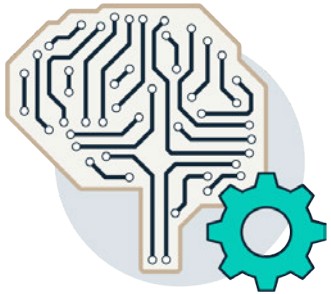
MACHINE LEARNING
-Introduction

Dr. M Lakshmi
Professor and Head.
Department of Data Science and Business Systems
School of Computing
SRM Institute of Science and Technology, Kattankulathur
Date : 24-07-2023

VS



VS

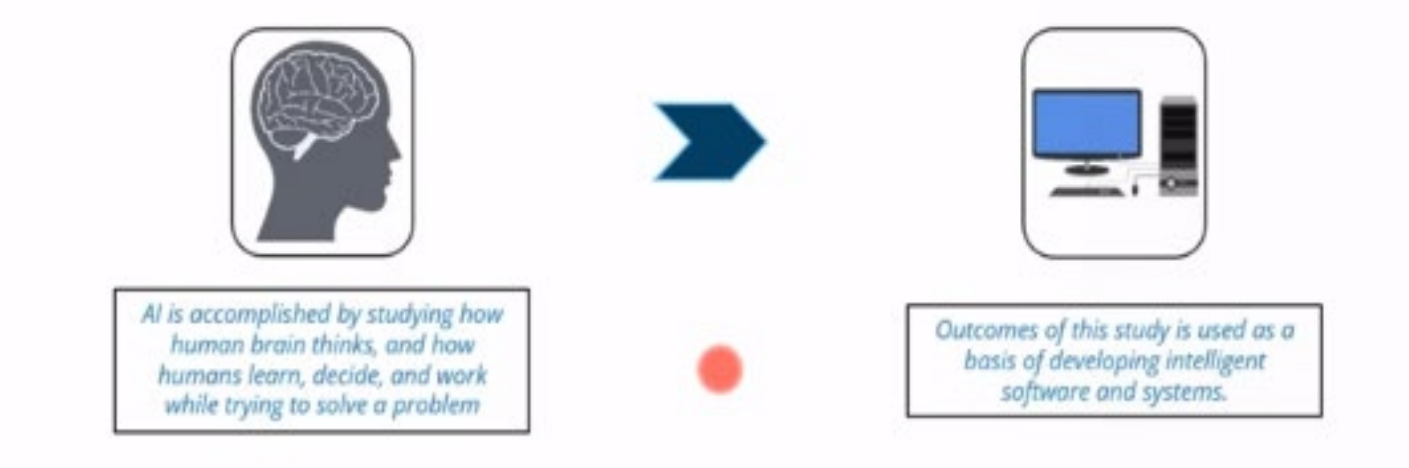
Artificial Intelligence

Machine Learning

Deep Learning

What is AI?

•AI is the capability of a machine to imitate intelligent human behavior



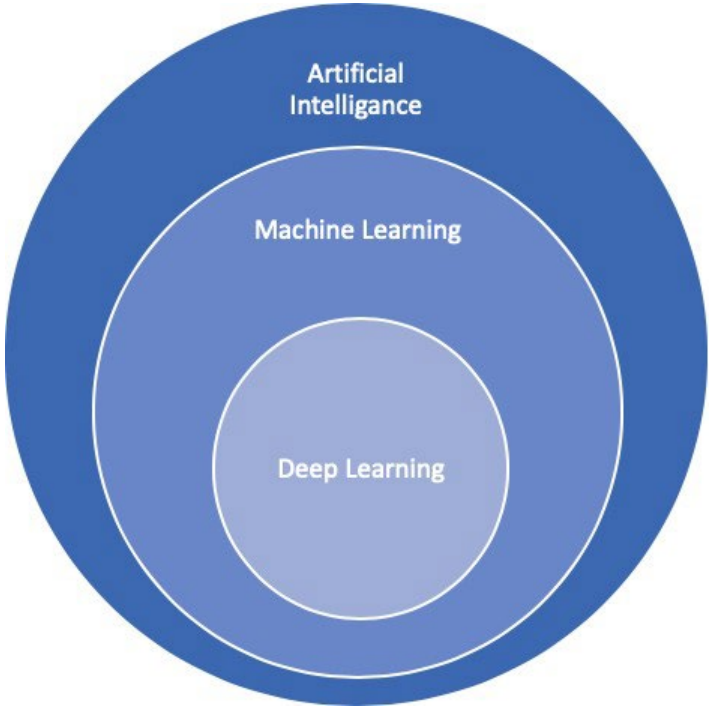
What is Artificial Intelligence?

Artificial Intelligence is a branch of Computer Science that is concerned with building smart & intelligent Machines

Non – intelligent machines

Intelligent machines

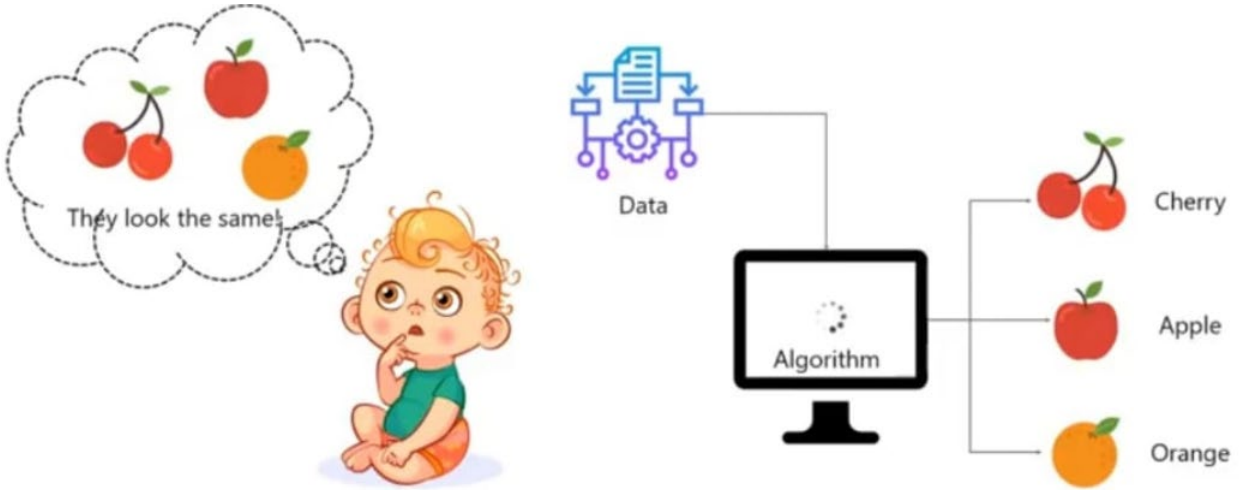
AI vs ML vs DL



Machine Learning is a subset of Artificial Intelligence

Deep Learning is a subset of Machine Learning

Child Learning



Learning

- Past experience
- Learning from others

- Machines need instruction from humans

Machine Learning

Machine Learning is a technique to implement AI that can learn from the data by themselves without being explicitly programmed.

Iron Man

Captain America

Machine Learning – In general

- A subset of Artificial Intelligence
- Machines that learn automatically & improve from experience.
- No explicit programme is needed

Definition - ML

A set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.

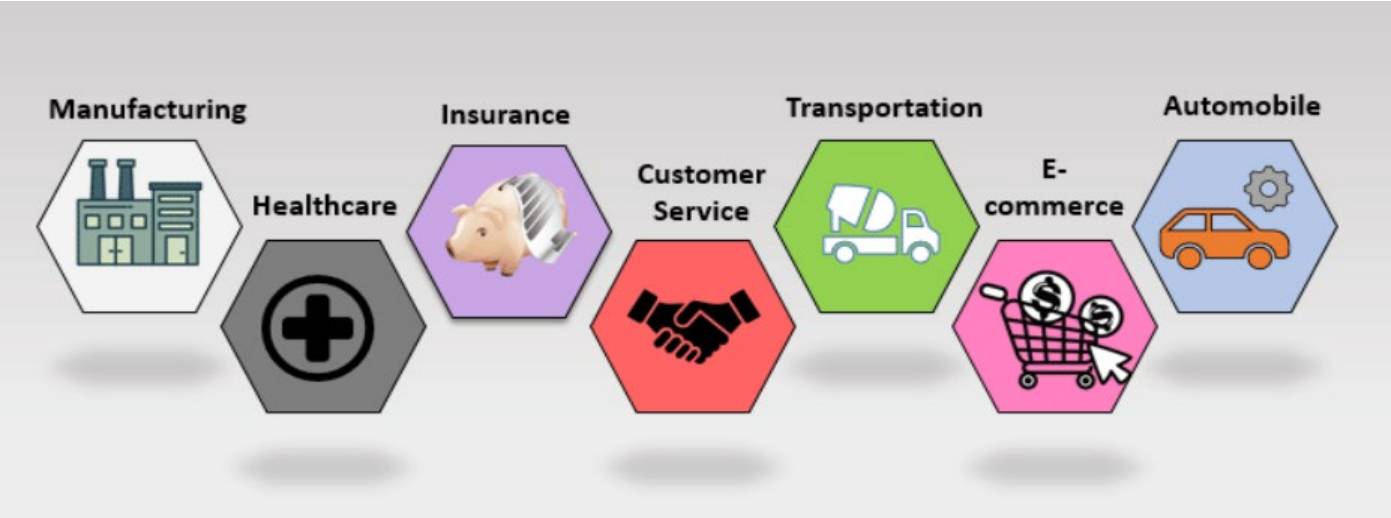
Need for ML

- Huge amount of data (Big data)

- Automated methods of data analysis

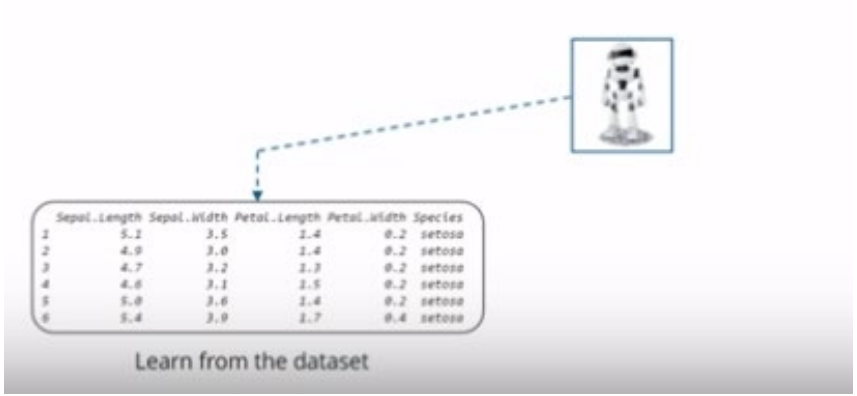

Machine Learning
Applications

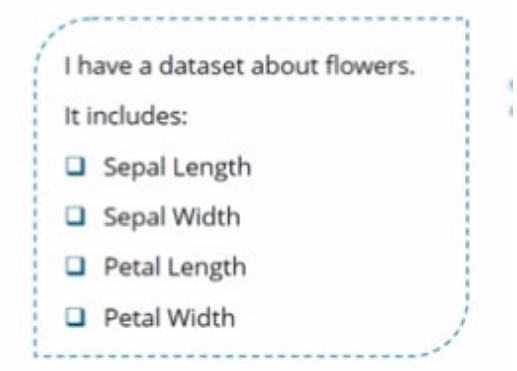
Few ML Applications



Machine Learning

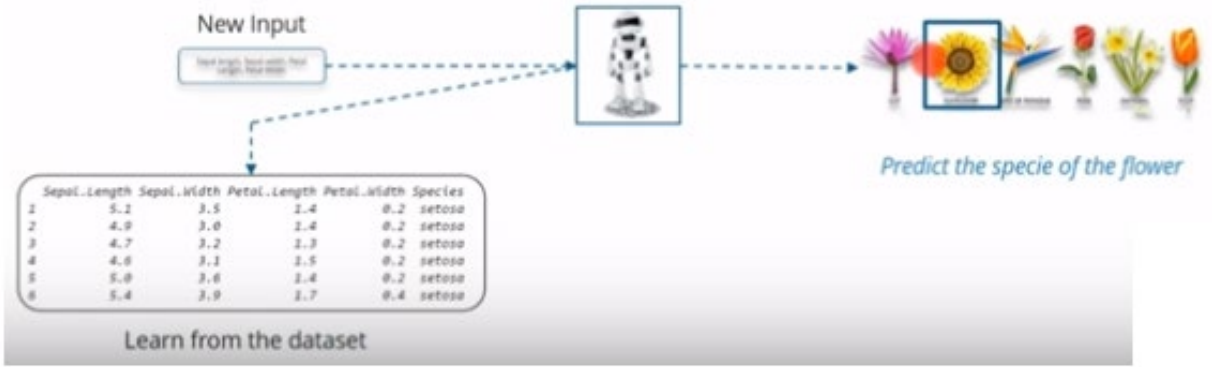
- Machine Learning is a type of AI that provides computers with the ability to learn without being explicitly programmed.
- Problem Statement : Determine the specie of the flower





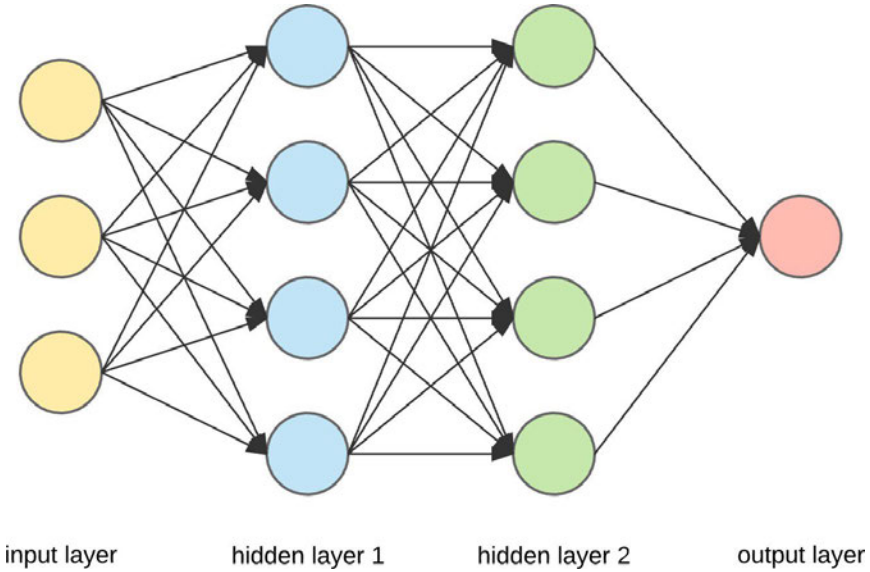
Machine Learning

- Machine Learning is a type of AI that provides computers with the ability to learn without being explicitly programmed.
- Problem Statement : Determine the specie of the flower



Deep Learning

Deep Learning is a subfield of Machine Learning that uses Artificial Neural Networks to learn from the data.



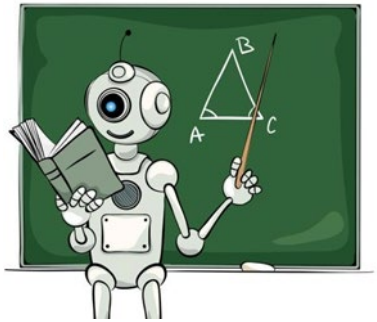
Machine Learning- Recap

Machine Learning is a technique to implement AI that can learn from the data by themselves without being explicitly programmed.

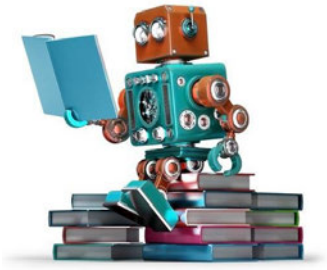
Types of Machine Learning

Machine Learning

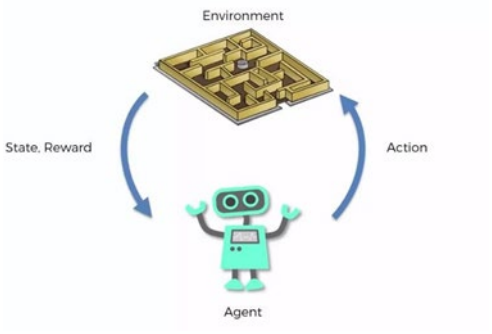
Supervised Learning



Unsupervised Learning



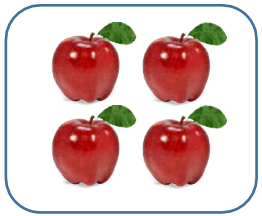
Reinforcement Learning



Supervised Learning

In Supervised Learning, the Machine Learning algorithm learns from ***Labelled Data***

Apples



Mangoes

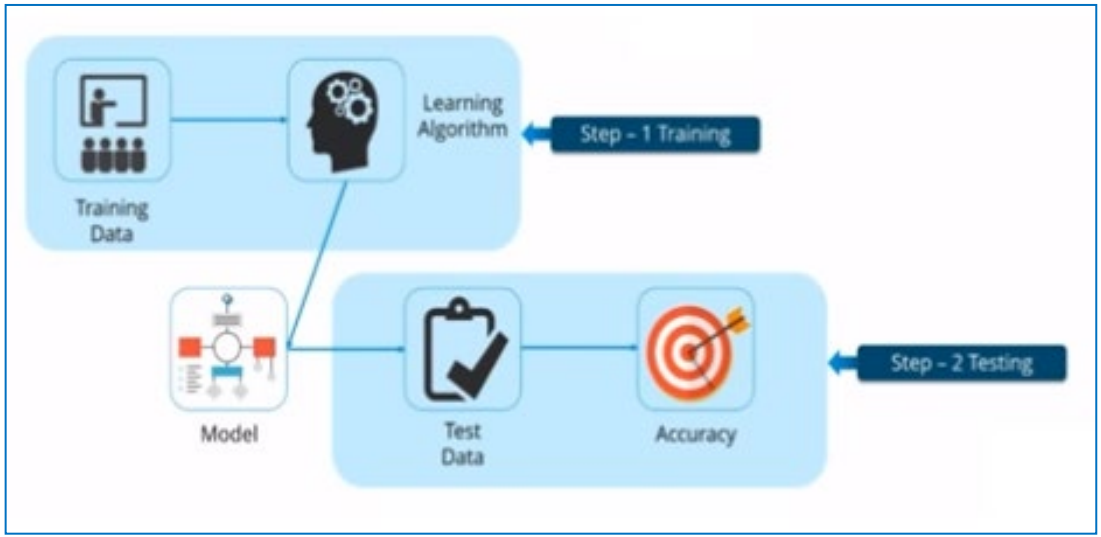
ML model

Apple

Unknown Image

Supervised Learning

- **Supervised Learning** – Where you have input variables (X) and out put variable (Y) and you use an algorithm to learn the mapping function from the input to the output

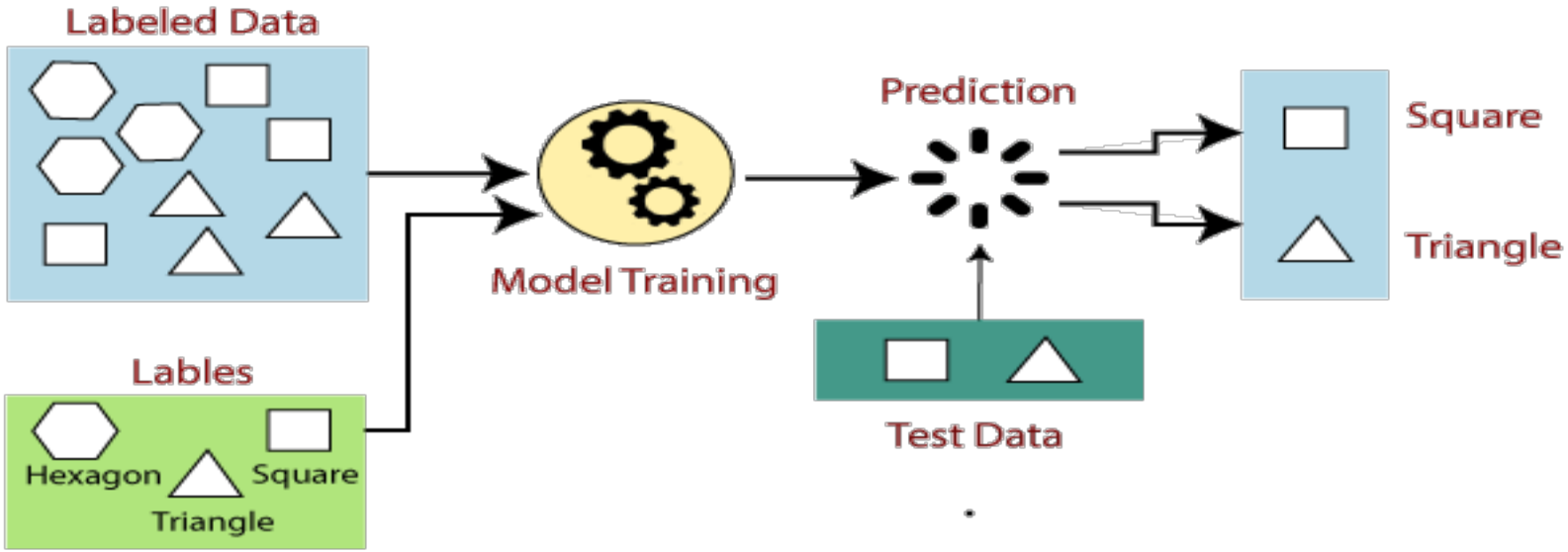


The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data.

