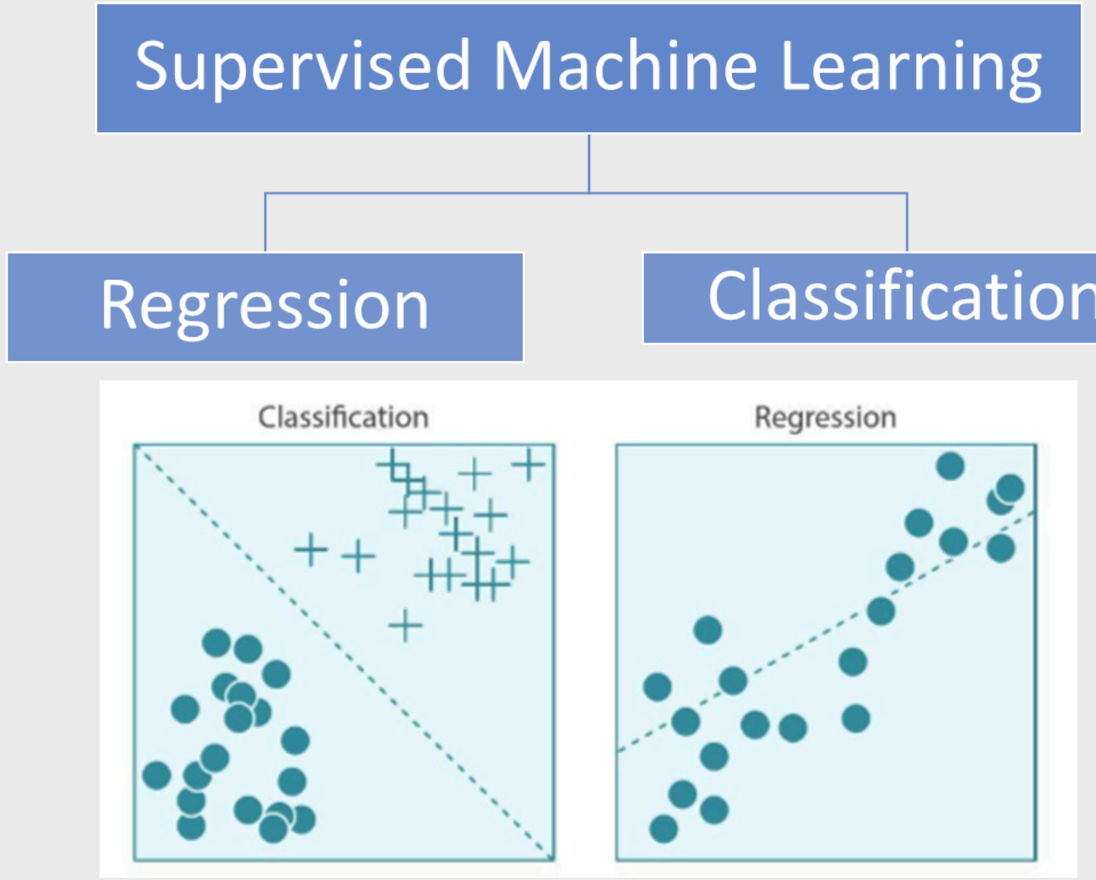
Supervised Machine Learning

- Here data mining approach is used to analyze the trained data to generalize the conclusion.
- It is a type of learning which uses labelled data set to train the algorithm.



Supervised Machine Learning





Most Commonly used Algorithms in Supervised ML

- Linear Regression
- Logistic Regression
- Naive Bayes
- Decision Trees
- k-Nearest Neighbours
- Support Vector Machine (SVM)
- Random Forest
- Xgboost