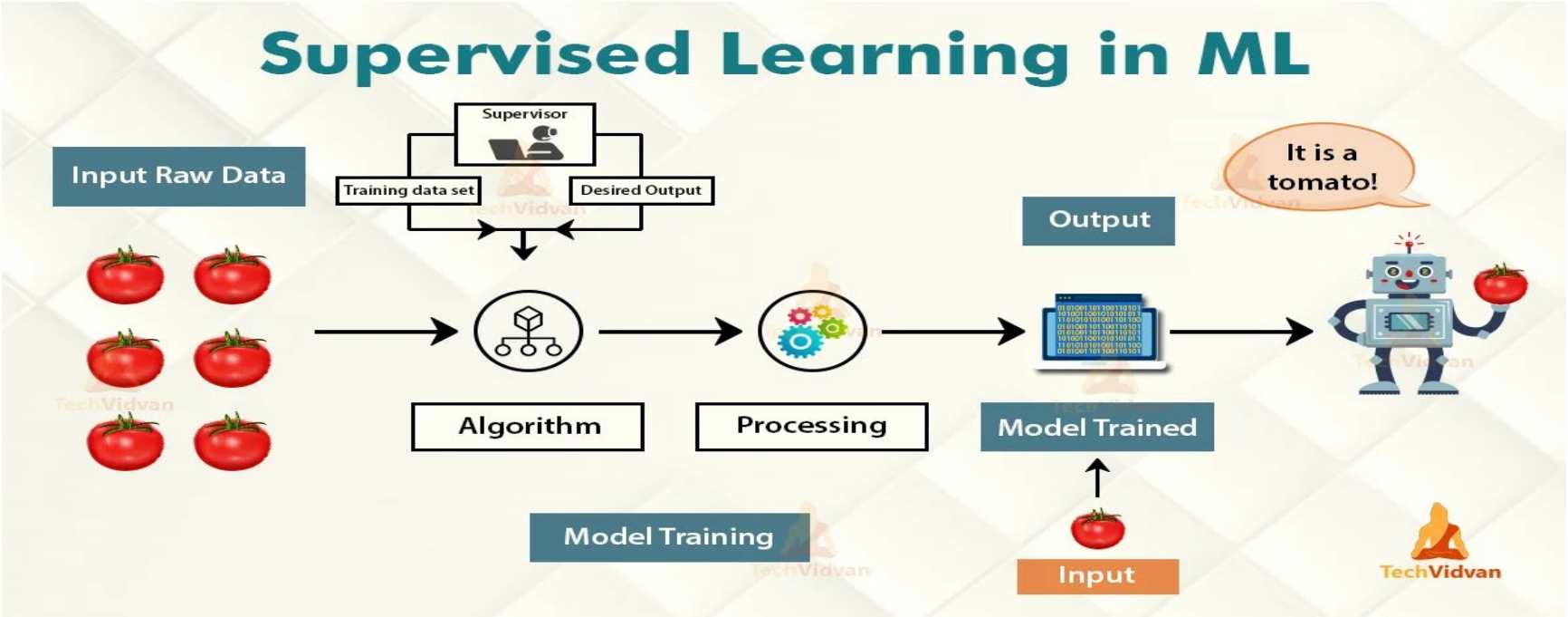
Supervised Learning

- Supervised Learning is a category of machine learning algorithms that are based upon the labelled data set.
- Predictive analytics is achieved for this category of algorithms where the outcome of the algorithm that is known as the dependent variable depends upon the value of independent data variables.
- It is based upon the training dataset, and it improves through iterations.

Supervised Learning

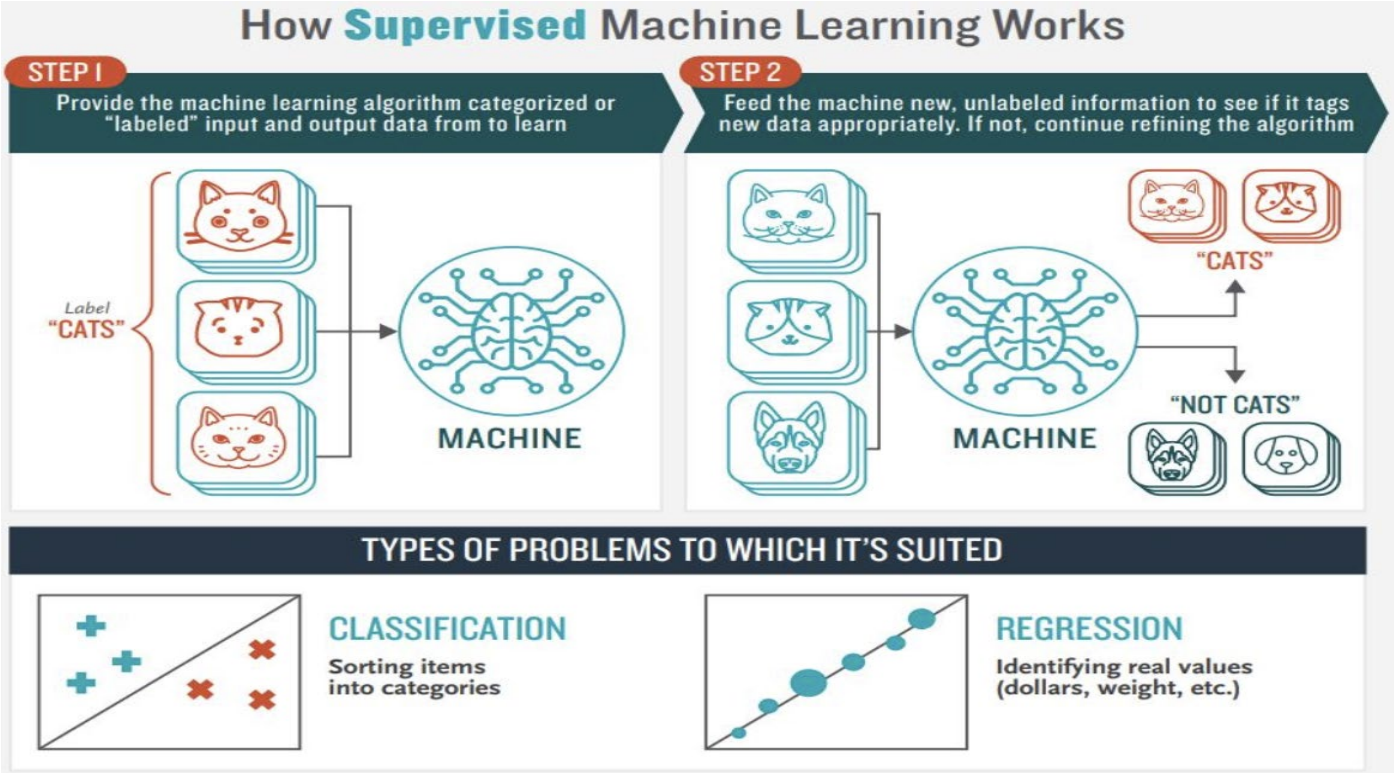


Supervised Learning



Courtesy: TechVidvan

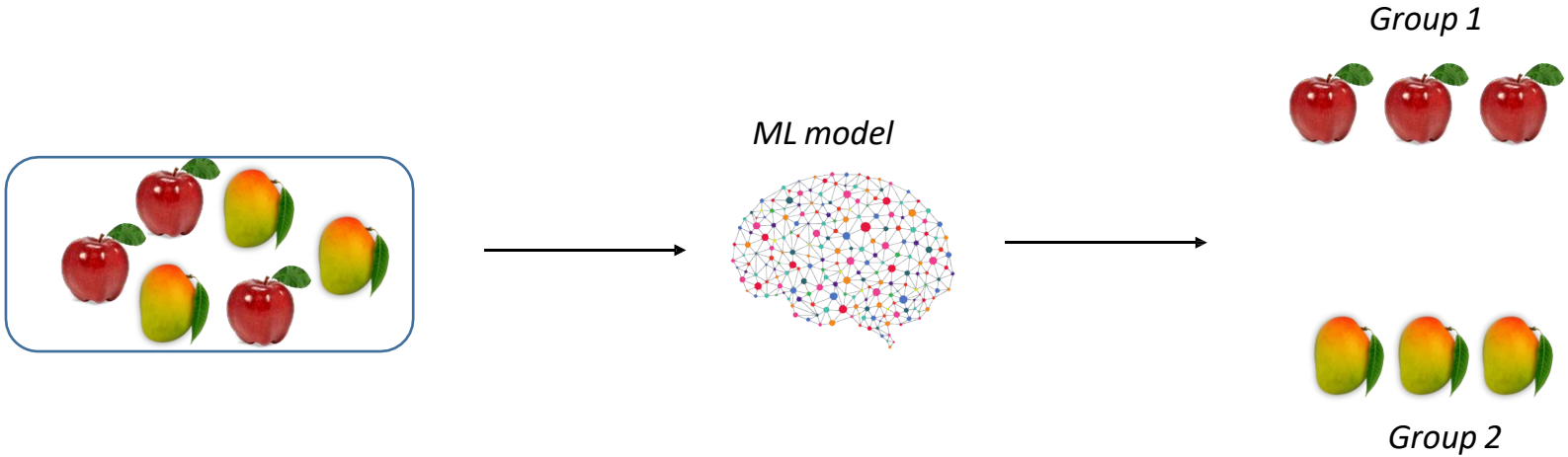
Supervised Learning



Courtesy: https://www.kaggle.com/discussions/getting-started/340445

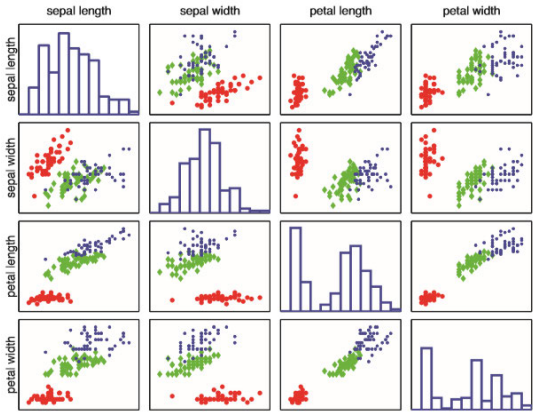
Unsupervised Learning

In Unsupervised Learning, the Machine Learning algorithm learns from ***Unlabelled Data***



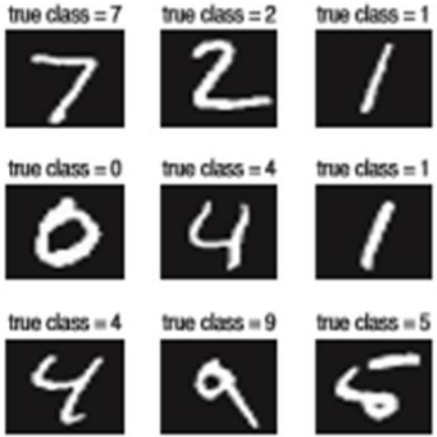
APPLICATIONS (SUPERVISED LEARNING)

Flower Classification



Handwriting Character recognition

Multiclass classification task



Email Spam Filtering

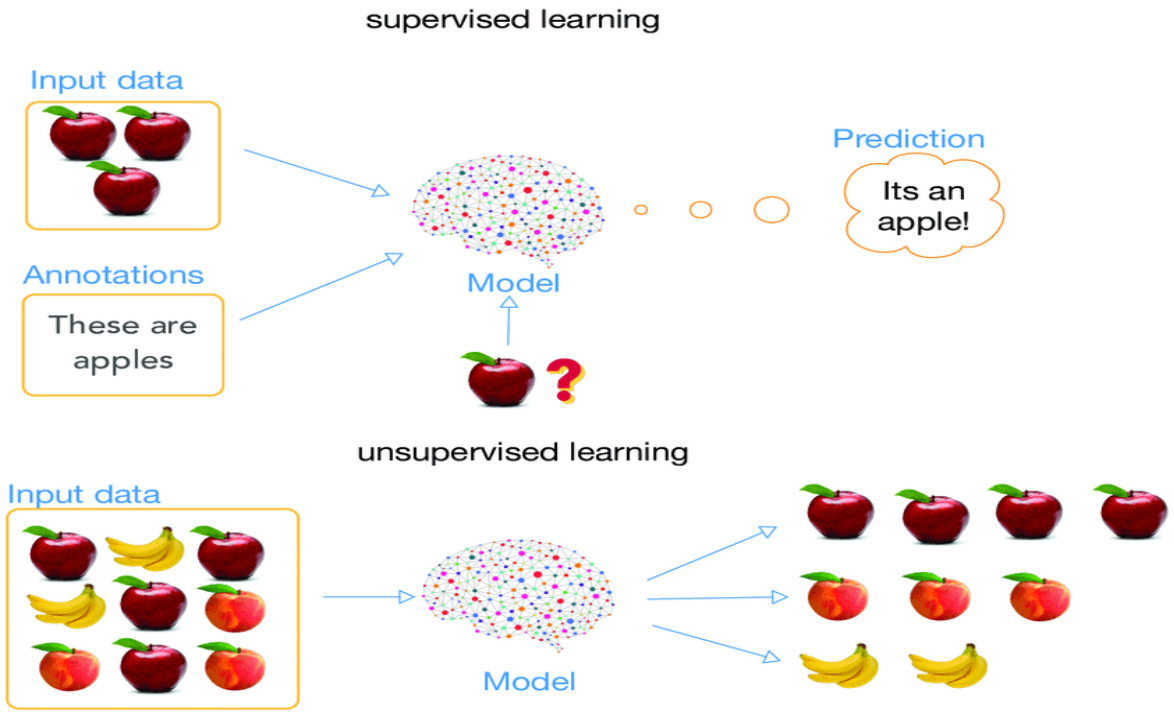
- A corpus of labeled emails, emails that are correctly marked as spam or not-spam
- To predict whether a new email belongs to either of the two categories

Predicting Real Estate

Email Spam Filtering

Weather Forecasting

Supervised vs Unsupervised



Unsupervised Learning

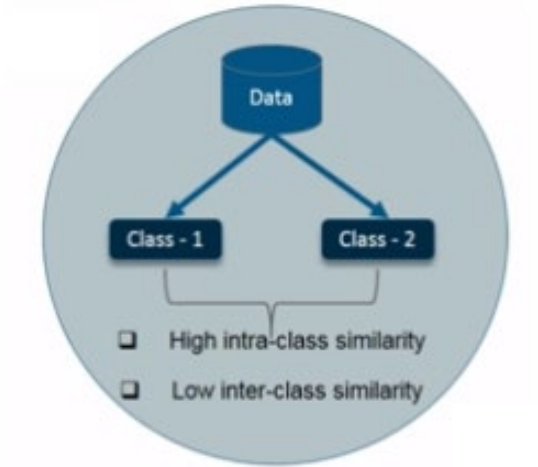
- Unsupervised ML is a method in which the machine is trained on unlabelled data without any guidance.

In other words.....

- For just given output data, without any inputs. The goal is to discover “interesting structure” in the data; this is sometimes called **knowledge discovery** .

Unsupervised Learning

- Is the training of a model using information that is neither classified nor labeled
- This model can be used to cluster the input data in classes on the basis of their statistical properties.



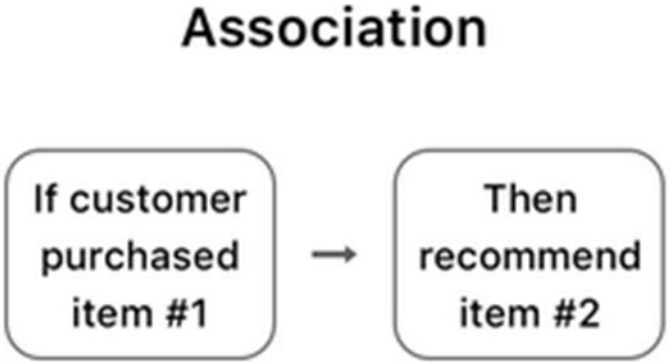
Categories

Clustering

- Grouping based on similarity

Association

Discovering patterns in data



Association

Association rules allow you to establish associations amongst data objects inside large databases.

This unsupervised technique is about discovering interesting relationships between variables in large databases.

For example, people that buy a new home most likely to buy new furniture.

Clustering

Clustering: is the assignment of a set of observations into subsets (called **clusters**) so that observations in the same **cluster** are similar in some sense.

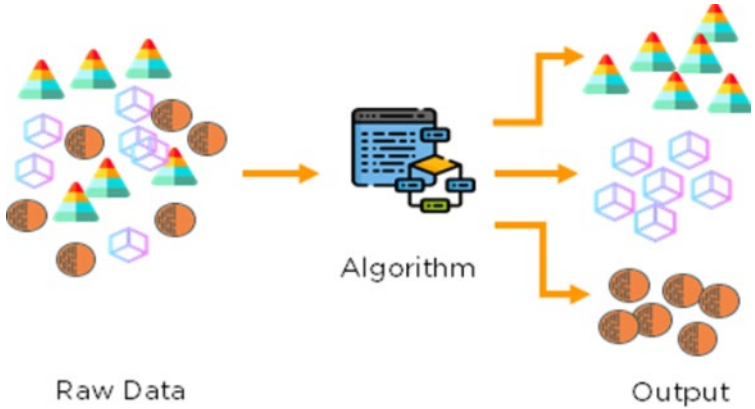
Clustering is a method of unsupervised **learning**, and a common technique for statistical data analysis used in many fields.

Data

- Unlabelled Input Data
- Patten Output Data (grouping)

Unsupervision of Training

- Needs **NO External Supervision** from training data



- I

Vegetable clustering

Clustering

sample

Cluster/group

Identifying fraudulent or criminal activity

- In this scenario, we are going to focus on fraudulent taxi driver behavior.
- **What is the problem:** You need to look into fraudulent driving activity. The challenge is how do you identify what is true and which is false?
- **How clustering works:** By analysing the GPS logs, the algorithm is able to group similar behaviors. Based on the characteristics of the groups you are then able to classify them into those that are real and which are fraudulent.

Algorithms

- K – means
- C- means
- Apriori algorithm

Reinforcement Learning

Reinforcement Learning is an area of Machine Learning concerned with how intelligent agents take actions in an environment to maximize its rewards.

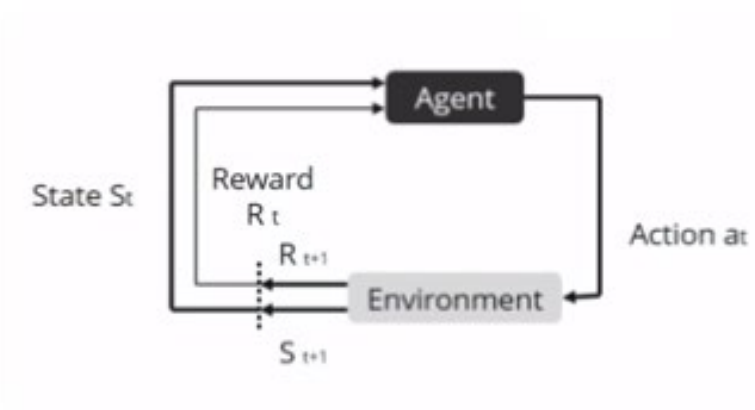
1. *Environment*
2. *Agent*
3. *Action*
4. *Reward*

Agent

Environment

Reinforcement Learning

- Is learning by interacting with a space or an environment.
- An RL agent learns from the consequences of its action rather than from being taught explicitly. It selects its actions on the basis of its past experiences(exploitation) and also by new choices (exploration).

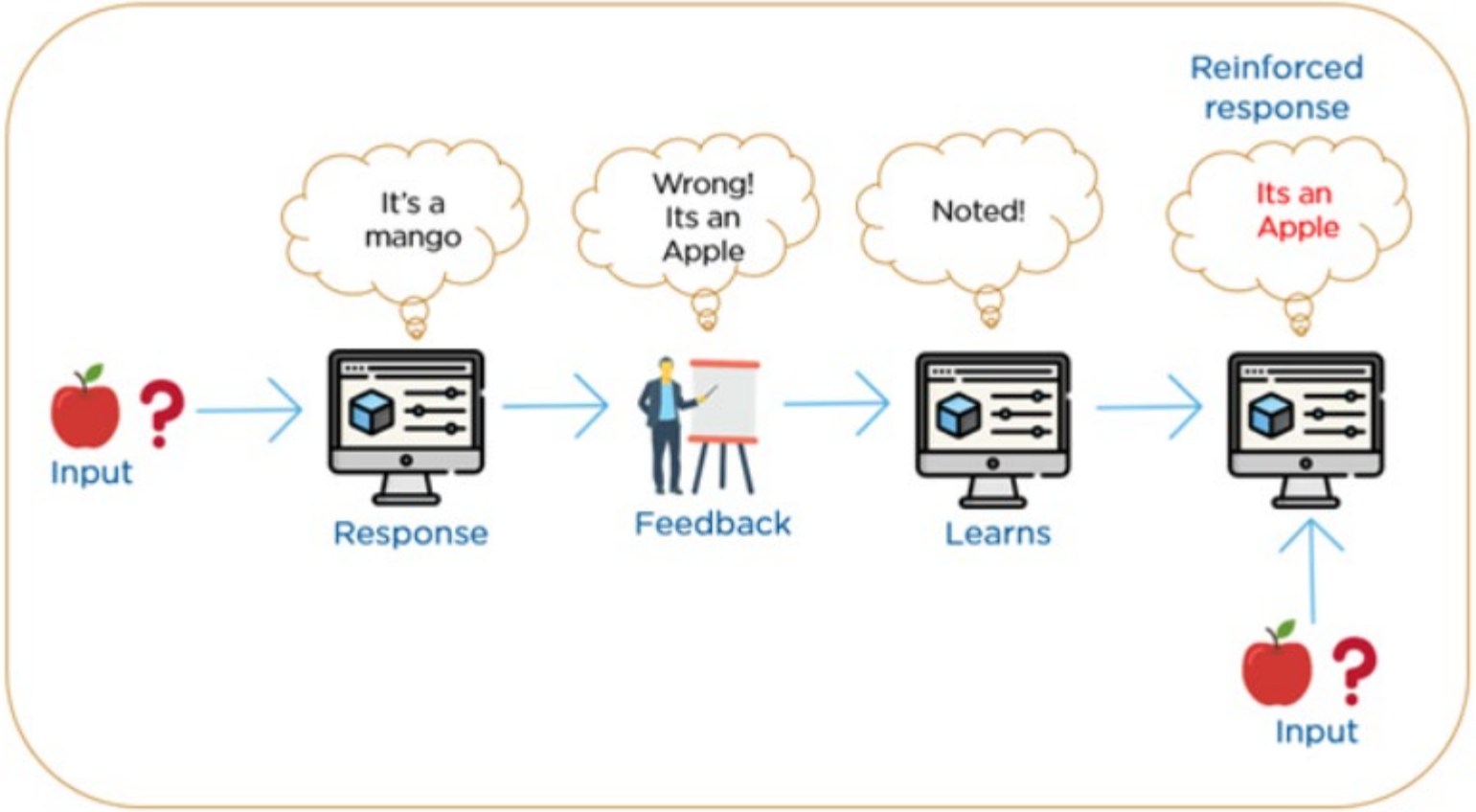


Reinforcement Learning

- Reinforcement learning (RL) is a type of ML which is all about taking suitable action to maximize reward in a particular situation.
- It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.
- RL means to establish or encourage a pattern of behavior.

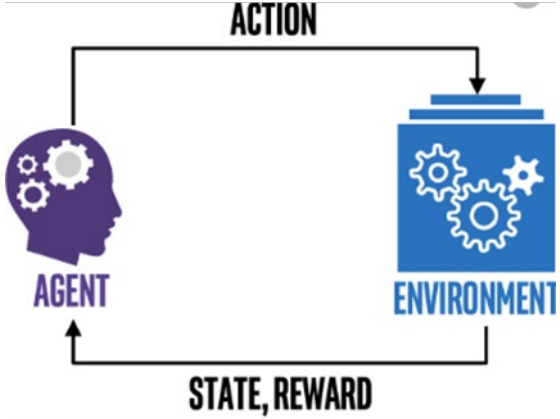
Reinforcement Learning is a subfield of machine learning that teaches an agent how to choose an action from its action space, within a particular environment, in order to maximize rewards over time.

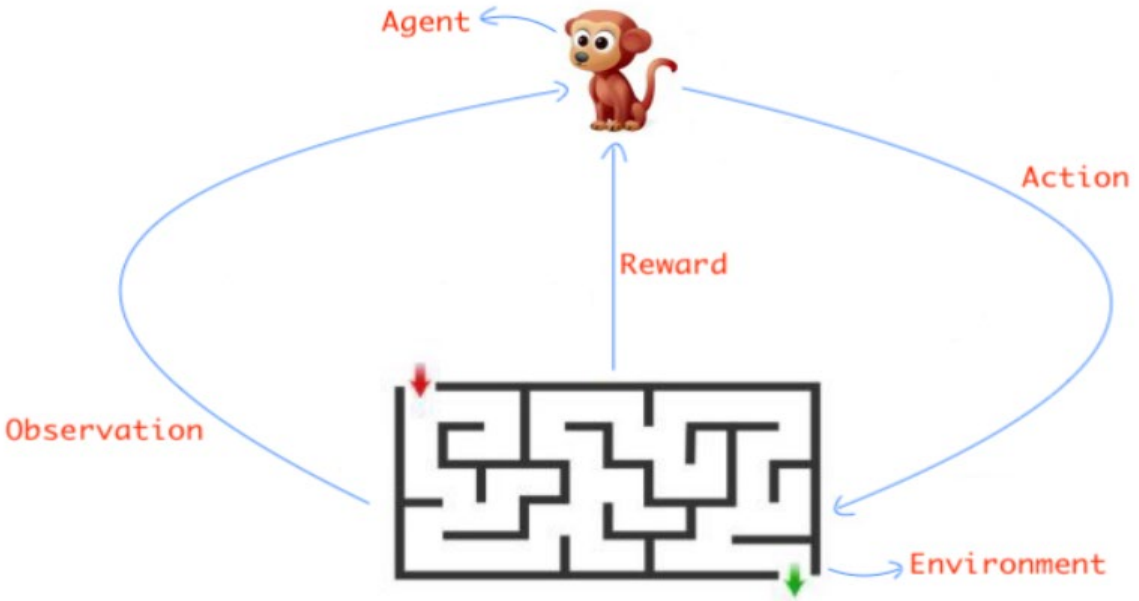
Reinforcement Learning



Essential Elements

- **Agent.** The program you train, with the aim of doing a job you specify.
- **Environment.** The world, real or virtual, in which the agent performs actions.
- **Action.** A move made by the agent, which causes a status change in the environment.
- **Rewards.** The evaluation of an action, which can be positive or negative.




Agent
Action
Reward
Observation
Environment

RL Works

- Observation of the environment
- Deciding how to act using some strategy
- Acting accordingly
- Receiving a reward or penalty
- Learning from the experiences and refining our strategy
- Iterate until an optimal strategy is found

Placement of Ads

- **Agent:** The program making decisions on how many ads are appropriate for a page.
- **Environment:** The web page.
- **Action:** One of three: 1. Putting another ad on the page. 2. Dropping an ad from the page. 3. Neither adding nor removing.
- **Reward:** Positive when revenue increases; negative when revenue drops.

Controlling A Walking Robot

- **Agent:** The program controlling a walking robot.
- **Environment:** The real world.
- **Action:** One out of four moves

1. Forward 2. Backward 3. Left and 4. Right.

- **Reward:** Positive when it approaches the target destination; negative when it wastes time, goes in the wrong direction or falls down.
- In this final example, a robot can teach itself to move more effectively by adapting its policy based on the rewards it receives.

Train and Test datasets in Machine Learning

- *Train and Test datasets are the two key concepts of machine learning, where the training dataset is used to fit the model, and the test dataset is used to evaluate the model .*

Training Dataset

- The *training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model .*

Test Dataset

- Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. *The test dataset is another subset of original data, which is independent of the training dataset .*

Train datasets in Machine Learning

- For **Unsupervised learning** , the training data contains unlabeled data points, i.e., inputs are not tagged with the corresponding outputs. Models are required to find the patterns from the given training datasets in order to make predictions.

- For **Supervised learning** , the training data contains labels in order to train the model and make predictions.

- The type of training data that we provide to the model is highly responsible for the model's accuracy and prediction ability. It means that the better the quality of the training data, the better will be the performance of the model.

Test datasets in Machine Learning

- Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world.
- We can also check and compare the testing accuracy with the training accuracy, which means how accurate our model is with the test dataset against the training dataset.
- If the accuracy of the model on training data is greater than that on testing data, then the model is said to have overfitting.
- The testing data should:
  - Represent or part of the original dataset.
  - It should be large enough to give meaningful predictions.

Need of Splitting dataset into Train and Test set

- Splitting the dataset into train and test sets is one of the important parts of data pre-processing, as by doing so, we can improve the performance of our model and hence give better predictability.

For splitting the dataset, we can use the train_test_split function of scikit-learn.


```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.2, random_state=0)
```