UN-SUPERVISED MACHINE LEARNING

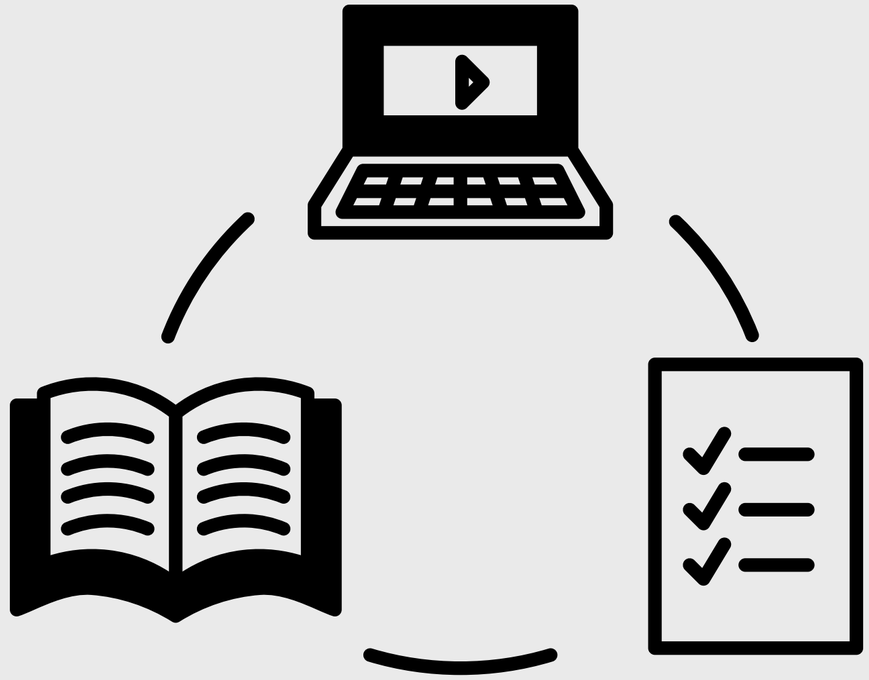
UNSUPERVISED MACHINE LEARNING

- It uses unlabelled (raw) data because training set is not available to train the algorithms.
- In this type of learning machine is complex and time consuming as the data is not pre-classified.
- Its main purpose is to analyse the data and extract structured data.
- In this learning machine learns by observing data and discover patterns within data.

Most Commonly used Algorithms in Unsupervised ML

- K-means Clustering
- Hierarchical clustering
- Anomaly detection
- PCA(principal component analysis)
- Apriori algorithms

REINFORCEMENT MACHINE LEARNING



REINFORCEMENT MACHINE LEARNING

- Without any prior knowledge input given this algorithm has to identify the situation on its own.
- Algorithm learns from situation from a series of trial and error to identify the reward gaining data sets.
- This technique is mostly used in gaming, navigation and robotics.

Most Commonly used Algorithms in Reinforcement ML

- Q-Learning
- SARSA
- DDPG-Deep deterministic policy gradient
- A3C
- PPO

many more...

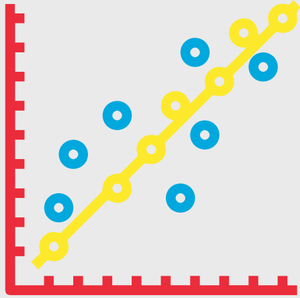
Parametric ML Algorithms

- Assumption can greatly simplify the learning process
- Algorithms that simplify the function
- eg. Logistic regression, Linear Discriminant Analysis, Perceptrons, Linear Regression.

Non-parametric ML Algorithms

- Algorithms that do not have strong assumptions about the form mapping function
- By not making assumptions, they are free to learn any functional form from training data
- eg. Decision Tree like CART and C4.5, Naive Bayes, SVM