Need of Splitting dataset into Train and Test set

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.2, random_state=0)
```

x_train: It is used to represent features for the training data
x_test: It is used to represent features for testing data
y_train: It is used to represent dependent variables for training data
y_test: It is used to represent independent variable for testing data

In the train_test_split() function, we have passed four parameters.

The last parameter, random_state, is used to set a seed for a random generator so that you always get the same resul

Overfitting and Underfitting issues

- A model can be said as overfitted when it performs quite well with the training dataset but does not generalize well with the new or unseen dataset.
- The issue of overfitting occurs when the model tries to cover all the data points and hence starts caching noises present in the data. Due to this, it can't generalize well to the new dataset.
- Because of these issues, the accuracy and efficiency of the model degrade. Generally, the complex model has a high chance of overfitting.
- There are various ways by which we can avoid overfitting in the model, such as Using the Cross-Validation method, early stopping the training, or by regularization, etc.

Overfitting and Underfitting issues

- the model is said to be under-fitted when it is not able to capture the underlying trend of the data.
- It means the model shows poor performance even with the training dataset.
- In most cases, underfitting issues occur when the model is not perfectly suitable for the problem that we are trying to solve.
- To avoid the underfitting issue, we can either increase the training time of the model or increase the number of features in the dataset.

Cross-Validation in Machine Learning

Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. ***We can also say that it is a technique to check how a statistical model generalizes to an independent dataset*** .

Hence the basic steps of cross-validations are:

- Reserve a subset of the dataset as a validation set.
- Provide the training to the model using the training dataset.
- Now, evaluate model performance using the validation set. If the model performs well with the validation set, perform the further step, else check for the issues.

Methods used for Cross - Validation

1. **Validation Set Approach**
2. **Leave-P-out cross-validation**
3. **Leave one out cross-validation**
4. **K-fold cross-validation**
5. **Stratified k-fold cross-validation**

Validation Set Approach

- We divide our input dataset into a training set and test or validation set in the validation set approach.
- Both the subsets are given 50% of the dataset.
- It has one of the big disadvantages that we are just using a 50% dataset to train our model, so the model may miss out to capture important information of the dataset. It also tends to give the underfitted model.

Leave-P-out cross-validation

- In this approach, the p datasets are left out of the training data. It means, if there are total n datapoints in the original input dataset, then n-p data points will be used as the training dataset and the p data points as the validation set.
- This complete process is repeated for all the samples, and the average error is calculated to know the effectiveness of the model.
- There is a disadvantage of this technique; that is, it can be computationally difficult for the large p.

Leave one out cross-validation

- This method is similar to the leave-p-out cross-validation, but instead of p, we need to take 1 dataset out of training. It means, in this approach, for each learning set, only one datapoint is reserved, and the remaining dataset is used to train the model. This process repeats for each datapoint. Hence for n samples, we get n different training set and n test set. It has the following features:

- In this approach, the bias is minimum as all the data points are used.
- The process is executed for n times; hence execution time is high.
- This approach leads to high variation in testing the effectiveness of the model as we iteratively check against one data point.

K-Fold Cross-Validation

- K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called folds. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set. This approach is a very popular CV approach because it is easy to understand, and the output is less biased than other methods.
- The steps for k-fold cross-validation are:
    - Split the input dataset into K groups
    - For each group:
    - Take one group as the reserve or test data set.
    - Use remaining groups as the training dataset
    - Fit the model on the training set and evaluate the performance of the model using the test set.

- I

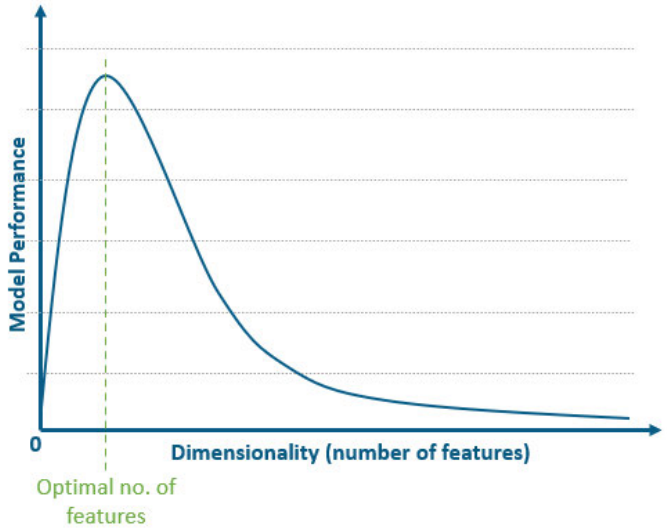
Curse of Dimensionality

Curse of Dimensionality

- The curse of dimensionality basically refers to the difficulties a machine learning algorithm faces when working with data in the higher dimensions.

- This happens because when you add dimensions (features), the minimum data requirements also increase rapidly.

- This means, that as the number of features (columns) increases, you need an exponentially growing number of samples (rows) to have all combinations of feature values well-represented in our sample.



Curse of Dimensionality

With the increase in the data dimensions, your model –

- would also increase in complexity.
- would become increasingly dependent on the data it is being trained on.
- This leads to overfitting of the model, so even though the model performs really well on training data, it fails drastically on any real data.

Dimensionality Reduction

- Data features are usually correlated. Hence, the higher dimensional data is dominated by a rather small number of features. If we can find a subset of the ***superfeatures*** that can represent the information just as well as the original dataset, we can remove the curse of dimensionality!

- – a process of reducing the dimension of your data to a few principal features.

- Fewer input dimensions often correspond to a simpler model, referred to as its **degrees of freedom** . A model with larger degrees of freedom is more prone to overfitting. So, it is desirable to have more generalized models, and input data with fewer features.

Why is Dimensionality Reduction necessary?

- **Avoids overfitting** – the lesser assumptions a model makes, the simpler it will be.
- **Easier computation** – the lesser the dimensions, the faster the model trains.
- **Improved model performance** – removes redundant features and noise, lesser misleading data improves model accuracy.
- Lower dimensional data requires **less storage space** .
- Lower dimensional data can **work with other algorithms** that were unfit for larger dimensions.

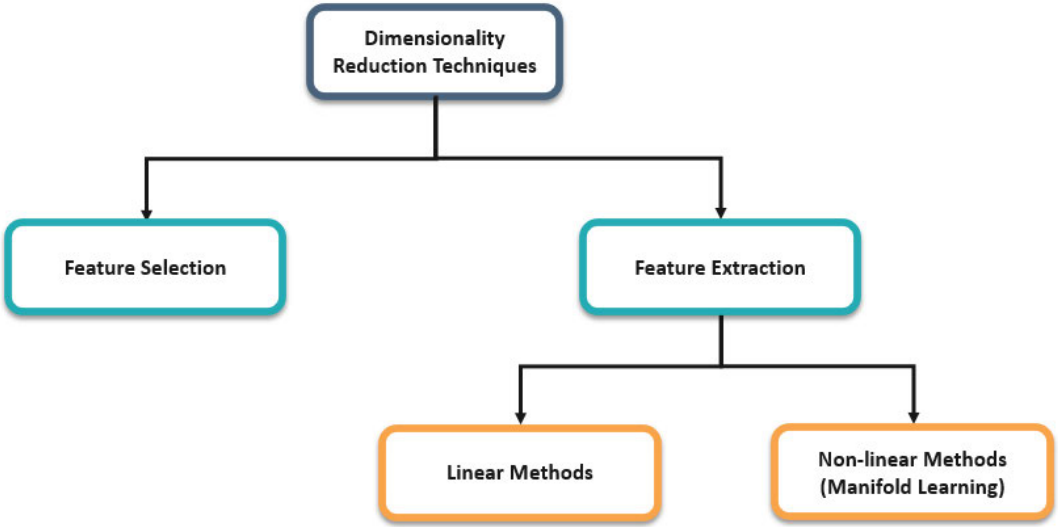
How is Dimensionality Reduction done?

DR Techniques are divided into two broad categories:

- **Feature Selection:** Choosing the most important features from the data
- **Feature Extraction:** Combining features to create new *superfeatures*

Feature Selection
-Low Variance Filter
-High correlation Filter
-Multicollinearity
-Feature Ranking
-Forward Selection

Feature Extraction
– Principal Component Analysis (PCA)
-Factor Analysis
-Independent Component Analysis



Curse of Dimensionality

Dimension

Features

Attributes

SCENARIO

Real estate rate prediction

Features

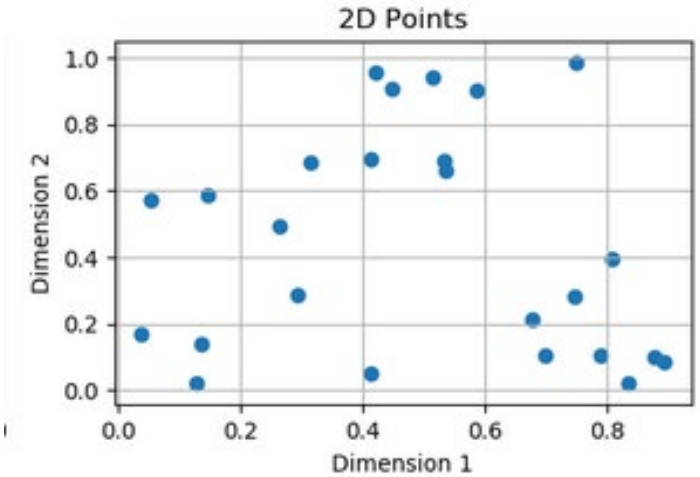
- House age
- No. of rooms
- No. of bed rooms
- No. of floors
- Type
- Location
- Cost

One Dimension

Number of Rooms

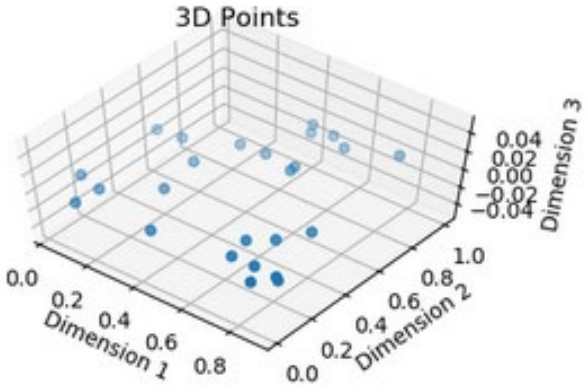
Two Dimension

Number of Rooms vs Costs



N- Dimension

Number of Rooms
Vs
Number of Floors
Vs
Costs



Definition

The curse of dimensionality refers to the phenomena that occur when classifying, organizing, and analyzing high dimensional data that does not occur in low dimensional spaces, specifically the issue of data sparsity and "closeness" of data.

Cont..

- Sparsity of data occurs when moving to higher dimensions.
- The volume of the space represented grows so quickly that the data cannot keep up and thus becomes sparse.

Cont..

- As the data space seen above moves from one dimension to two dimensions and finally to three dimensions, the given data fills less and less of the data space.
- In order to maintain an accurate representation of the space, the data for analysis **grows exponentially** .
- **Issues** with sorting or classifying the data. In low dimensional spaces, data may seem very similar but the higher the dimension the further these data points may seem to be.
- **Infinite Features Requires Infinite Training**

Drawbacks (Summary)

- As the dimension  data become more sparse.
- Hard to generalize
- Need more training data
- If the dimension  ,every data point is equidistant from all other points.

Bias and Variance in Machine Learning

- **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.
- **Noise:** Noise is unnecessary and irrelevant data that reduces the performance of the model.
- **Bias:** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.
- **Variance:** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset , then variance occurs.

Bias and Variance in Machine Learning

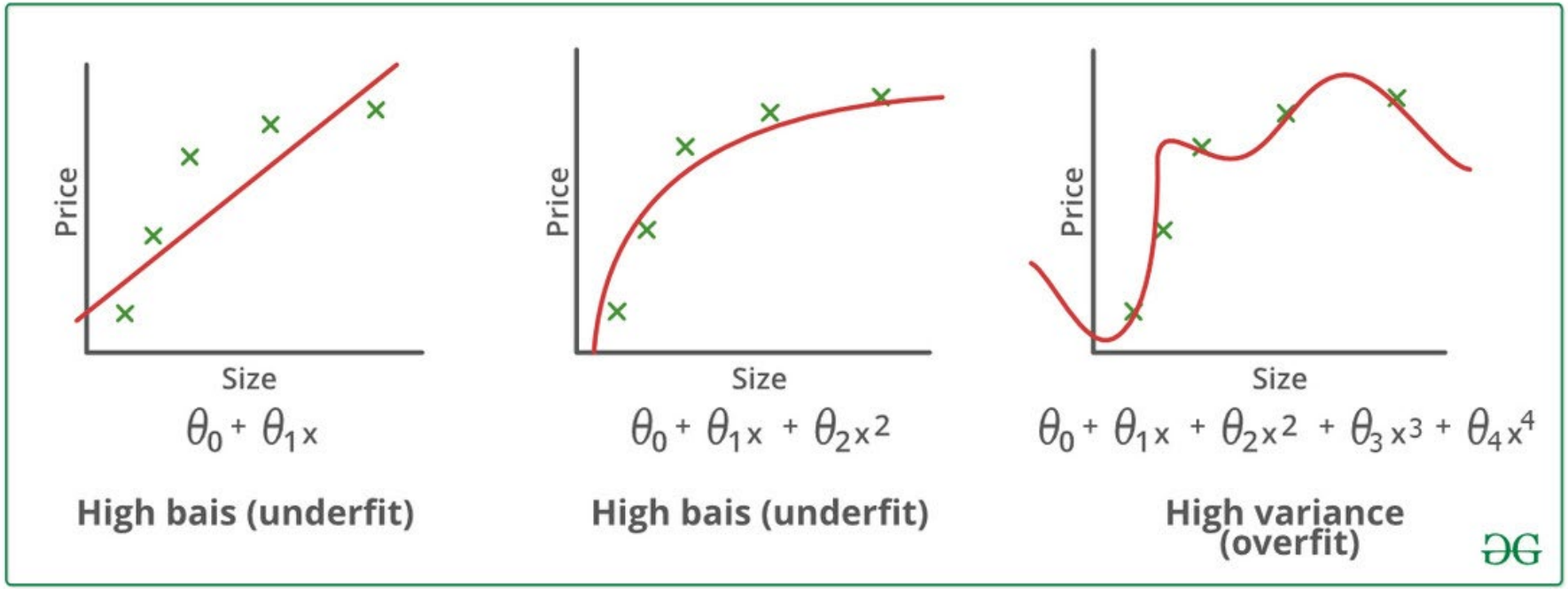
- **Bias**: It is actually the error rate of the training data. When the error rate has a high value, we call it High Bias and when the error rate has a low value, we call it low Bias.

- **Variance** : The difference between the error rate of training data and testing data is called variance. If the difference is high then it’s called high variance and when the difference in errors is low then it’s called low variance. Usually, we want to make a low variance for generalized our model.

Bias and Variance in Machine Learning

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Bias and Variance



Underfitting

Underfitting = High bias+Low variance

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.

- Underfitting destroys the accuracy of our machine-learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.

Reasons for Underfitting

- High bias and low variance.
- The size of the training dataset used is not enough.
- The model is too simple.
- Training data is not cleaned and also contains noise in it.

Handling Underfitting

•Get more training data.
•Increase the size or number of parameters in the model.
•Increase the complexity of the model.
•Increasing the training time, until cost function is minimised.

Techniques to Reduce Underfitting

- Increase model complexity.
- Increase the number of features, performing feature engineering.
- Remove noise from the data.
- Increase the number of epochs or increase the duration of training to get better results.

Overfitting

Overfitting = Low bias+High variance

- A statistical model is said to be overfitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance.
- Then the model does not categorize the data correctly, because of too many details and noise
- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

Reasons for Overfitting:

- High variance and low bias.
- The model is too complex.
- The size of the training data.

Techniques to Reduce Overfitting

- Increase training data.
- Reduce model complexity.
- Early stopping during the training phase
- Ridge Regularization and Lasso Regularization.
- Use dropout for neural networks to tackle overfitting.

Handling Overfitting

1. Cross-validation

2. Regularization

This is a form of regression, that regularizes or shrinks the coefficient estimates towards zero. This technique discourages learning a more complex model.

3. **Early stopping**

When training a learner with an iterative method, you stop the training process before the final iteration. This prevents the model from memorizing the dataset.

4. Pruning (applies to decision trees)

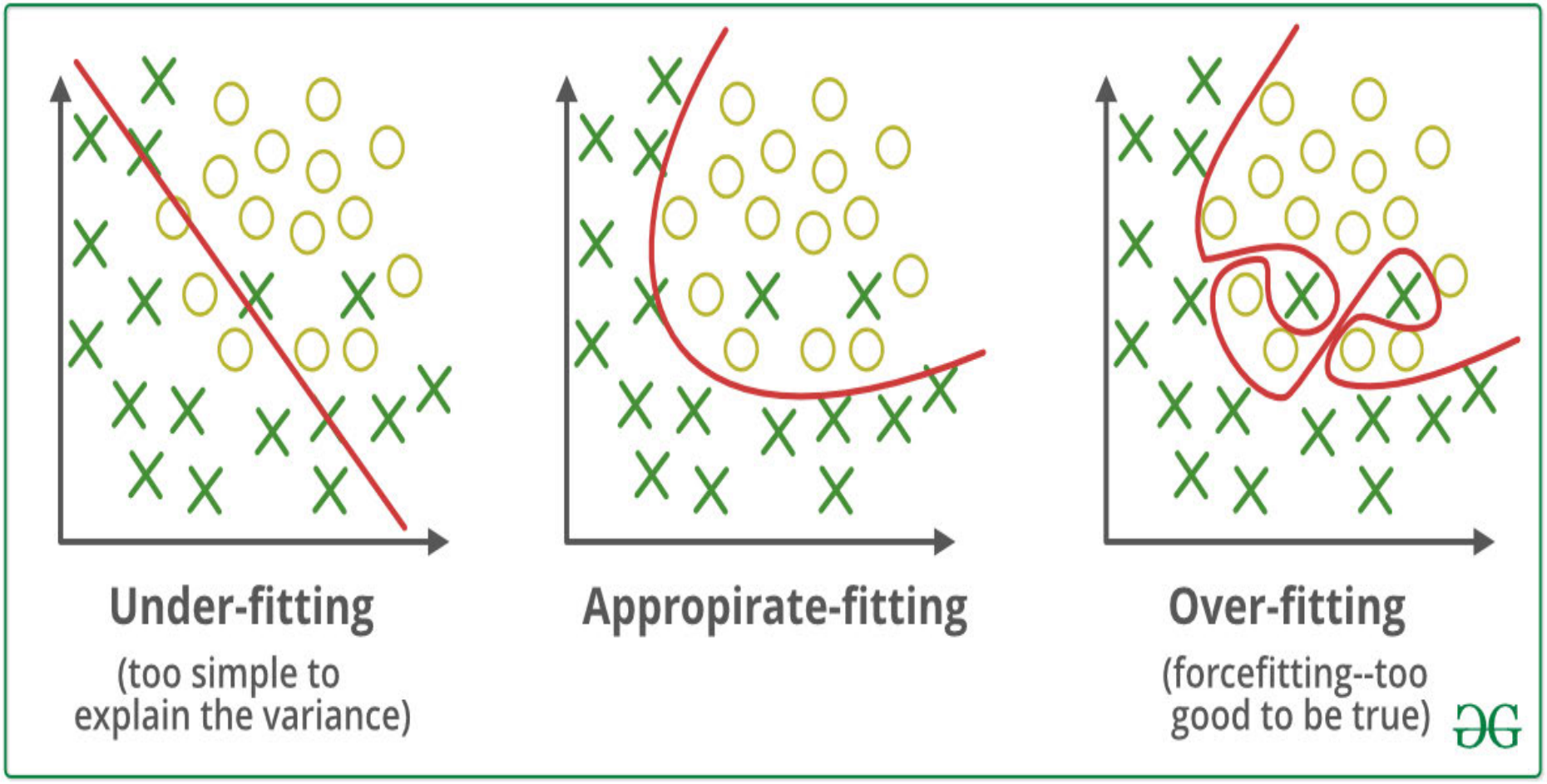
Pre-pruning: Stop 'growing' the tree earlier before it perfectly classifies the training set.

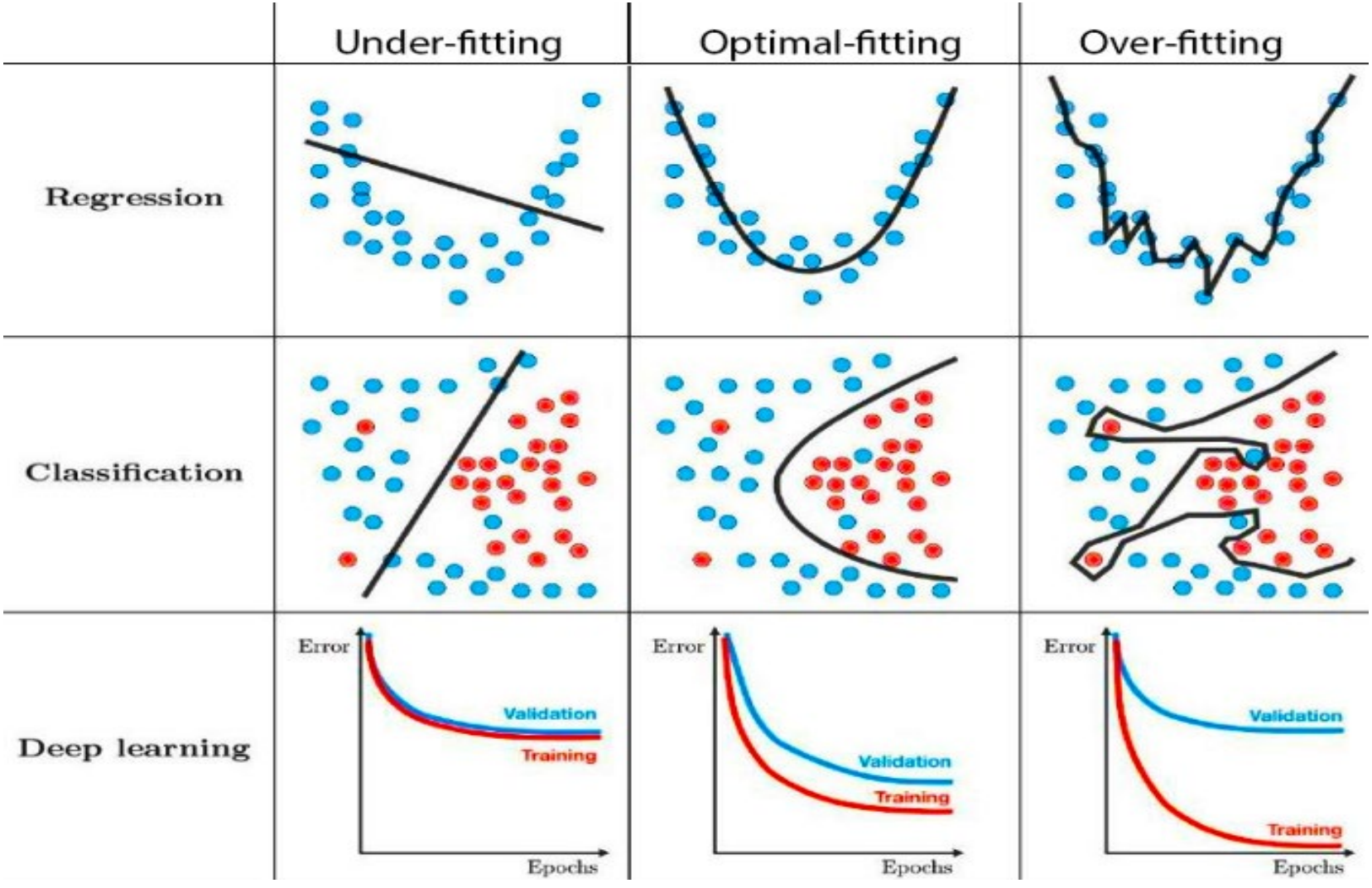
Post-pruning: Allows the tree to 'grow', perfectly classify the training set and then post prune the tree.

5. Dropout

This is a technique where randomly selected neurons are ignored during training.

6. Regularize the weights.


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)


Under-fitting
Optimal-fitting
Over-fitting
Regression
Classification
Deep learning
Error
Validation
Training
Epochs
Error
Validation
Training
Epochs
Error
Validation
Training
Epochs

Regularization

- It is a technique to prevent the model from overfitting by adding extra information to it.

- This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

- It mainly regularizes or reduces the coefficient of features toward zero.

- "*In regularization technique, we reduce the magnitude of the features by keeping the same number of features.*"

How does Regularization Work?

- Regularization works by adding a penalty or complexity term to the complex model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + b$$

- Y represents the value to be predicted
- X1, X2, …Xn are the features for Y.
- β0,β1,…..βn are the weights or magnitude attached to the features, respectively. Here represents the bias of the model, and b represents the intercept.

Linear regression models try to optimize the β0 and b to minimize the cost function. The equation for the cost function for the linear model is given below:

How does Regularization Work?

Linear regression models try to optimize the β0 and b to minimize the cost function. The equation for the cost function for the linear model is given below:

$$\sum_{i=1}^{M}\left(y_i - y'_i\right)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * Xij\right)^2$$

we will add a loss function and optimize parameter to make the model that can predict the accurate value of Y. The loss function for the linear regression is called as **RSS or Residual sum of squares.**

Techniques of Regularization

- **Ridge Regression**
- **Lasso Regression**

Ridge Regression

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
- In this technique, the cost function is altered by adding the penalty term to it.
- The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j{}^2$$

Ridge Regression

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j{}^2$$

- the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.
- if the values of **λ tend to zero, the equation becomes the cost function of the linear regression model.** Hence, for the minimum value of λ, the model will resemble the linear regression model.
- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
- It helps to solve the problems if we have more parameters than samples.

Lasso Regression:

- It stands for **Least Absolute and Selection Operator.**
- It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.
- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as **L1 regularization.** The equation for the cost function

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|^{\square}$$

- Some of the features in this technique are completely neglected for model evaluation.
- Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

Elastic Net Regression

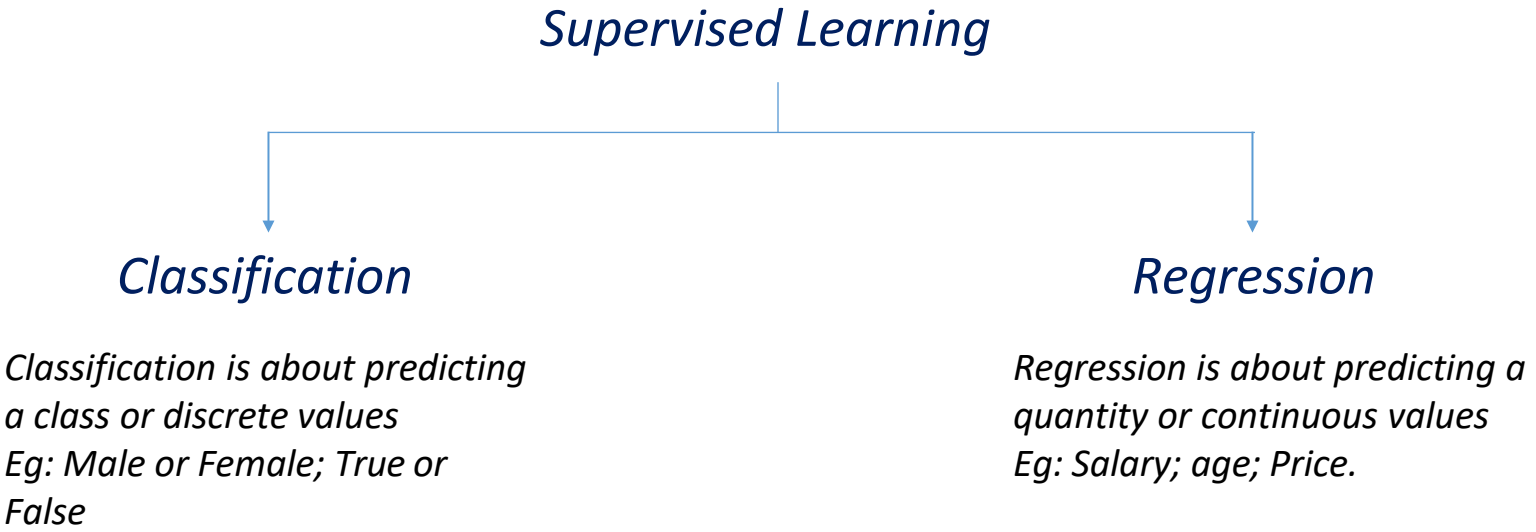
- This model is a combination of L1 as well as L2 regularization.
- That implies that we add the absolute norm of the weights as well as the squared measure of the weights.
- With the help of an extra hyperparameter that controls the ratio of the L1 and L2 regularization.

$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left((1-\alpha)\sum_{i=1}^{m}|w_i| + \alpha\sum_{i=1}^{m}w_i^2\right)$$

Key Difference between Ridge Regression and Lasso Regression

- **Ridge regression** is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.
- **Lasso regression** helps to reduce the overfitting in the model as well as feature selection.

Types of Supervised Learning



Types of Supervised Learning

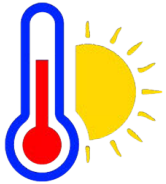
Classification:

Dog

Cat

(Dog or Cat)

Regression:



Temperature

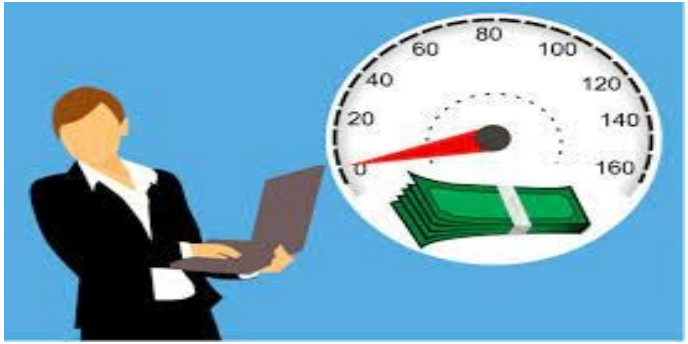
Rainfall in cm

Rainfall in cm

Regression

- These problems are used for **continuous data** .
- For example, predicting the price of a piece of land in a city, given the area, location, number of rooms, etc. And then the input is sent to the machine for calculating the price of the land according to previous examples.
    - Linear Regression
    - Nonlinear Regression
    - Bayesian Linear Regression

Regression Problems



Regression

- Examples:
  - person’s age, height, or income
  - the value of a house
  - the price of a stock.

Regression

- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
- It is mainly used for **prediction, forecasting, time series modeling , and determining the causal-effect relationship between variables** .

Regression

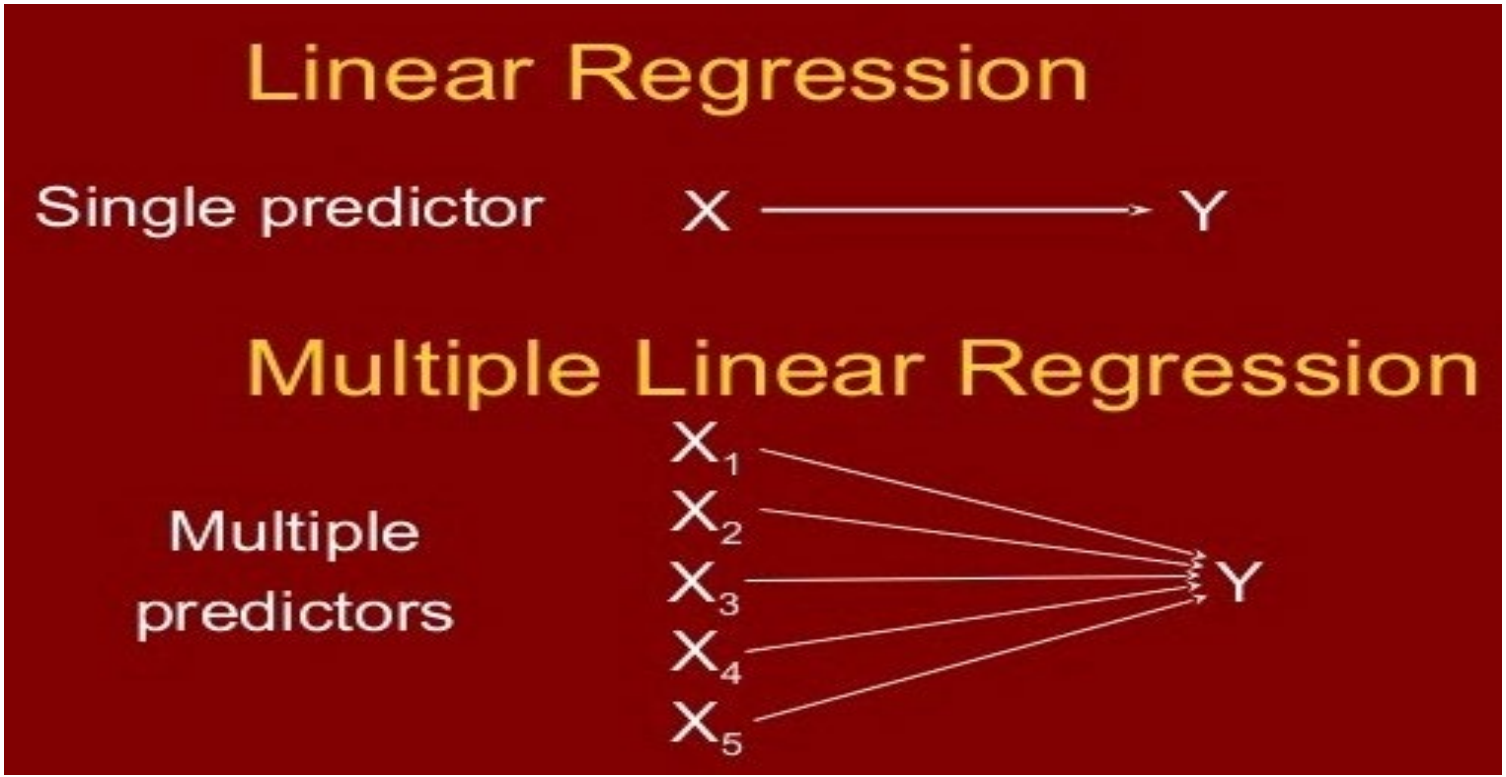
- *Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum.*

Linear Regression

- **Simple Linear Regression**
  - If there is only one input variable (x)
- **Multiple linear regression**
  - If there is more than one input variable

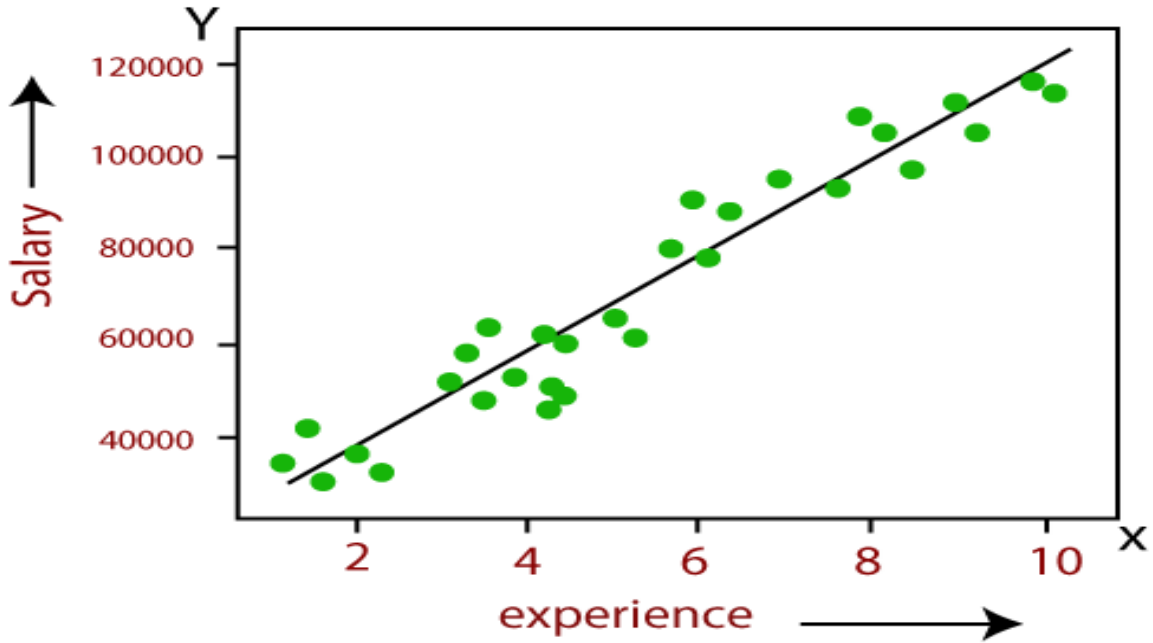
    i.e., relationship between one continuous dependent variable and two or more independent variables.

Linear Regression



Linear Regression

Predicting the salary of an employee on the basis of the year of experience.



Linear Regression

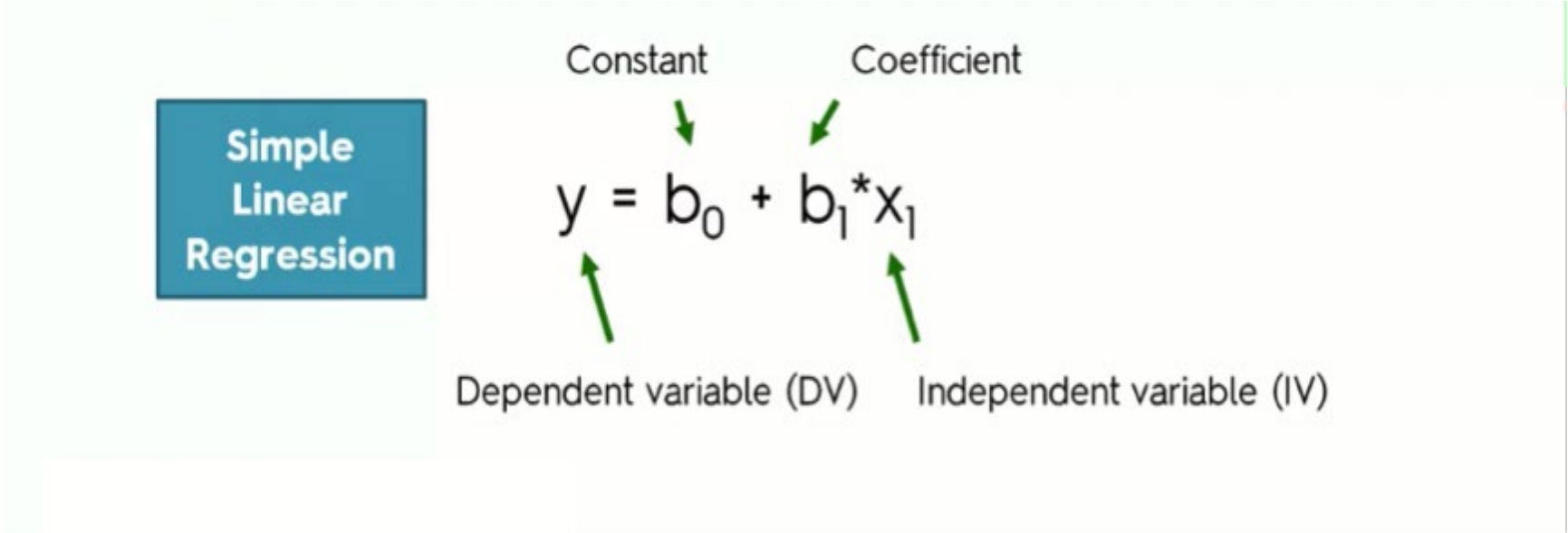
The mathematical equation for Linear regression:

$$Y = aX + b$$

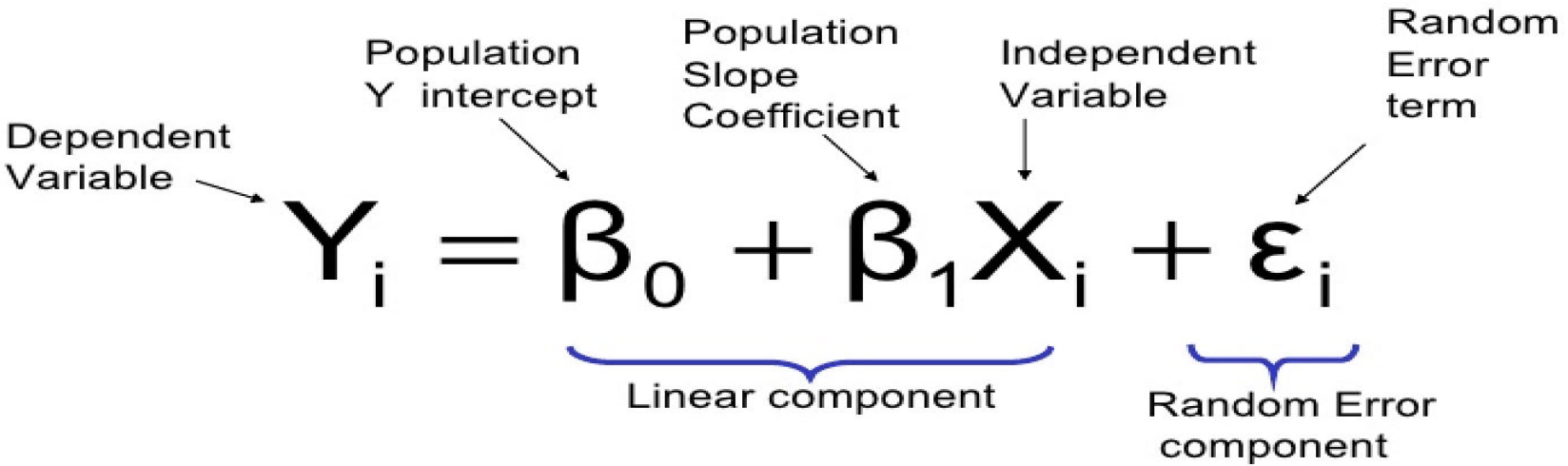
- Y = dependent variables (target variables)
- X= Independent variables (predictor variables)
- a and b are the linear coefficients

The regression dependent variable can be called as outcome variable or criterion variable or an endogenous variable. The independent variable can also be called an exogenous variable.

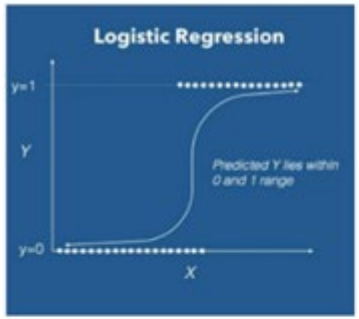
Linear Regression

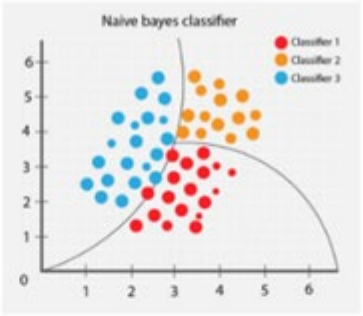


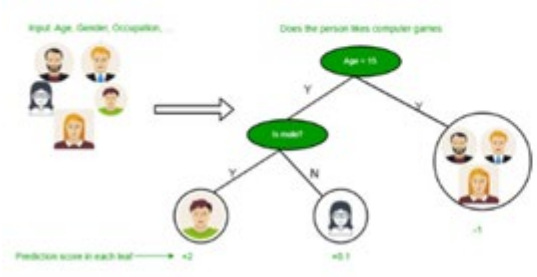
Linear Regression



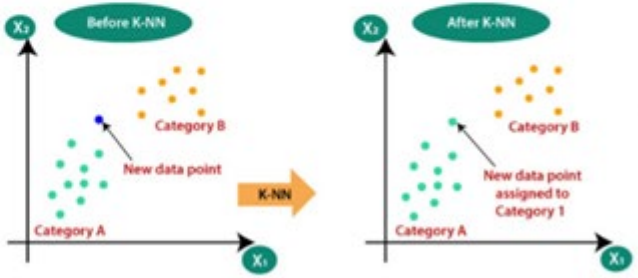