Supervised Machine Learning

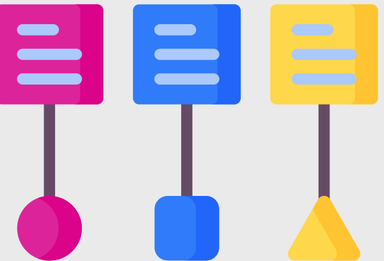


Supervised ML is classified as-

1) Classification
2) Regression

In both classification and Regression, there are two types of variables-

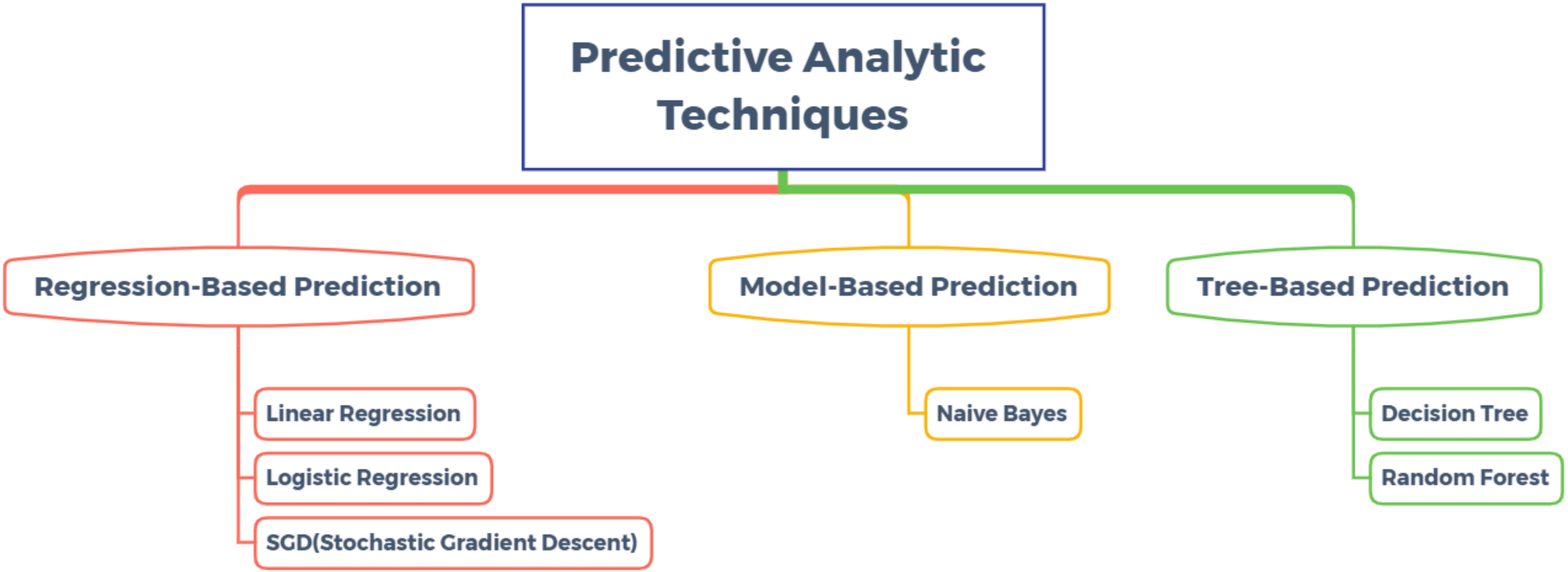
1) Target Variables (output, Dependent Variable)
2) Predictive Variables (input, Independent Variables)



Difference Between Classification & Regression

Classification	Regression
• Helps to predict discrete outcomes • Continuous values may be predicted but they will be in the form of probabilities. • can be estimated using accuracy.	• Helps to predict continuous outcomes. • Discrete values may be predicted but they will be in form of an integer quantity. • can be estimated using root mean squared error.

On the basis of method of Analysis



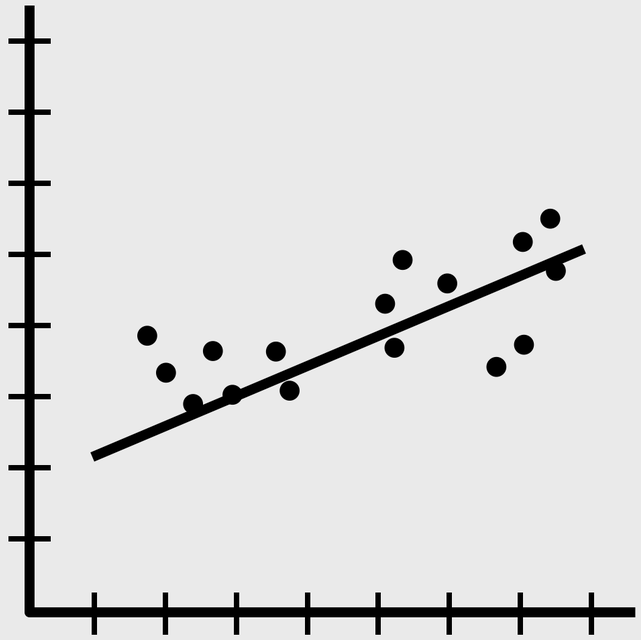
Supervised ML Algorithms

1. Linear Regression (Simple, Multiple)
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Support Vector Machine
6. Naive Bayes Classifier
7. K-Nearest Neighbour
8. Extreme Gradient Boost

Linear Regression

- Regression Analysis is a statistical techniques for investigating and modelling the relationship between variables
- It is introduced by Francis Galton
- Types of Linear regression:
  1) Simple Linear Regression
  2) Multiple Linear Regression
  3) Logistic Regression
  4) Ordinal Regression
  5) Multinomial logistic regression.

Simple Linear Regression



Simple Linear Regression

- Linear regression is a statistical analysis to show relationship between two variables.
- It tries to find relationship between two variables by using linear equation.
- The one variable is explanatory variable while another variable is dependent variable.
- For instance, develop a graph to relate individual's weight with their heights using linear regression model.

Steps in Regression Analysis

Step:1] Statement of the problem under consideration
Step:2] Choice of relevant variables
Step:3] Collection of the data
Step:4] Specification of the model
Step:5] Choice of method for fitting data
Step:6] Fitting Model
Step:7] Model evaluation
Step:8] Using chosen model(s) for solution of proposed problem and forecasting

Simple Linear Regression

- Used to estimate the relationship between two quantitative variables
- Formula:

y=β0+ β1x+ ϵ

where

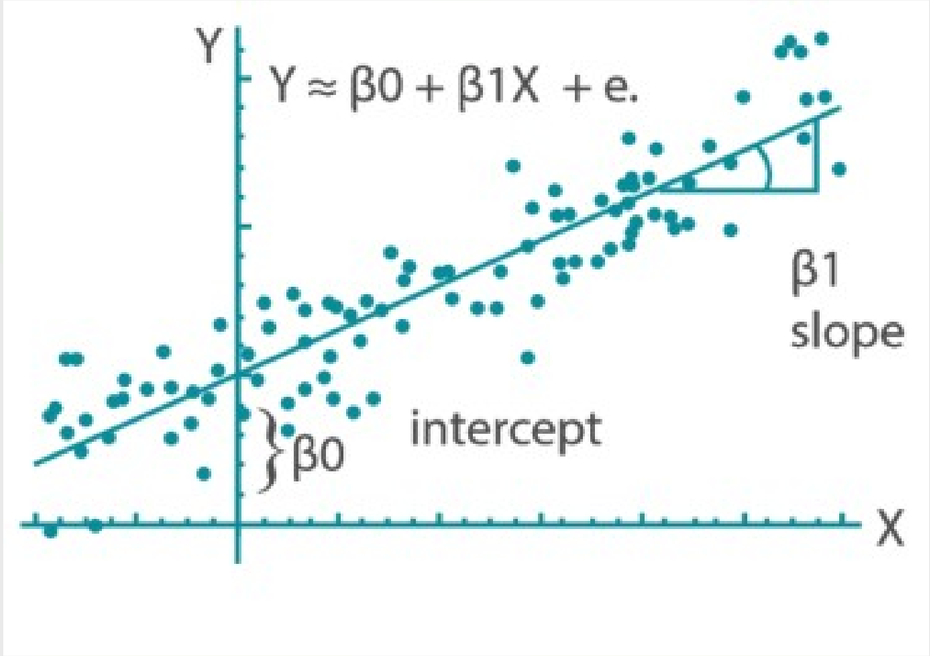
y - dependent variable

β0 - intercept

β1 - slope(regression coefficient)