Classification





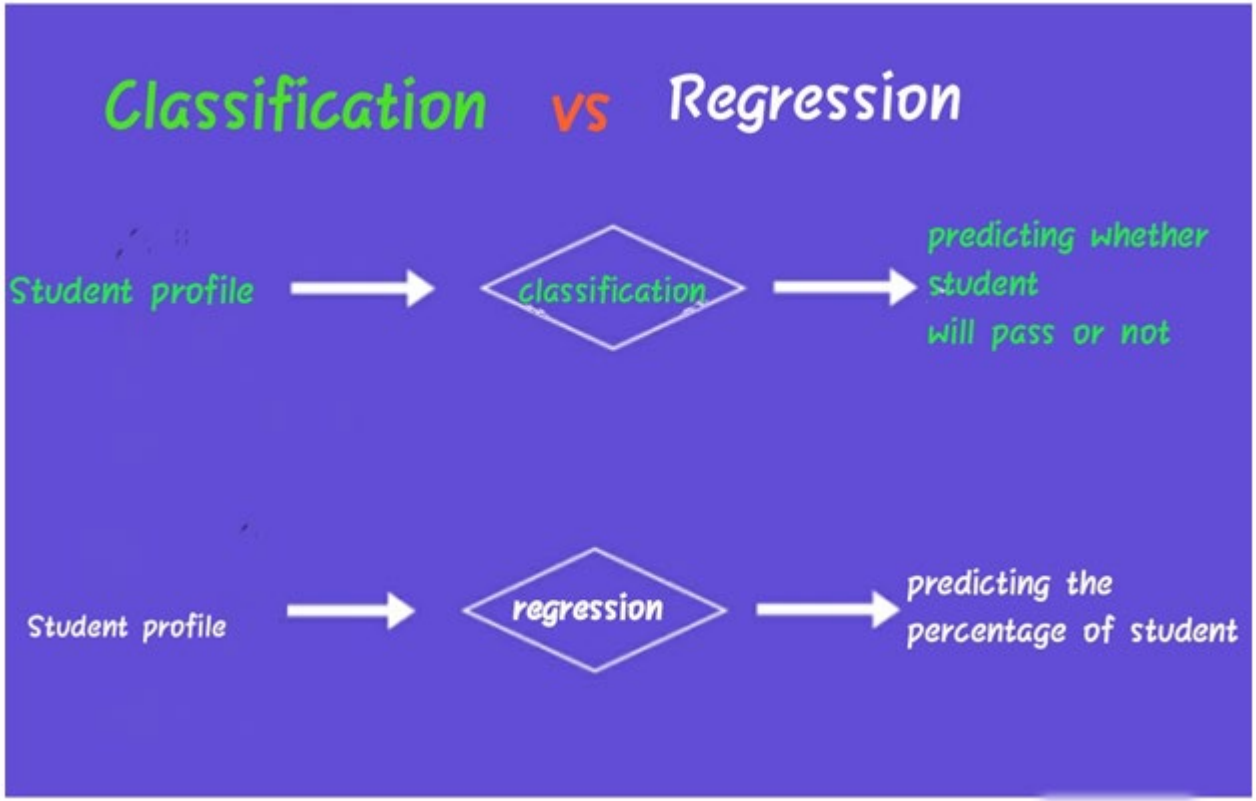


Decision Tree

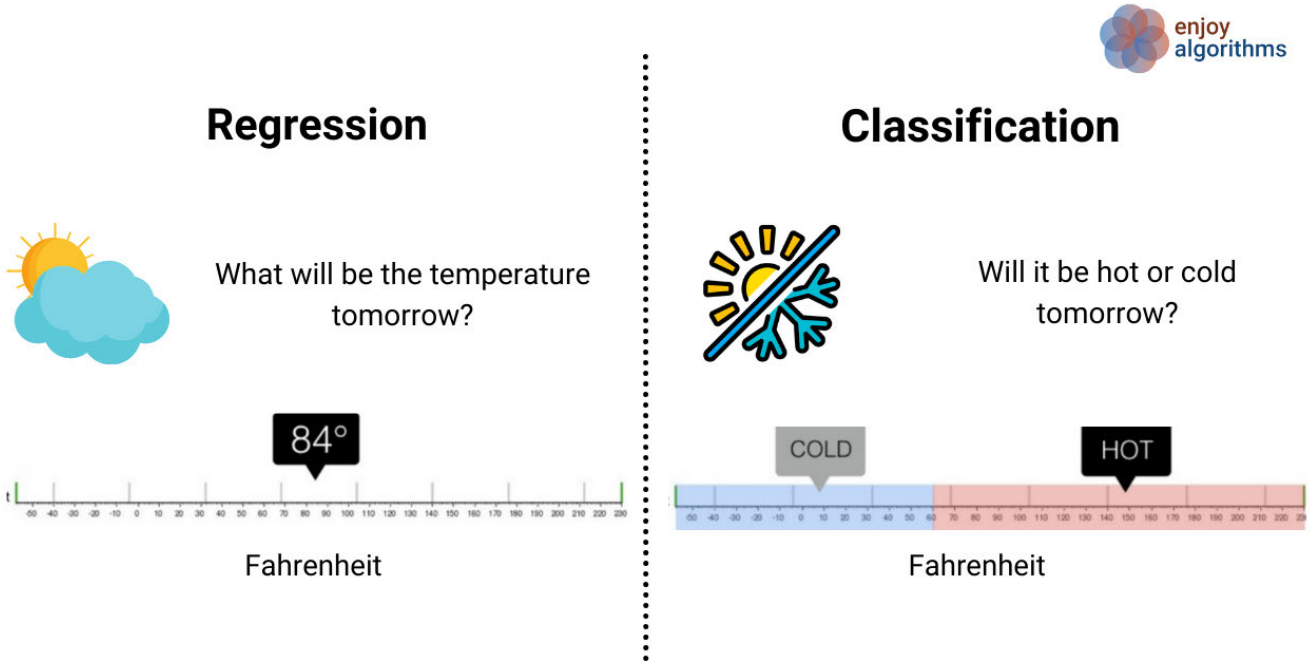


K-Nearest Neighbors

Classification vs Regression

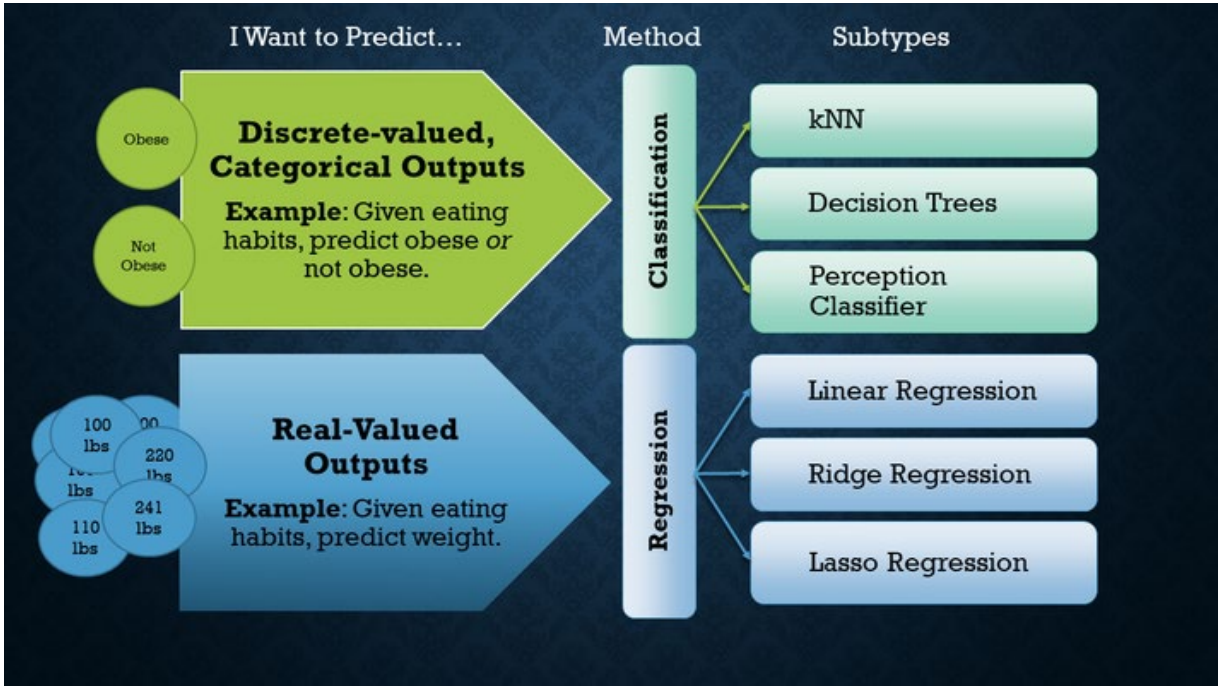


Regression vs Classification



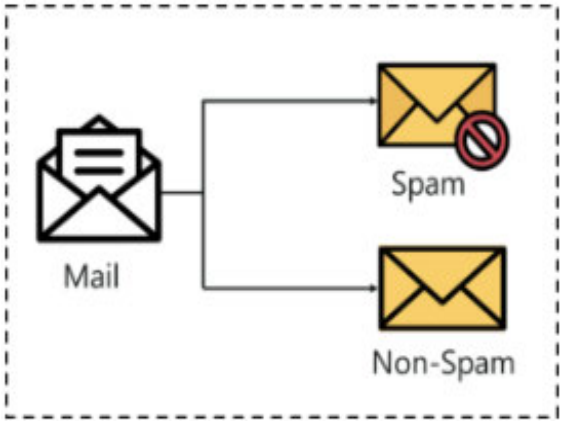
Courtesy: https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

I want to Predict….



Courtesy: https://www.datasciencecentral.com/classification-vs-regression/

What is Classification In Machine Learning



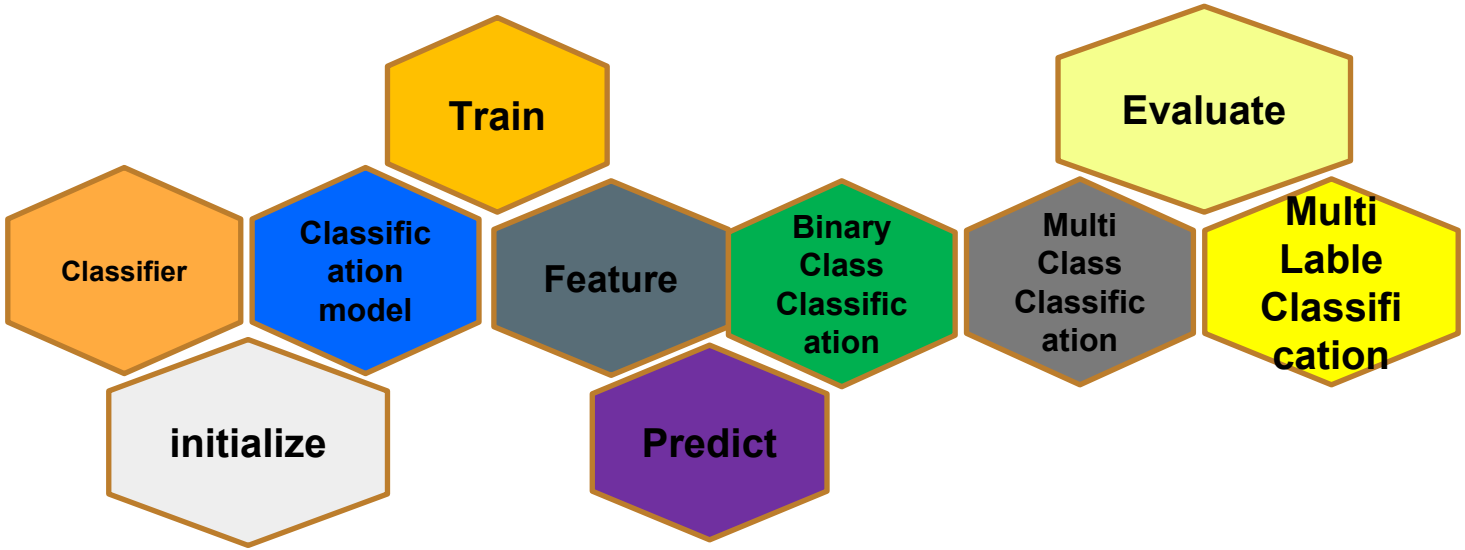
Classification is a process of categorizing a given set of data into classes.

It can be performed on both structured or unstructured data.

The process starts with predicting the class of given data points.

The classes are often referred to as target, label or categories.

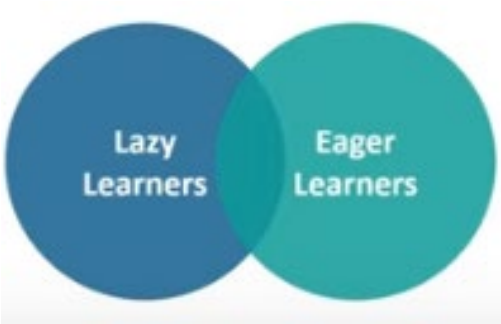
Classification Terminologies



Types of Learners in Classification

Lazy Learners –

simply store the training data and wait until a testing data appears.



Eager Learners –

construct a classification model based on the given training data before getting data for predictions.

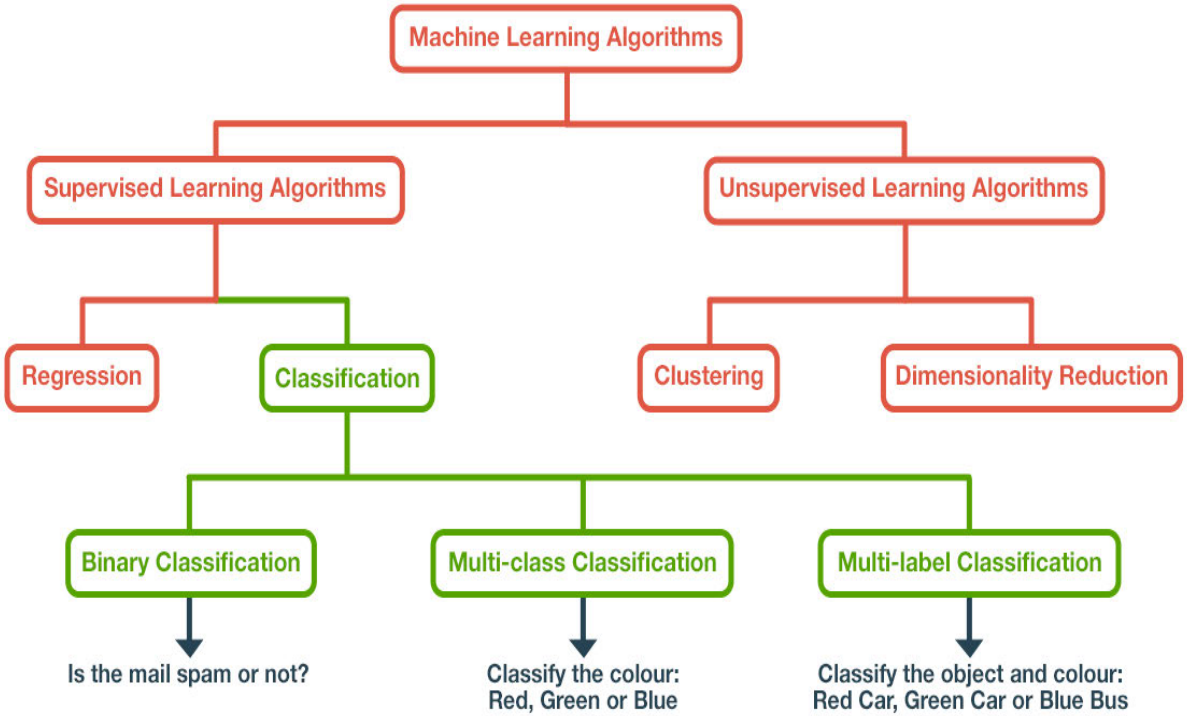
Classification

- This algorithm helps to **predict a discrete value** .
- It can be thought, the input data as a member of a particular class or group.
- For instance, taking up the photos of the fruit dataset, each photo has been labelled as a mango, an apple, etc. Here, the algorithm has to **classify** the new images into any of these categories.

Examples:

- a person's gender (male or female),
- the brand of product purchased (brand A, B, or C),
- whether a person defaults on a debt(yes or no),
- cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia).

Classification



Classification

Binary Classification

- Spam
- Not spam

Multiclass Classification

- Dog
- Cat
- Horse
- Fish
- Bird

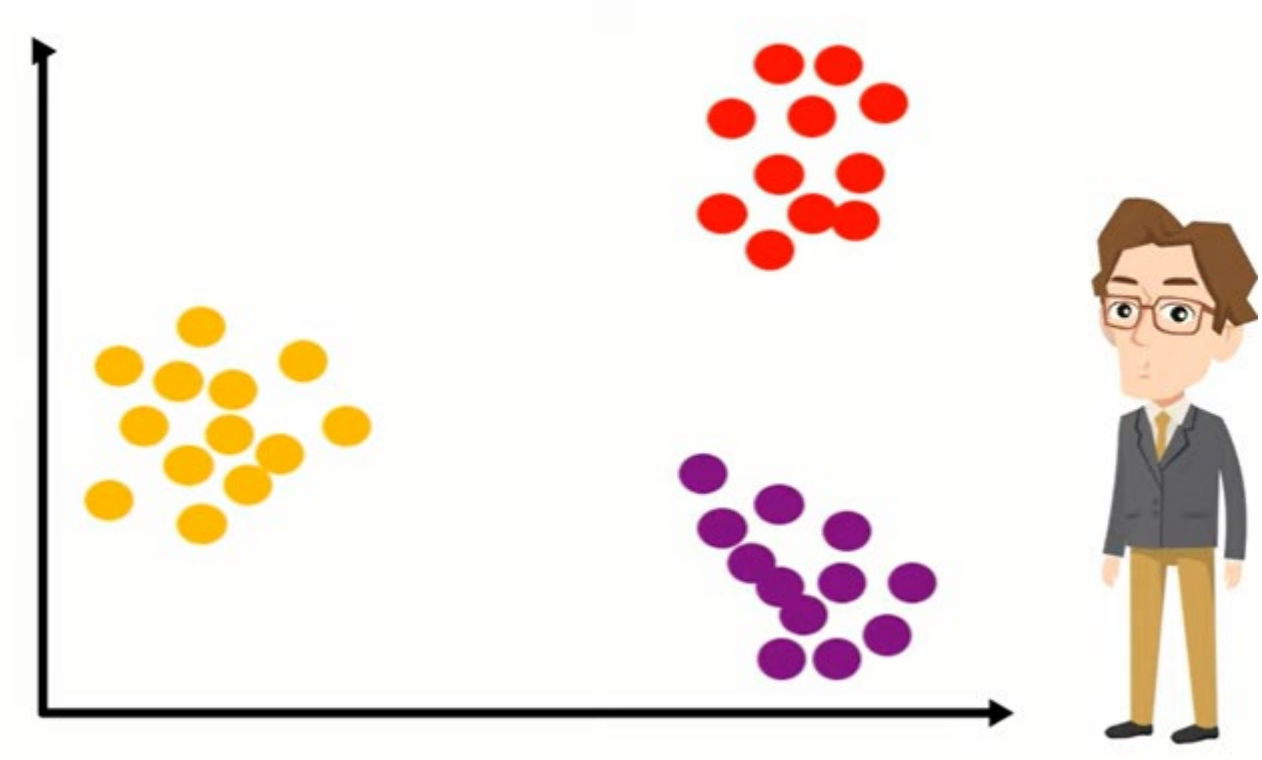
Multi-label Classification

- Dog
- Cat
- Horse
- Fish
- Bird

Classification Data

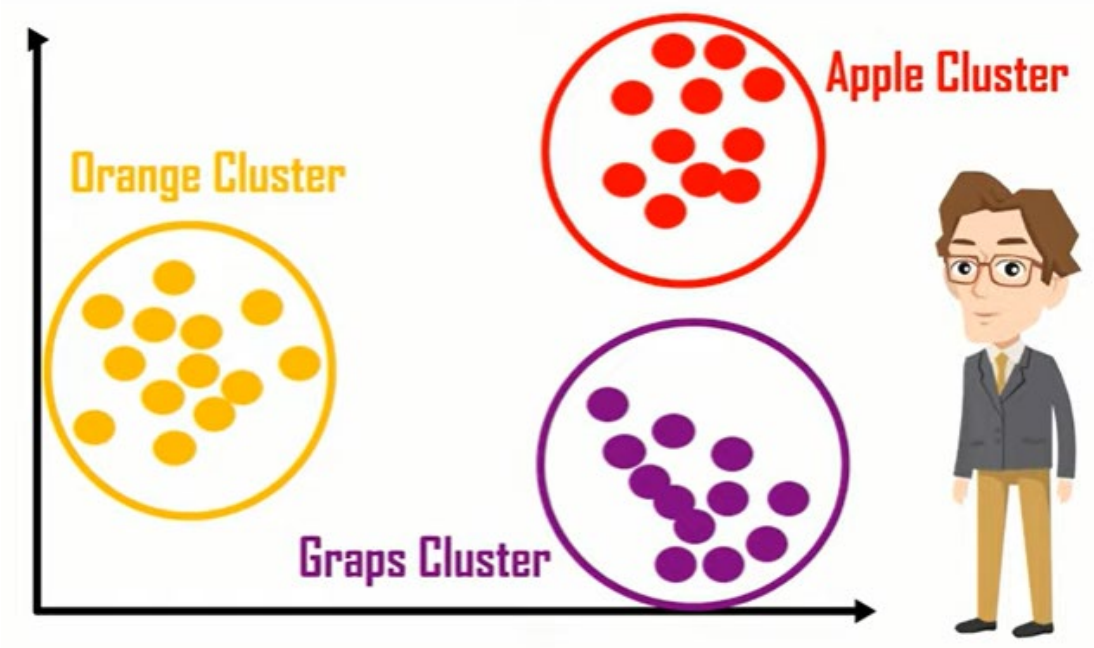
Diameter	Weight	Red	Green	Blue	Name
2.96	86.76	172	85	2	Orange
3.91	88.05	166	78	3	Orange
5.43	108.54	157	98	2	Apple
5.51	109.49	150	98	5	Grape
11.06	191.08	151	57	6	Apple
11.06	191.08	151	57	6	Grape
11.06	191.08	151	57	6	Grape
11.06	191.08	151	57	6	Orange
13.17	223.49	162	79	13	Grape
13.17	223.51	163	74	23	Apple
13.17	223.52	140	66	22	Orange
13.17	223.55	165	75	26	Orange
13.17	223.56	125	69	24	Apple

Classification



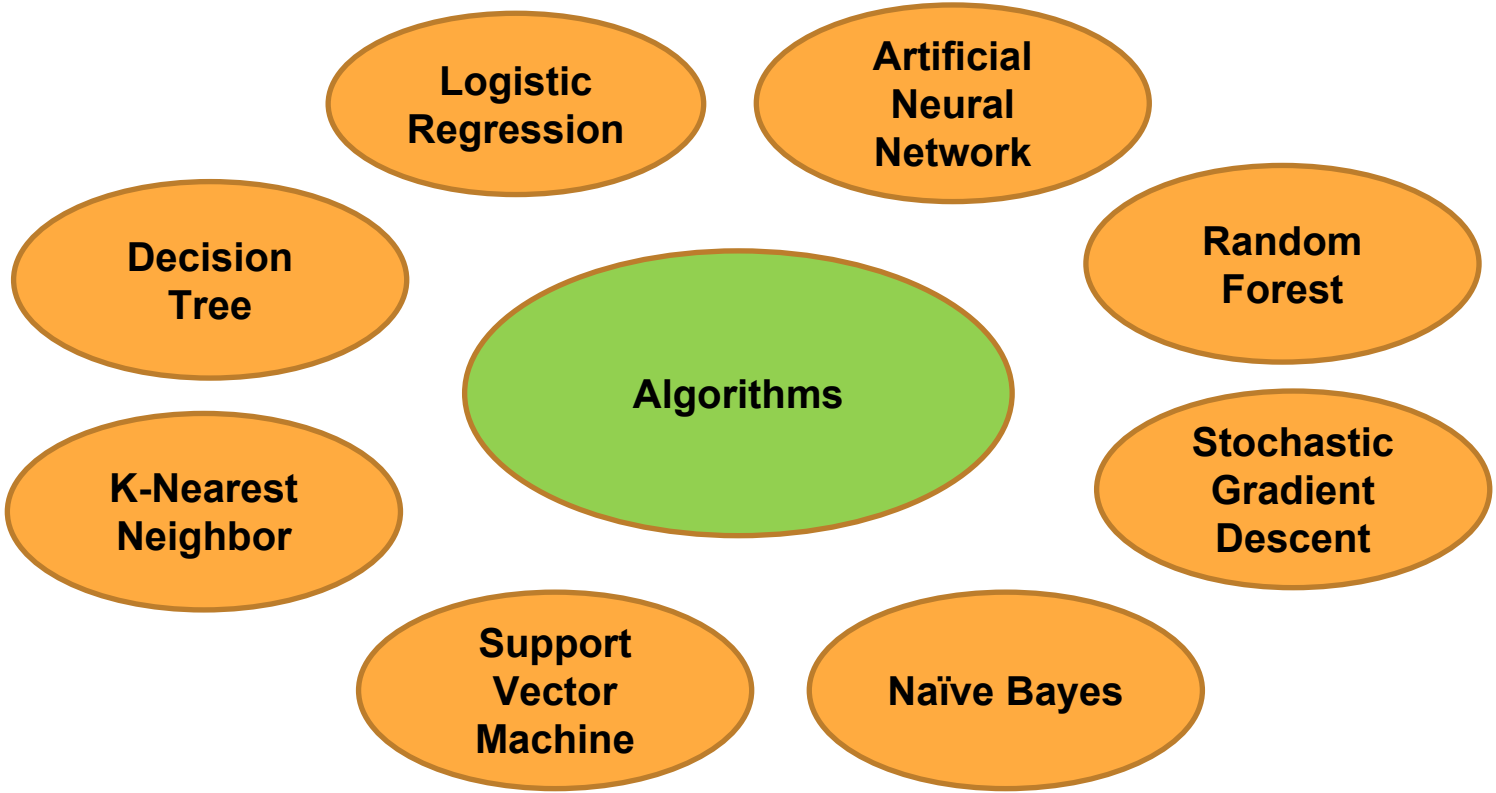
Courtesy: https://www.youtube.com/@tecnoport5886

Classification



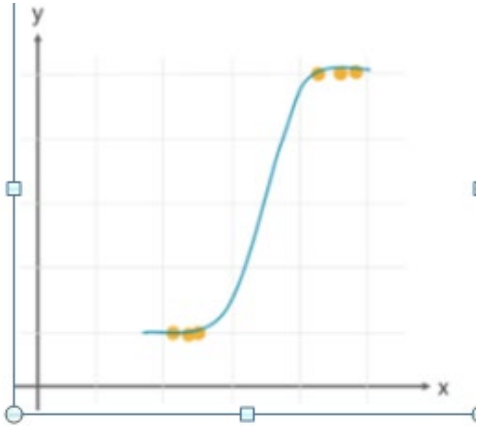
Courtesy: https://www.youtube.com/@tecnoport5886

Classification Algorithm



Logistic Regression

- It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with a dichotomous variable meaning **it will have only two possible outcomes**.



Use Cases

- Identifying risk factors for diseases
- Word classification
- Weather Prediction
- Voting Applications

Logistic Regression

Advantages and Disadvantages

- Logistic regression is specifically meant for classification, it is useful in understanding how a set of independent variables affect the outcome of the dependent variable.

- The main disadvantage of the logistic regression algorithm is that it only works when the predicted variable is binary, it assumes that the data is free of missing values and assumes that the predictors are independent of each other.

Naïve Bayes

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

It is a classification algorithm based on **Bayes's theorem** which gives an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Use Cases

- Disease Predictions
- Document Classification
- Spam Filters
- Sentiment Analysis

Naïve Bayes

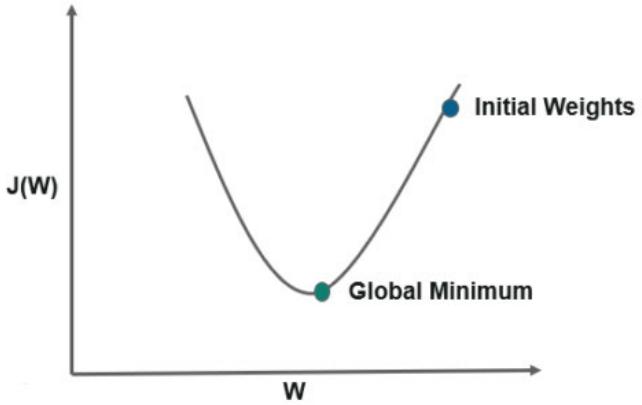
Advantages and Disadvantages

- The Naive Bayes classifier requires a small amount of training data to estimate the necessary parameters to get the results. They are extremely fast in nature compared to other classifiers.

- The only disadvantage is that they are known to be a bad estimator.

Stochastic Gradient Descent

- It is a very effective and simple approach to fit linear models. Stochastic Gradient Descent is particularly useful when the **sample data is in a large number**. It supports different loss functions and penalties for classification.



Use Cases

- Internet Of Things
- Updating the parameters such as weights in neural networks or coefficients in linear regression

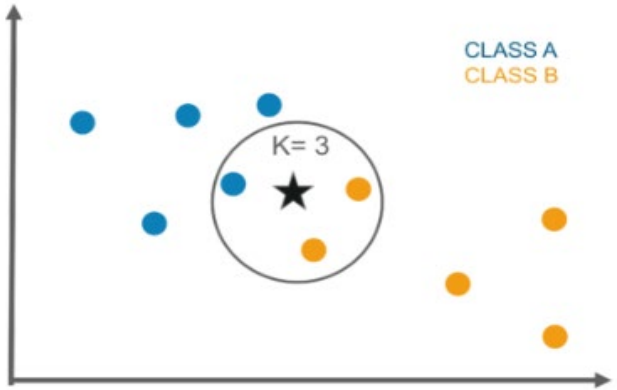
Stochastic Gradient Descent

Advantages and Disadvantages

- The only advantage is the ease of implementation and efficiency whereas a major setback with stochastic gradient descent is that it requires a number of hyper-parameters and is sensitive to feature scaling.

K-Nearest Neighbor

- It is a lazy learning algorithm that **stores all instances corresponding to training data in n-dimensional space**. It is a **lazy learning algorithm** as it does not focus on constructing a general internal model, instead, it works on storing instances of training data



Use Cases

- Industrial applications to look for similar tasks in comparison to others
- Handwriting detection applications
- Image recognition
- Video recognition
- Stock analysis

K-Nearest Neighbor

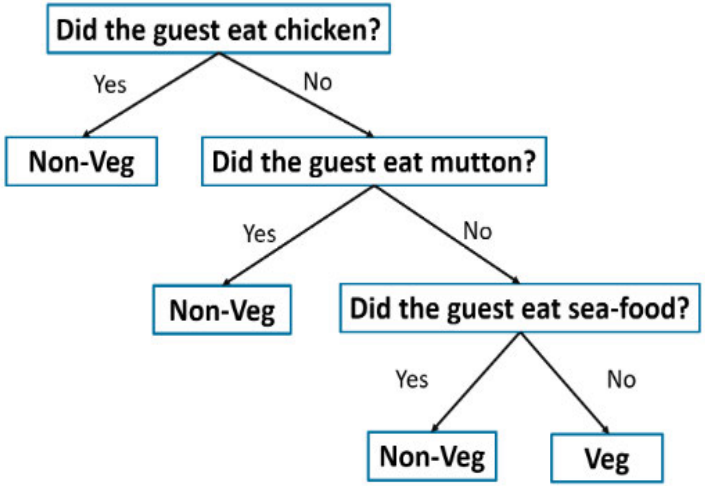
Advantages And Disadvantages

- This algorithm is quite simple in its implementation and is robust to noisy training data. Even if the training data is large, it is quite efficient.

- The only disadvantage with the KNN algorithm is that there is tricky determine the value of K and computation cost is pretty high compared to other algorithms.

Decision Tree

- The decision tree algorithm builds the classification model in the form of a **tree structure**. It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification.



Use Cases

- Data exploration
- Pattern Recognition
- Option pricing in finances
- Identifying disease and risk threats

Decision Tree

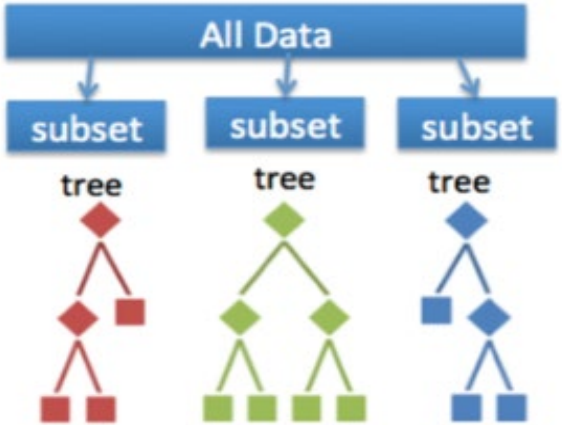
Advantages and Disadvantages

- A decision tree gives an advantage of simplicity to understand and visualize, it requires very little data preparation as well.

- The disadvantage that follows with the decision tree is that it can create complex trees that may not categorize efficiently. They can be quite unstable because even a simplistic change in the data can hinder the whole structure of the decision tree.

Random Forest

- Random decision trees or random forest are an **ensemble learning method** for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.



Use Cases

- Industrial applications such as finding if a loan applicant is high-risk or low-risk
- For Predicting the failure of mechanical parts in automobile engines
- Predicting social media share scores
- Performance scores

Random Forest

Advantages and Disadvantages

- The advantage of the random forest is that it is more accurate than the decision trees due to the reduction in the over-fitting.

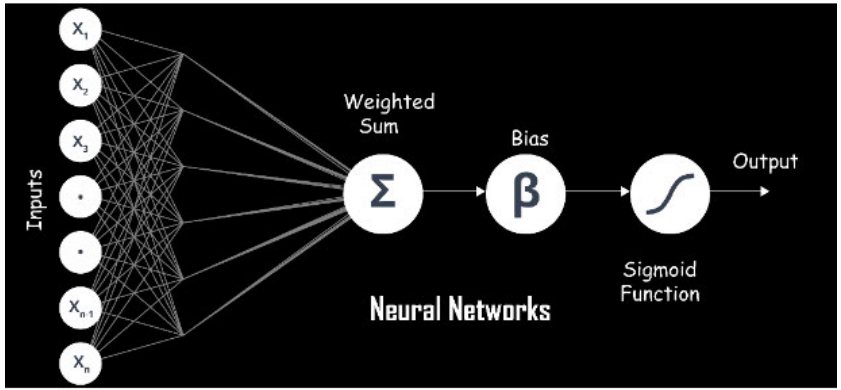
- The only disadvantage with the random forest classifiers is that it is quite complex in implementation and gets pretty slow in real-time prediction.

Artificial Neural Networks

innovate achieve lead

- A neural network consists of neurons that are **arranged in layers**, they take some input vector and convert it into an output. The process involves each neuron taking input and applying a function which is often a non-linear function to it and then passes the output to the next layer.

.



Use Cases

- Handwriting analysis
- Colorization of black and white images
- Computer vision processes
- Captioning photos based on facial features

Artificial Neural Networks

Advantages and Disadvantages

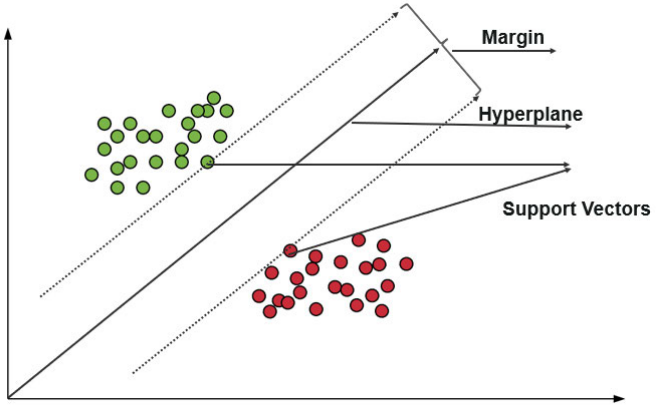
- It has a high tolerance to noisy data and able to classify untrained patterns, it performs better with continuous-valued inputs and outputs.

- The disadvantage with the artificial neural networks is that it has poor interpretation compared to other models.

.

Support Vector Machine

- The support vector machine is a classifier that represents the **training data as points in space** separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.



Use Cases

- Business applications for comparing the performance of a stock over a period of time
- Investment suggestions
- Classification of applications requiring accuracy and efficiency

Support Vector Machine

Advantages and Disadvantages

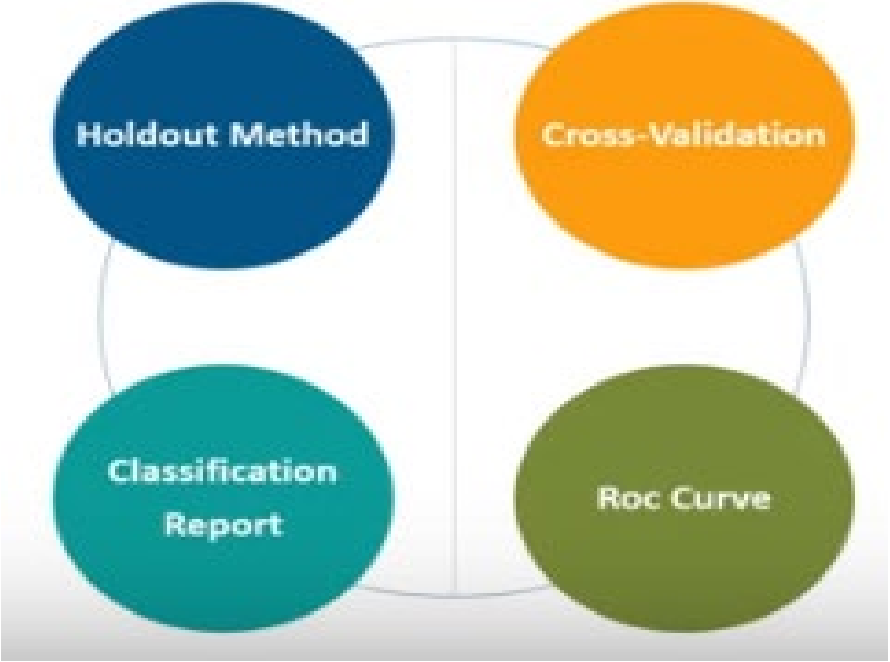
- It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces.

- The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

.

Classifier Evaluation

Classifier Evaluation

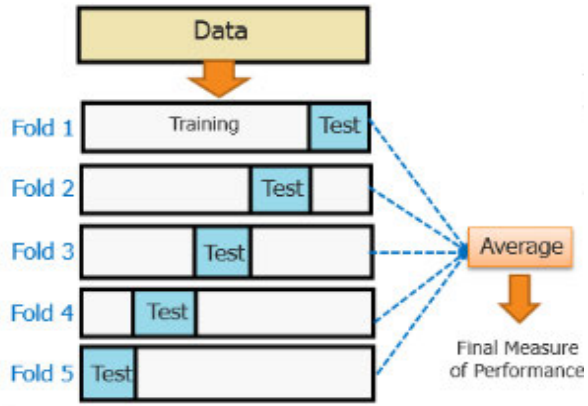


Holdout Method

- This is the most common method to evaluate a classifier. In this method, the given data set is divided into two parts as a test and train set 20% and 80% respectively.
- The train set is used to train the data and the unseen test set is used to test its predictive power

Cross-Validation

- Over-fitting is the most common problem prevalent in most of the machine learning models. K-fold cross-validation can be conducted to verify if the model is over-fitted at all.



Classification Report

Accuracy

Accuracy is a ratio of correctly predicted observation to the total observations

Precision And Recall

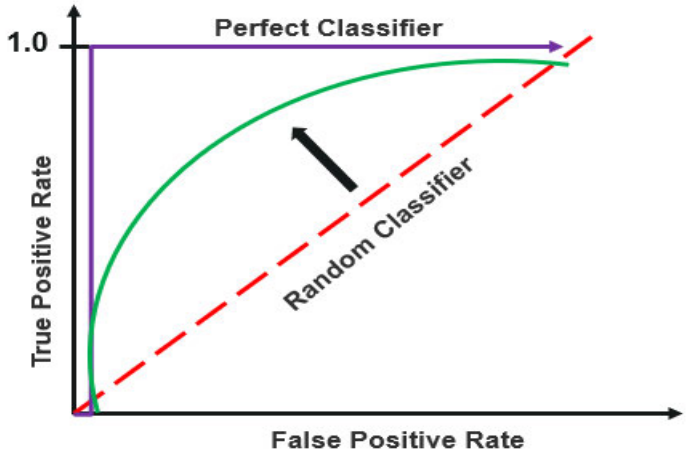
Precision is the **fraction of relevant instances among the retrieved instances**, while recall **is the fraction of relevant instances that have been retrieved over the total number of instances.**

F1- Score

It is the weighted average of precision and recall

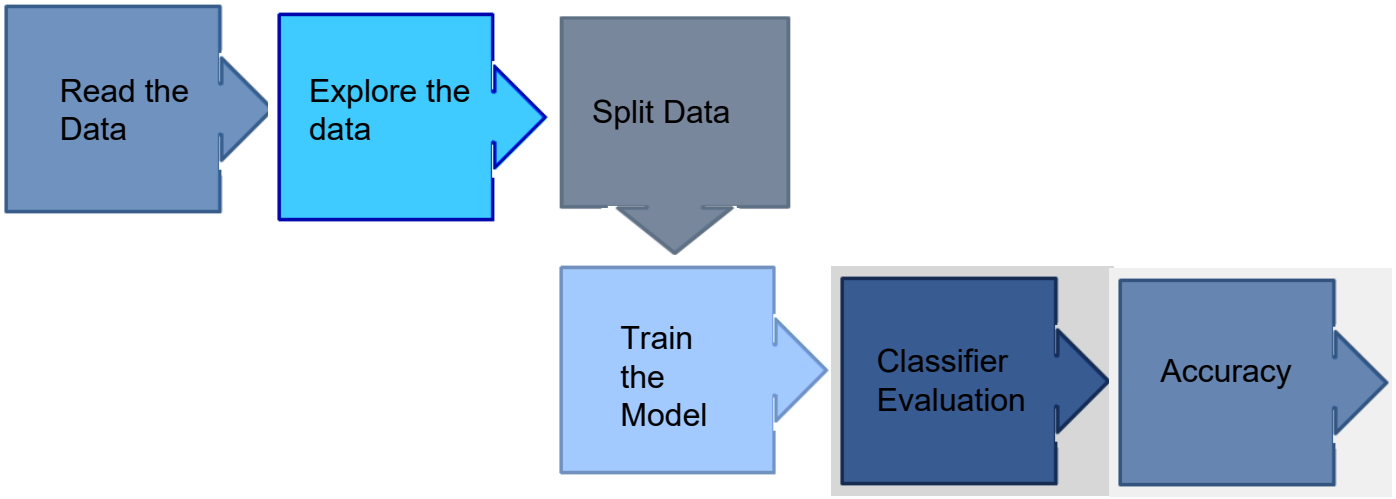
ROC Curve

- Receiver operating characteristics or ROC curve is used for visual comparison of classification models, which shows the relationship between the true positive rate and the false positive rate. The area under the ROC curve is the measure of the accuracy of the model.



Algorithm Selection

Algorithm Selection



Algorithm Selection

- Read the data
- Create dependent and independent data sets based on our dependent and independent features
- Split the data into training and testing sets
- Train the model using different algorithms such as KNN, Decision tree, SVM, etc
- Evaluate the classifier
- Choose the classifier with the most accuracy.

Standard Process for Data Science Project

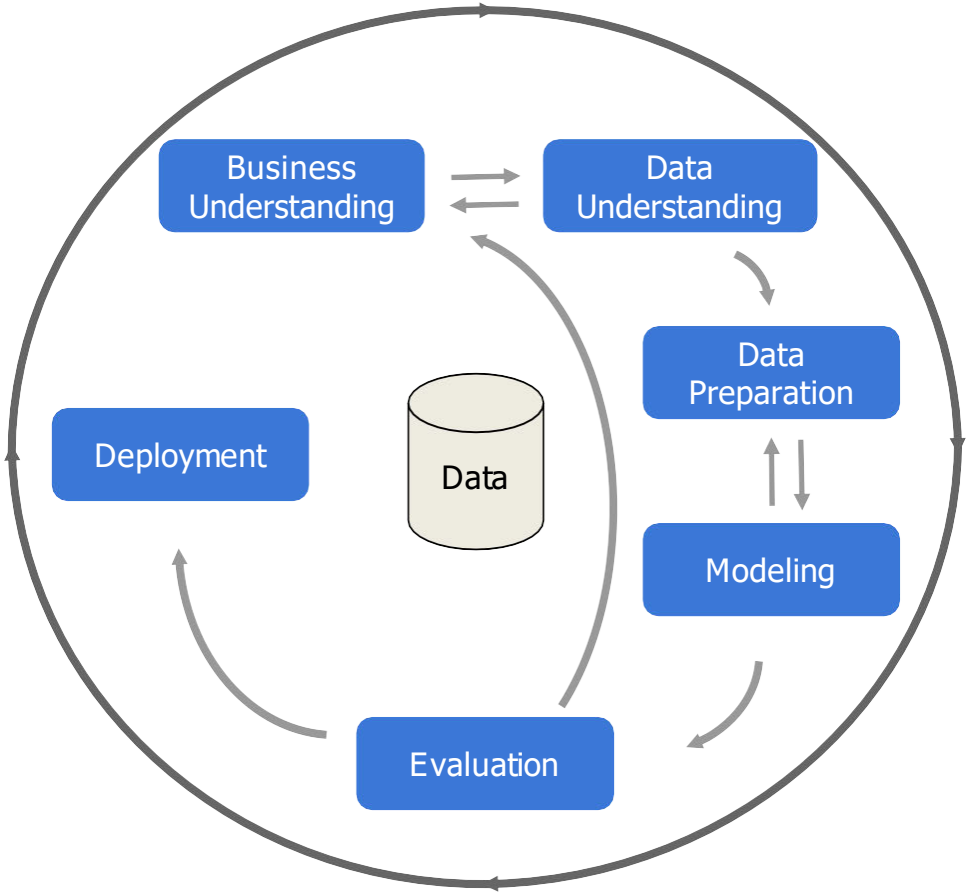
CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) is a standard process used for data mining

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

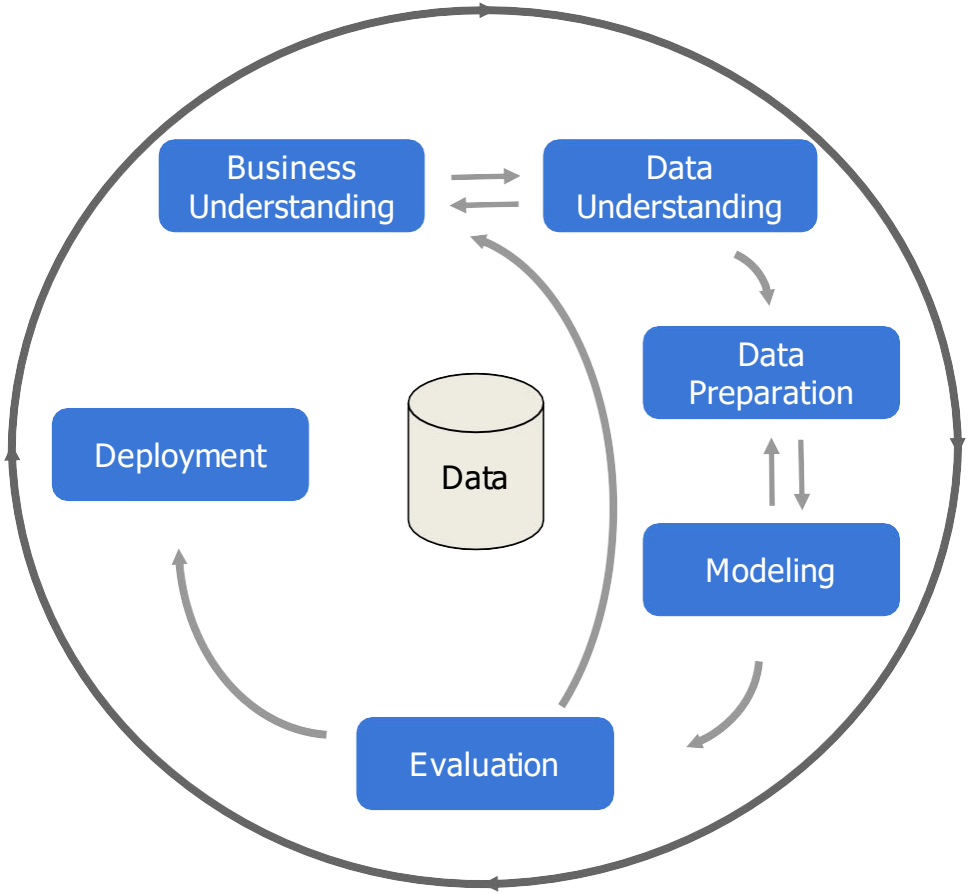
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- **Business Understanding**
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



Business understanding

In this phase we define what problem we are trying to solve.

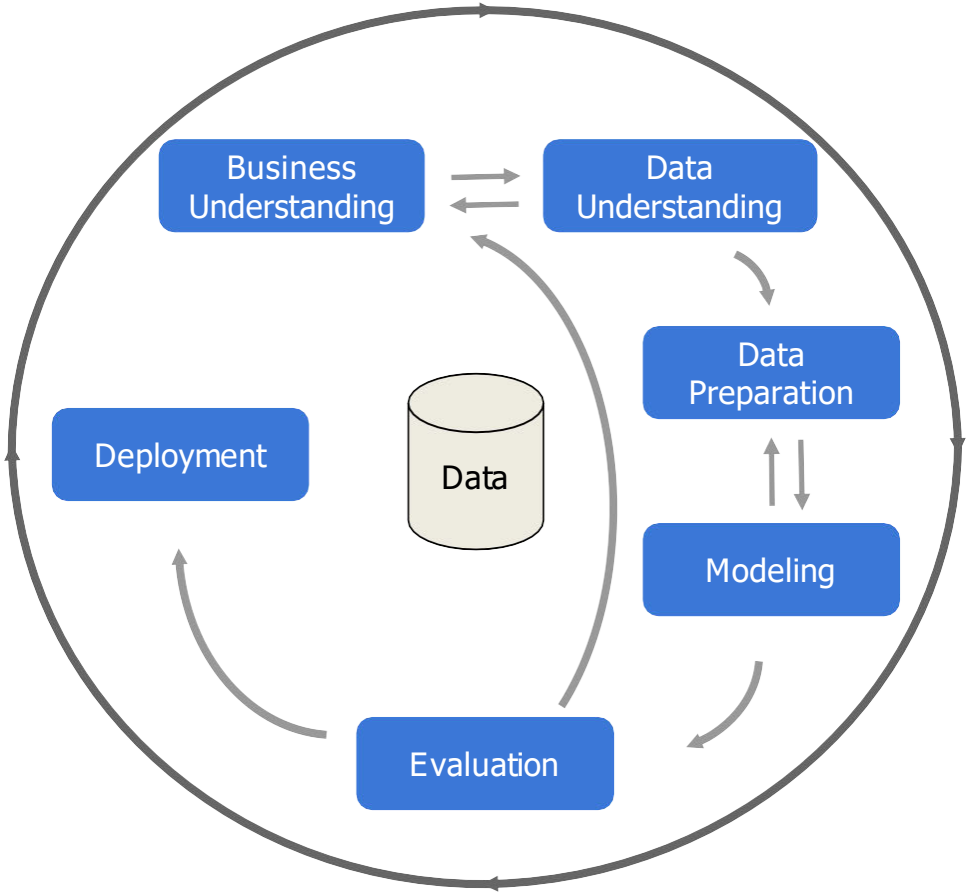
Example: Consider the example where we have inventory data for an online retailer which includes number of orders, type and demand area for each item. The aim is to classify the items based on if they have a high or low demand. So we can define clear business problems like:

- Is type of item related to the demand for the item?
- Can attributes in the considered data be used to classify the entire inventory list with reasonable accuracy?

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- **Data Understanding**
- Data Preparation
- Modeling
- Evaluation
- Deployment



Data understanding

This phase involves understanding the data considered for finding the solution.

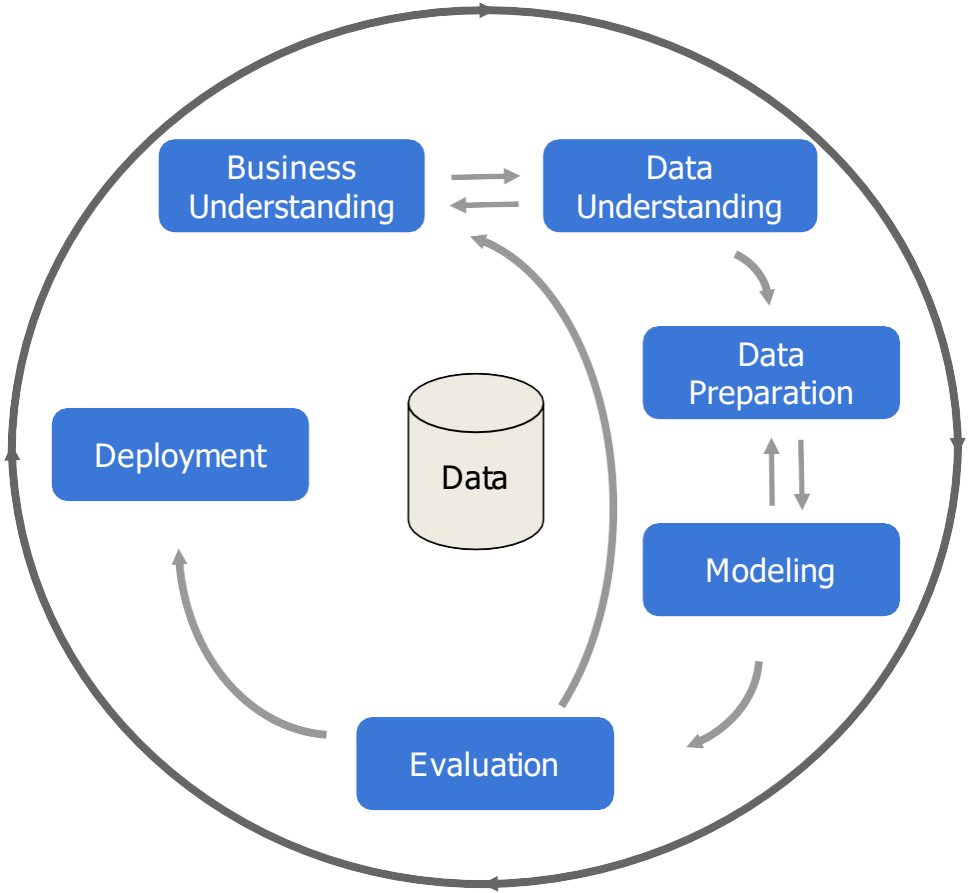
Example: Consider the example where we have inventory data for an online retailer which includes number of orders, type and demand area for each item. The aim is to classify the items based on if they have a high or low demand.

It is important to know if the items are perishable or non-perishable. For instance, items like dairy, cosmetics, and so on, can not be stocked, and hence adequate inventory should be available to meet the demand.

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- Data Understanding
- **Data Preparation**
- Modeling
- Evaluation
- Deployment



Data preparation

This phase involves cleaning and processing the data to be in a format suitable for the model used to solve the problem.

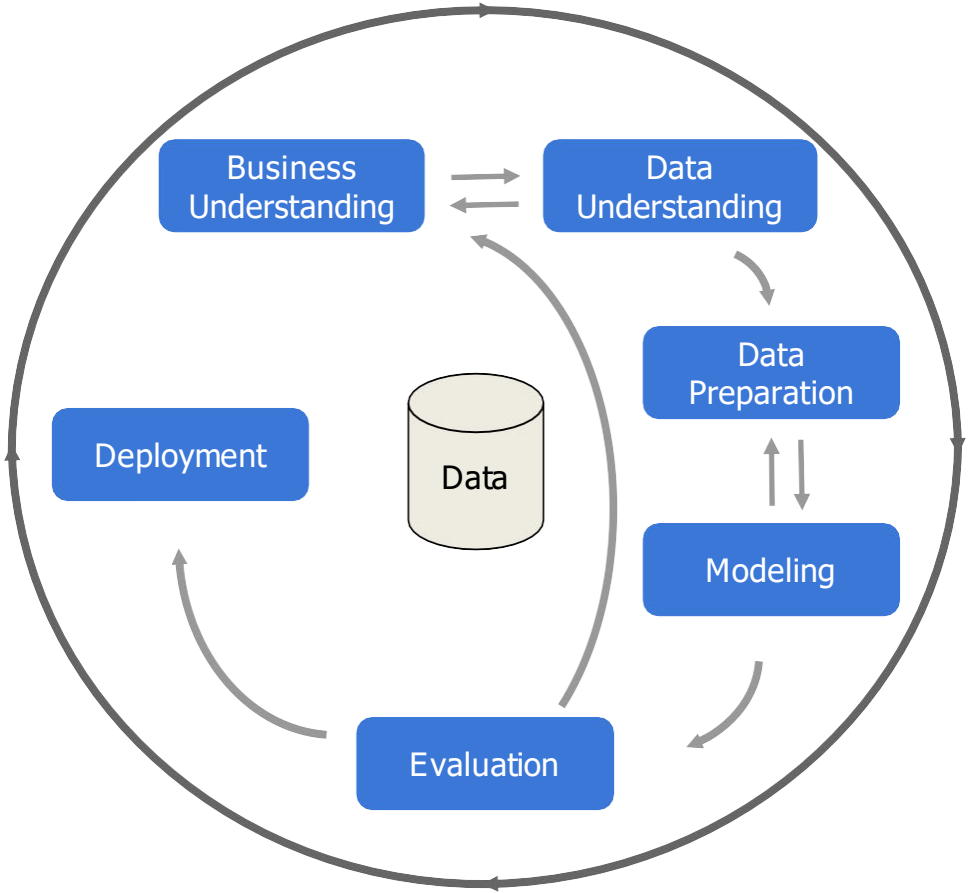
Example: Consider the example to classify the items based on if they have a high or low demand. We can prepare the data as follows:

- Treat the missing values
- The categorical variables need to be dummy encoded
- Check for correlation among variables

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- **Modeling**
- Evaluation
- Deployment



Modeling

- This phase involves finding the model that captures the solution to the business problem using available data

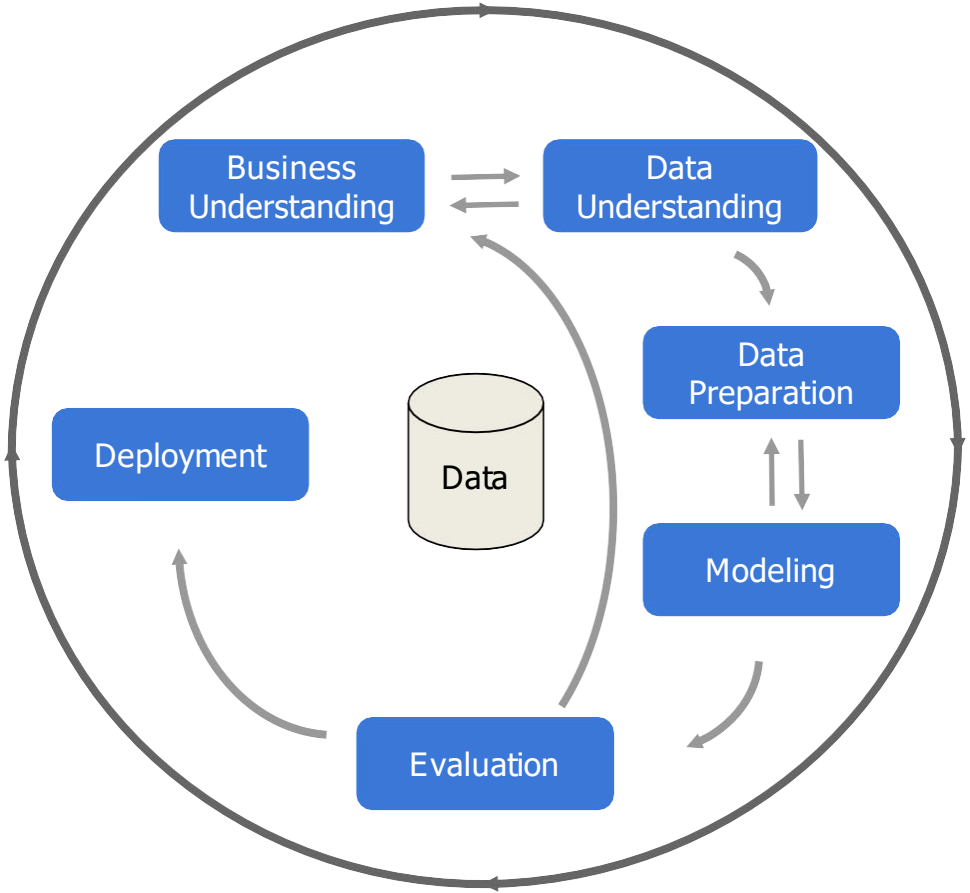
- We may have to try multiple models and go back and forth between data preparation and modelling to choose the correct model

Example: In the modelling phase we try to find a function that maps the attributes like number of orders, type, etc from the data to demand for the item.

CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- **Evaluation**
- Deployment



Evaluation

Once the model is built we need to check how good the model performs on unseen data. This process is done during the evaluation phase.

Example: We can check the model performance on data for which we know the actual demand. Using that data we can compare the predicted and actual values and evaluate.

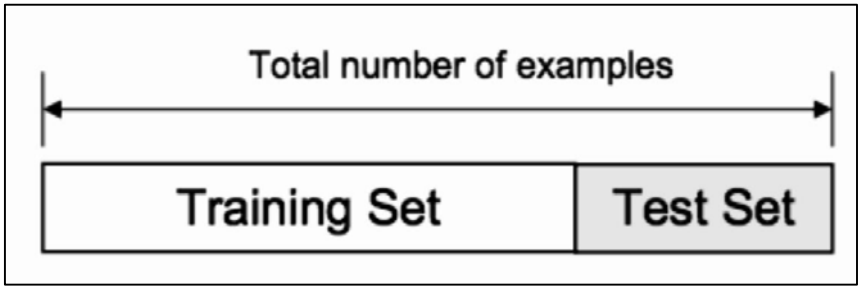
Train-Test Split

Train-test split

- The most straightforward technique that is used to evaluate the performance of a machine learning algorithm is to use different subsets of a dataset
- We can split our original dataset into two parts (train and test)
- Build the model on the training dataset, make predictions on the test dataset and evaluate the forecasts against the expected results

Train-test split

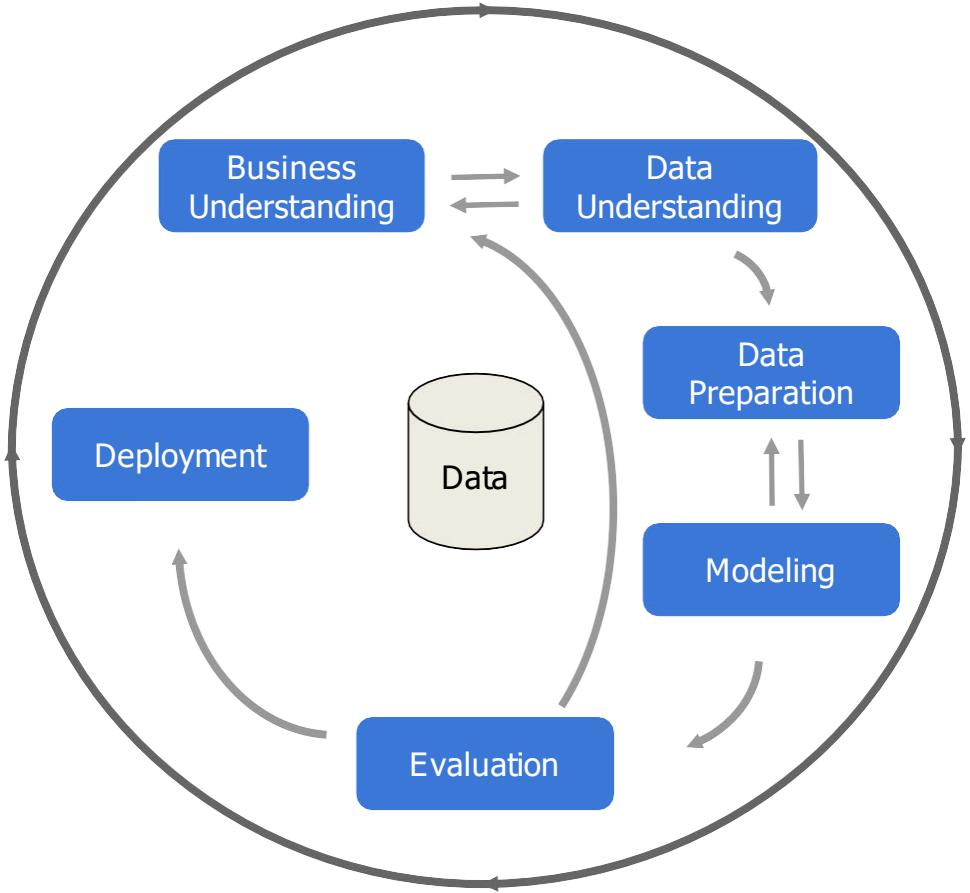
The size of the split can depend on the size of the dataset, although it is common to use 70% of the data for training and the remaining 30% for testing.



CRISP-DM phases

CRISP-DM breaks data mining into six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- **Deployment**



Deployment

If we are satisfied with the performance of the model from the previous phase we deploy it in the deployment phase

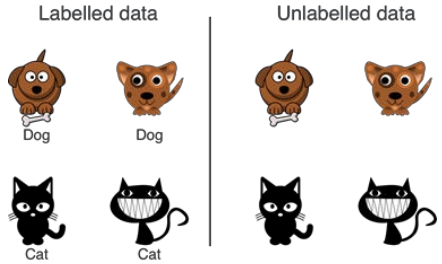
Example: For the considered example of predicting demand for an item, perhaps we could develop an app that takes input as the attribute values for an item and returns the demand for that item to the retailer.

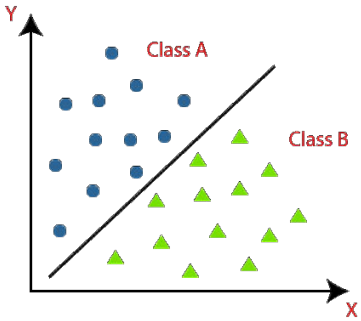
- Logistic Regression
- - intuition

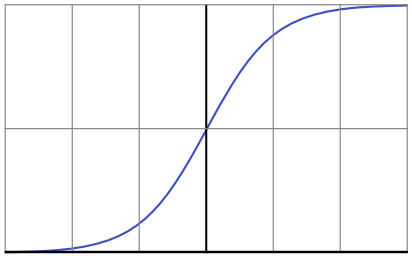
Logistic Regression

About Logistic Regression:

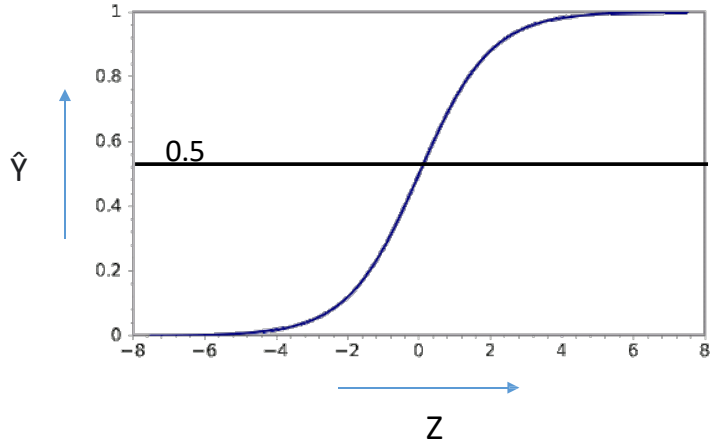
1. Supervised Learning Model
2. Classification model
3. Best for Binary Classification Problem
4. Uses Sigmoid function







Logistic Regression



$$\hat{Y} = \frac{1}{1 + e^{-Z}}$$

$$Z = w.X + b$$

Sigmoid Function

- Ŷ - Probability that (y = 1)
- Ŷ = P(Y=1 | X)

  Ŷ = σ (Z)
- X - input features
- w – weights

(number of weights is equal to the number of input features in a dataset)

- b - bias

Logistic Regression

Advantages:

1. Easy to implement
2. Performs well on data with linear relationship
3. Less prone to over-fitting for low dimensional dataset

Disadvantages:

1. High dimensional dataset causes over-fitting
2. Difficult to capture complex relationships in a dataset
3. Sensitive to Outliers
4. Needs a larger dataset

- Math behind
- Logistic Regression

Visiting Basics

Odds vs probability

Odds of an event are the ratio of number of observations in favour of an event to number of observations not in favour of the event

$$\text{odds} = \frac{\text{number of observations in favour of the event}}{\text{number of observations not in favour of the event}}$$

Probability of an event is the ratio of number of observations in favour of an event to all possible observations

$$\text{probability} = \frac{\text{number of observations in favour of the event}}{\text{number of observations}}$$

Odds vs probability

Plasma score	90	90	150	165	115	180	100	170	130	166
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes

For the above data, the odds of a patient having diabetes is given by,

$$\text{odds} = \frac{\text{number of patients having diabetes}}{\text{number of patients not having diabetes}} = \frac{6}{4}$$

For the above data,, the probability of a patient having diabetes is given by,

$$\text{probability} = \frac{\text{number of patients having diabetes}}{\text{Total number of patients}} = \frac{6}{10}$$

Log of odds

odds of having diabetes $= \frac{6}{4}$	odds of not having diabetes $= \frac{4}{6}$
log(odds of having diabetes) = ln (1.5) = 0.405	log(odds of not having diabetes) $= \ln(0.667) = -0.405$

- As we can see if we only consider odds value, the magnitude for each class value taken by variable is very different
- Hence the log(odds) value is consider so that no matter which class the magnitude is same
- Log of odds is the logit function used in logistic regression

Relation between odds and probability

If P(A) is probability of event A

$$\text{Odds} = \frac{\text{P(A)}}{\text{1-P(A)}}$$

$$\text{Probability} = \frac{\text{odds}}{\text{1+odds}}$$

$$\log(\text{Odds}) = \ln\left(\frac{\text{P(A)}}{\text{1-P(A)}}\right)$$

Odds ratio

- Odds ratio refers to the ratio of odds
- Odds ratio can be used to determine the impact of a feature on target variable
- For our considered example the odds ratio can be calculated as,

$$\text{odds ratio} = \frac{\text{odds of patient having diabetes}}{\text{odds of patient not having diabetes}} = \frac{\frac{6}{4}}{\frac{4}{6}} = \frac{9}{4}$$

Question:

Patients with high sugar diet are considered more susceptible to diabetes. How can we determine whether sugar content in diet has an impact on possibility of a patient getting diagnosed with diabetes? Consider the following sample data.