x - independent variable

ϵ - error of estimate

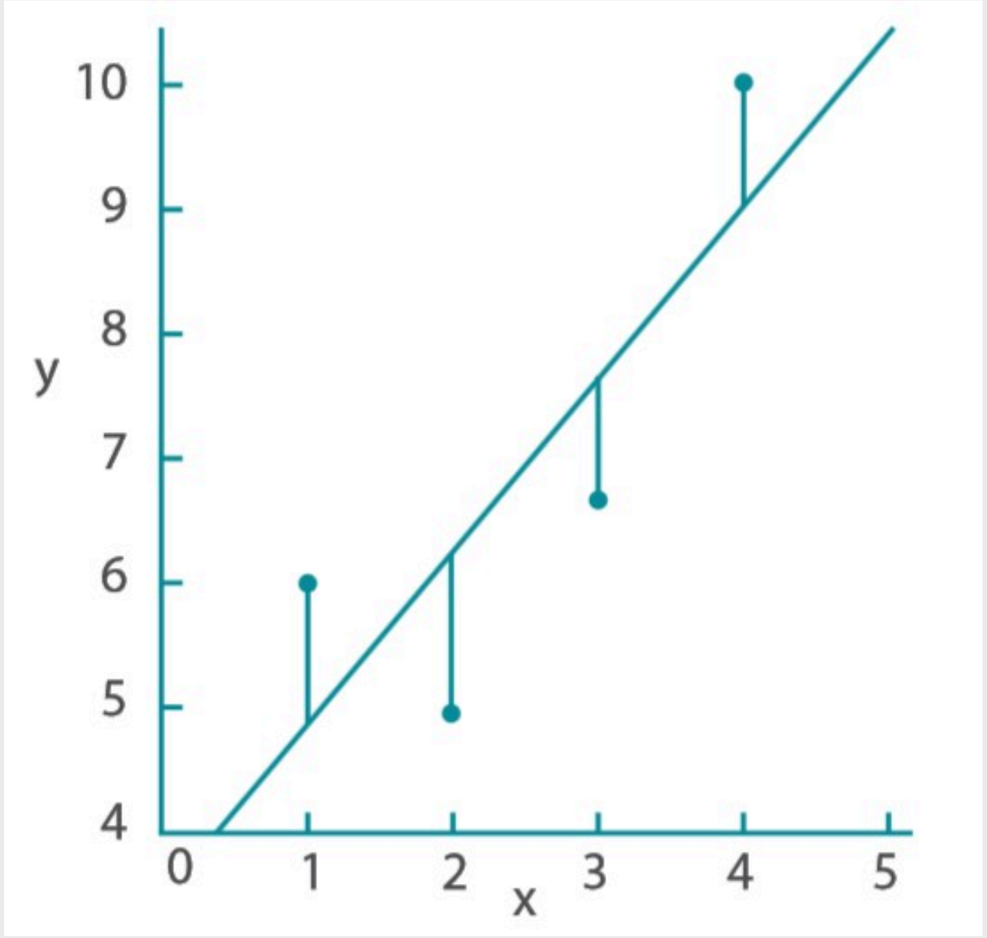


$$\beta 0 = \bar{y} - \beta 1 \bar{x}$$

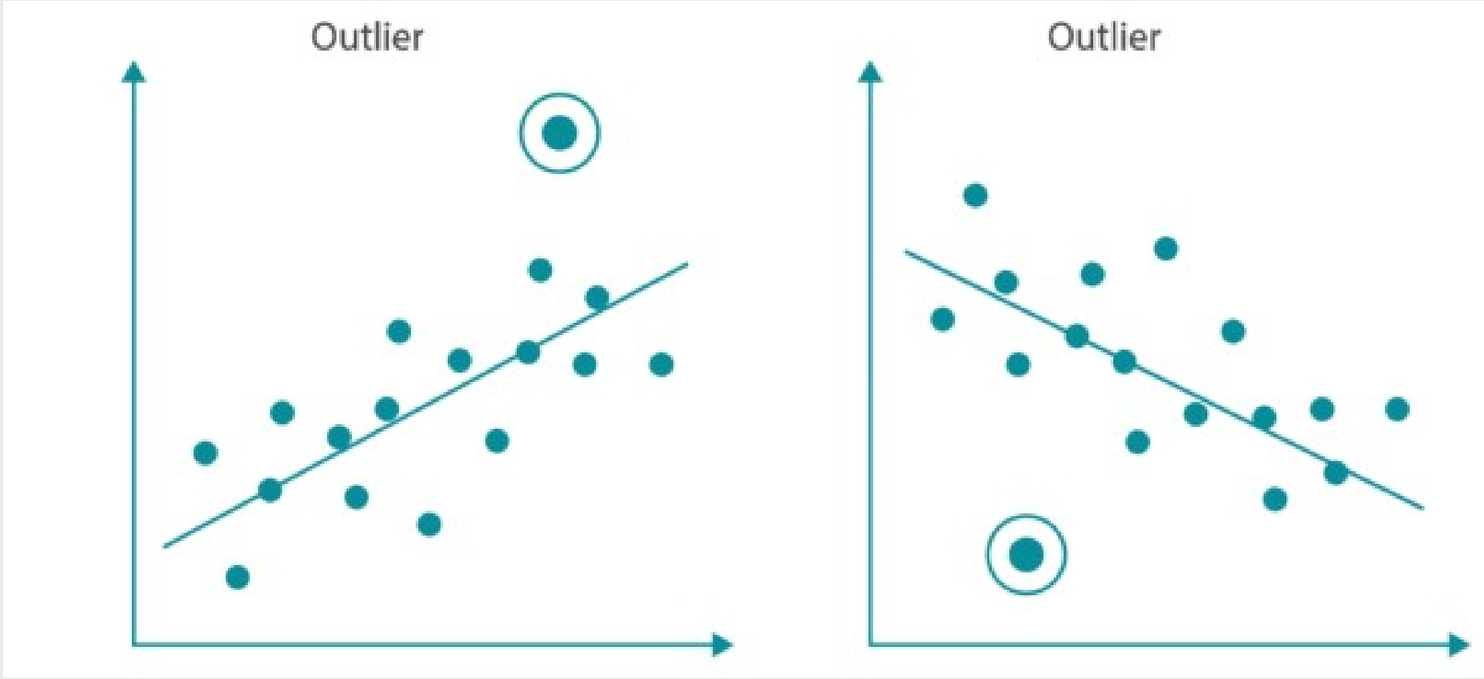
$$\beta 1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Least Square Method

- Least Square method is used to fit regression line.
- The best fit line is calculated by least square method to minimize sum of square of deviation.
- Main aim is to minimize distance between target values and regression line.



Outlier



After regression line has been plotted, there are some points that lie far away from line. These points are called Outliers. Such points show error data or poor regression line.

Practical Session

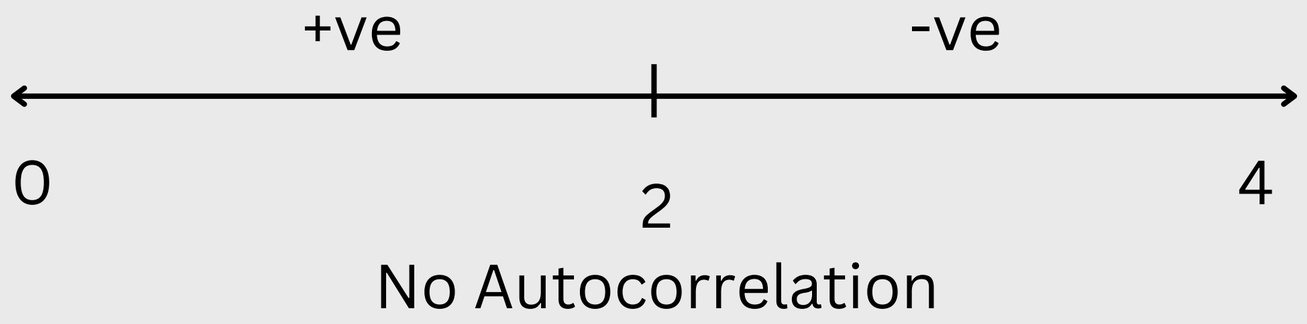
Simple Linear Regression

Assumptions of Linear Regression

1] The relationship between response y and the regression x is linear (approx.)
2] Error term has zero mean and constant variance.
3] Errors are uncorrelated.
4] Errors are normally distributed.

Autocorrelation

- Autocorrelation measures the relationship between a variable's current value and its past values
- -1 < autocorrelation < 1
- statistical test: Durbin watson test to detect presence of autocorrelation
- Durbin-Watson Test:
  1) Measures the amount of autocorrelation in residuals.
  2) always have values between 0 to 4.



Multicollinearity:

- It occurs when two or more independent variables have high correlation.
- Due to presence of multicollinearity estimated regression coefficients become unstable and difficult to interpret
- Multicollinearity may not affect the accuracy of ML model but might affect reliability in determination of effect of features.

Detection of Multicollinearity:

1) VIF - Variance Inflation Factor
2) VIF determines the strength of correlation between independent variables.
VIF = 1/(1-R^2)

3. VIF starts at 1 and has no upper limit
4. VIF = 1 --- (No correlation)
5. VIF > 5 or 10 ----(High multicollinearity)

Residual Analysis

A residual(Error) is the difference between predicted and actual value of the data.

$$e = y - \hat{y}$$

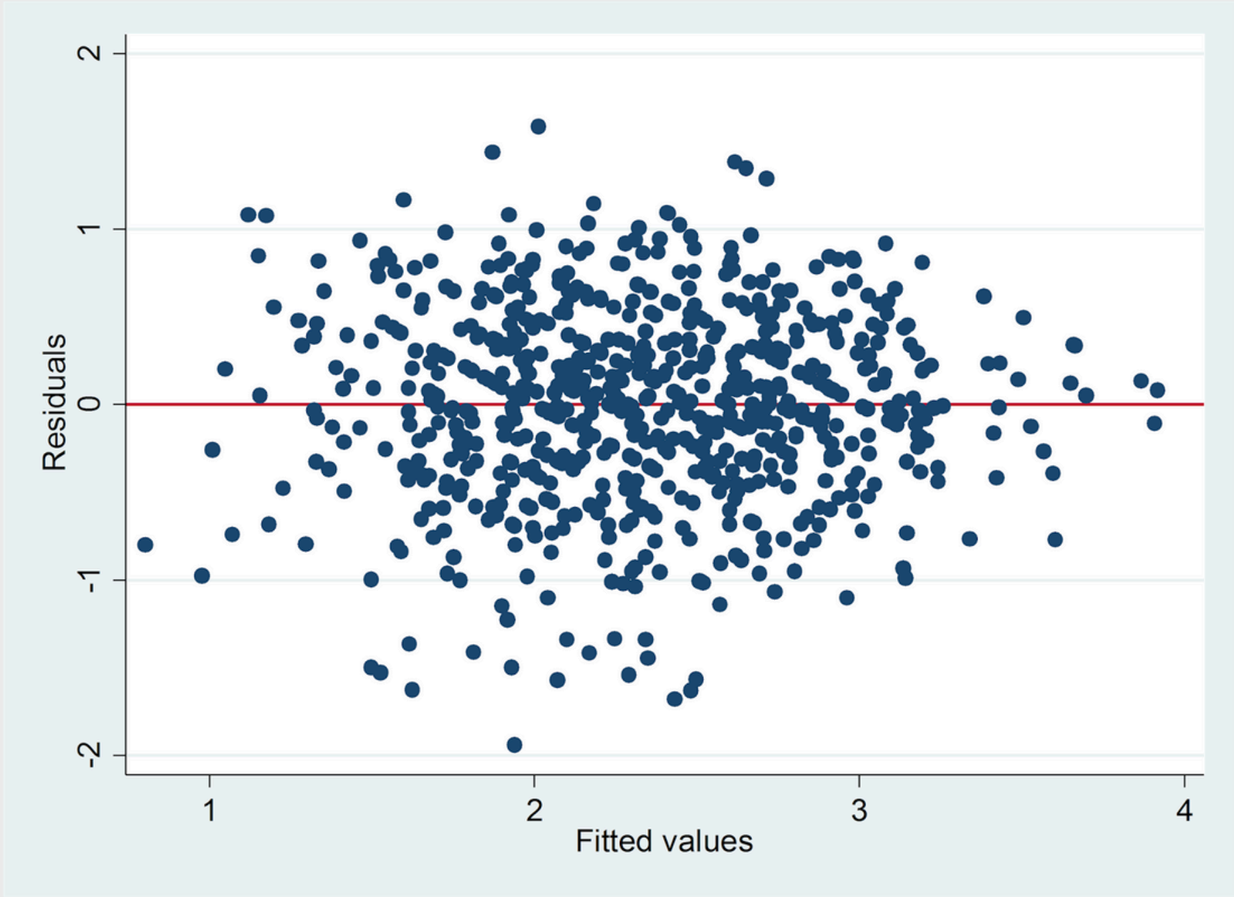
Where: e - residual
y - actual values
$\hat{y}$ - predicted

Residual Interpretation:

1) incredibly useful for determining best suited model
2) using residual plot we can determine whether a linear or non-linear model is preferable.

Residual Plot

1) It is a scatter plot with residuals of variable plotted on y-axis and value of variable x on x-axis.



Advantages

1. estimation of output process become simpler because of linearity.
2. It selects the best fit line which result in minimum error from all the points.
3. space complexity is very low.
4. good interpretability.

Disadvantages

1. process is limited to linear regression.
2. focuses on the mean of dependent process.
3. assumes data is independent.

Applications

1. Analysing trend and estimated sales of company.
2. Variation of price based on customer behaviour.