Sugar content in diet	High	High	Low	High	Low	High	High	Low	High	Low
Is the patient Diabetic?	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No

Solution:

From the given sample data we can calculate:

1. Odds of a patient having diabetes given he has high sugar diet

$$\frac{\text{number of patients having diabetes given he has high sugar diet}}{\text{number of patients not having diabetes given he has high sugar diet}} = \frac{4}{2}$$

2. Odds of a patient having diabetes given he has low sugar diet

$$\frac{\text{number of patients having diabetes given he has low sugar diet}}{\text{number of patients not having diabetes given he has low sugar diet}} = \frac{1}{3}$$

Solution continued:

From 1 and 2 we can calculate odds ratio:

$$\text{odds ratio} = \frac{\text{odds of a patient having diabetes given he has high sugar diet}}{\text{odds of a patient having diabetes given he has low sugar diet}} = \frac{\frac{4}{2}}{\frac{1}{3}} = 6$$

Thus from the odds ratio we can see that patients with a high sugar diet are 6 times more susceptible to diabetes compared to patients who have a low sugar diet.

Odds

- Odds ➔ Ratio of the chances of success to the chances of failure.
- As a result, in Logistic Regression, **a linear combination of inputs is translated to log(odds), with an output of 1.**

Probability and Odds

- The probability that an event will occur is the **fraction of times you expect to see that event in many trials.** If the probability of an event occurring is Y, then the probability of the event not occurring is 1-Y. Probabilities always range between 0 and 1.
- The odds are defined as the **probability that the event will occur divided by the probability that the event will not occur.** Unlike probability, the odds are not constrained to lie between 0 and 1, but can take any value from zero to infinity.

Binomial Logistic Regression

Question:

Consider the example below about whether or not a patient has diabetes based on plasma score. Can we use linear regression line to predict the whether the patient is diabetic?

Plasma score	90	90	150	165	115	180	100	170	130	166
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes

Answer:

The example about whether or not a patient has diabetes based on plasma score is a classification problem. Moreover, the target variable is categorical. Hence, we can not use linear regression line to predict the whether the patient is diabetic. We classify them as diabetic and non-diabetic.

Plasma score	90	90	150	165	115	180	100	170	130	166
Is the patient Diabetic?	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes

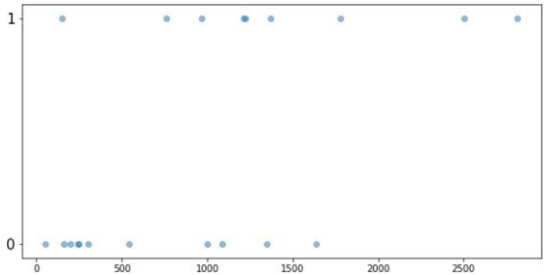
Logistic regression

- Here on we shall consider the adjacent data
- The data is tell us the presence of one fish depending on the density of other fish in a lake
- If they compete with each other, then higher density of one may suggest absence of the other whereas if they are symbiotic, high density of one may promote the other

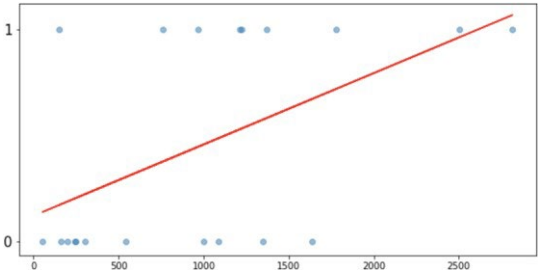
BKT kg/ha	Presence of fish
1085.33	0
1210	1
1780.62	1
52.4	0
200	0
2502.67	1
301.33	0
542	0
969.33	1
240.56	0
1640	0
247	0
999.99	0
1220.76	1
150.67	1
160	0
2816	1
760	1
1350	0
1370	1

Logistic regression

- Consider the scatter plot of the previous data



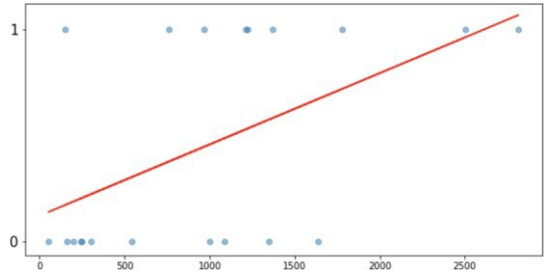
- Fit a linear regression line to it



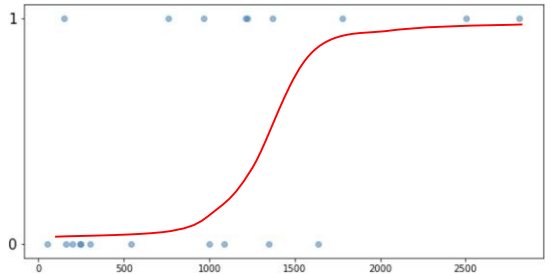
- Note the line is not a true representative of the data

Logistic regression

- A S-shaped curve as in the figure below gives the true relationship



- Such a curve is given be the sigmoid function

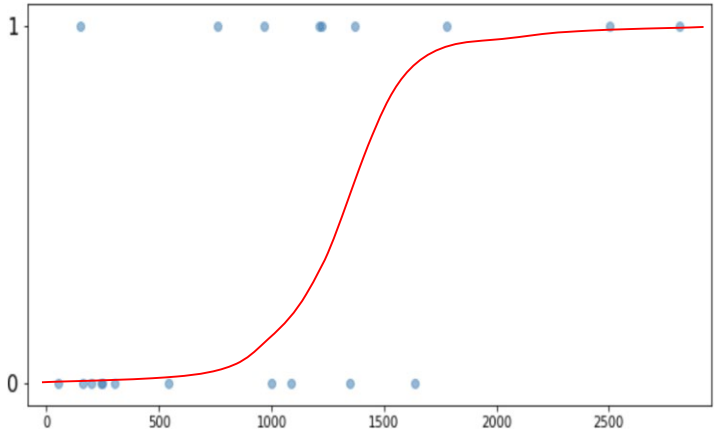


What is a sigmoid function?

- The sigmoid function is a mathematical function which is S-shaped and is given by

$$f(x) = \frac{1}{1-\exp^{-z}}$$

- It exists between 0 to 1



Logistic regression

- Logistic regression is a binary classification algorithm

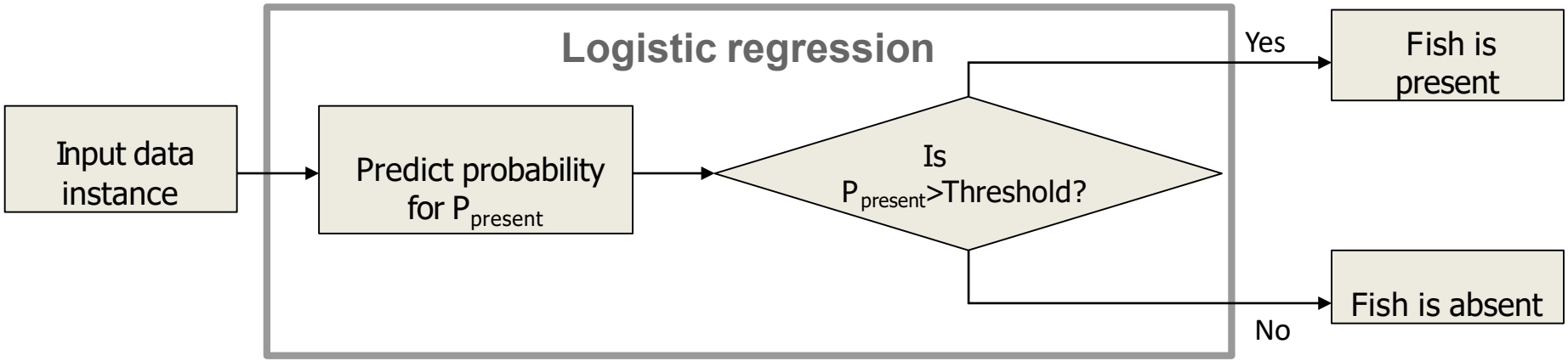
- It predicts the probability of occurrence of a class label. Based on these probabilities
- the data points are labelled

- A threshold (or cut-off; commonly a threshold of 0.5 is used) is fixed, then

	Classify as
threshold < probability	Presence of fish
Probability < threshold	Absence of fish

Main steps in logistic regression

Consider that logistic regression is used to identify whether or not a patient is suffering from diabetes



Logistic regression

- The logistics regression is given by

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

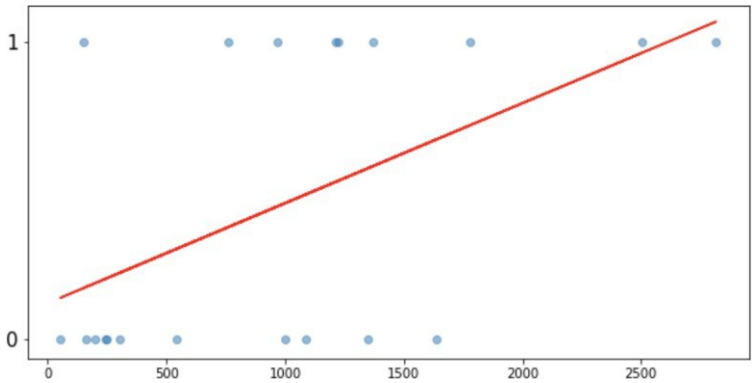
- Here, π(x) is the conditional expectation of the outcome given the values for independent variables , i.e. E(Y|X)

- It predicts the probability of occurrence of a class label by fitting the data to a function called logit function, hence called logit regression

Probability as output of logistic regression

- The logistic regression model is given by $\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$
- Taking limits tending to -∞ on both side, $\lim_{x \to -\infty} \frac{e^x}{1+e^x} = 0$
- Taking limits tending to ∞ on both side, $\lim_{x \to \infty} \frac{e^x}{1+e^x} = 1$
- Thus π(x) lies in-between 0 and 1, i.e. π(x) ∈ [0,1] and can be viewed as probability

Logistic regression



- Since we are predicting probabilities it is important for values to be between 0 and 1
- Consider the points A and B for which values for plasma score are 80 and 177 respectively, the probability values are out of the expected range 0 to 1
- Hence linear regression cannot be directly used to predict probabilities

Usage

- Classification:

  The ICU in a hospital is assigned on priority to high risk patients. Logistic regression can be used to classify the list of patients into high risk and low risk records.

- Profiling:

  Nuclear fuel companies moderate various factors like pressure, temperature, etc. to produce high yielding and low yielding fuels. Based on different parameters for the current day it can be predetermined whether the fuel produced will be high yielding or not. The company can then alter the parameters to produce high yielding fuel everyday.

Logistic regression

- Thus for a binary logistic classification where the target variable takes two values names 0 and 1, we have

Class labels	0	1
P[Y=y | X]	1-π(x)	π(x)

- π(x) denotes the probability that the response is present for the records for some combination of values that the independent variables take, i.e. for X=x

- 1-π(x) denotes the probability that the response is absent for the records for some combination of values that the independent variables take, i.e. for X=x

Linearization

- To estimate the parameter we need to linearize the function. We use the logit transformation

$$\eta = \ln \frac{\pi}{1-\pi}$$

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

- The ratio π/(1-π) is called the odds

- Hence the logit transformation is also known as the log-odds

Linearization

We have,

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\frac{\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1+\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}}{1-\frac{\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1+\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}} = \frac{\frac{\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{\cancel{1+\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}}}{\frac{1}{\cancel{1+\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}}} = \frac{\exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1}$$

That is,

$$\frac{\pi(x)}{1-\pi(x)} = \exp^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}$$

Linearization

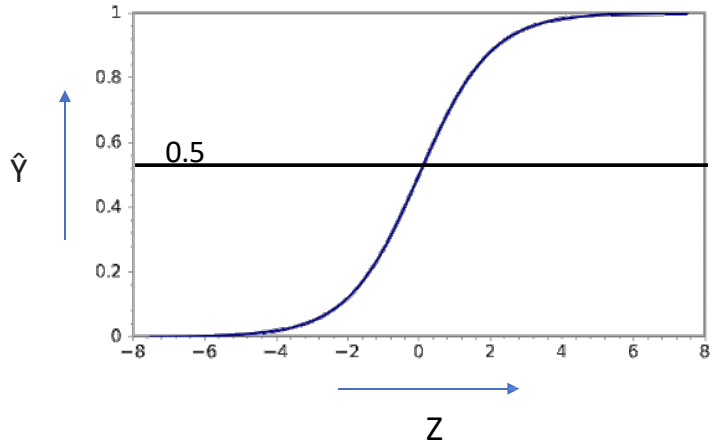
Taking natural log on both sides we have,

$$\ln \frac{\pi(x)}{1-\pi(x)} = \ln \exp^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}$$

$$\ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Thus, we have a linear relationship.

Logistic Regression



$$\hat{Y} = \frac{1}{1 + e^{-Z}}$$

$$Z = w.X + b$$

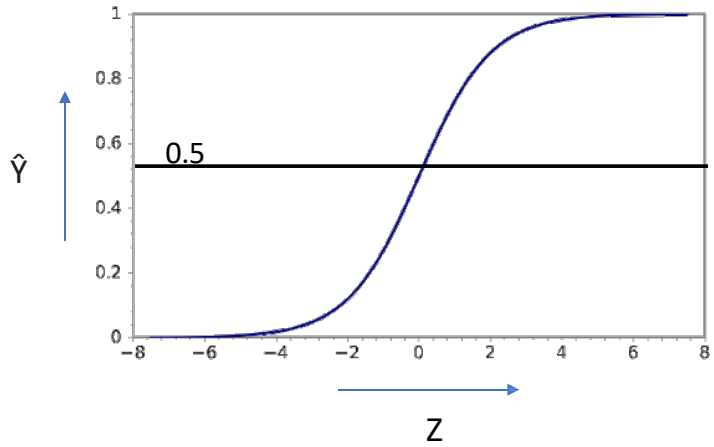
Sigmoid Function

- $\hat{Y}$ - Probability that (y = 1)
- $\hat{Y} = P(Y=1 \mid X)$
- X - input features
- w – weights
- (number of weights is equal to the number of input features in a dataset)

$\hat{Y} = \sigma(Z)$

- b - bias

Logistic Regression



$$\hat{Y} = \frac{1}{1 + e^{-Z}}$$

Sigmoid Function

Z = 5X + 10

Ŷ - Probability that (y = 1)

Ŷ = P(Y=1 | X)

X - input features

w – weights
(number of weights is equal to the number of input features in a dataset)

b - bias

Ŷ = σ (Z)

Logistic Regression

X	-9	-8	0	8	9
$\hat{Y}$					

$$Z = 5X + 10$$

$$\hat{Y} = \frac{1}{1+e^{-Z}}$$

X = -9

Z = 5(-9) + 10

Z = -35

$$\hat{Y} = \frac{1}{1+e^{35}}$$

$\hat{Y}$ = 0

X = -8

Z = 5(-8) + 10

Z = -30

$$\hat{Y} = \frac{1}{1+e^{30}}$$

$\hat{Y}$ = 0

X = 0

Z = 5(0) + 10

Z = 10

$$\hat{Y} = \frac{1}{1+e^{-10}}$$

$\hat{Y}$ = 1

X = 8

Z = 5(8) + 10

Z = 50

$$\hat{Y} = \frac{1}{1+e^{-50}}$$

$\hat{Y}$ = 1

X = 9

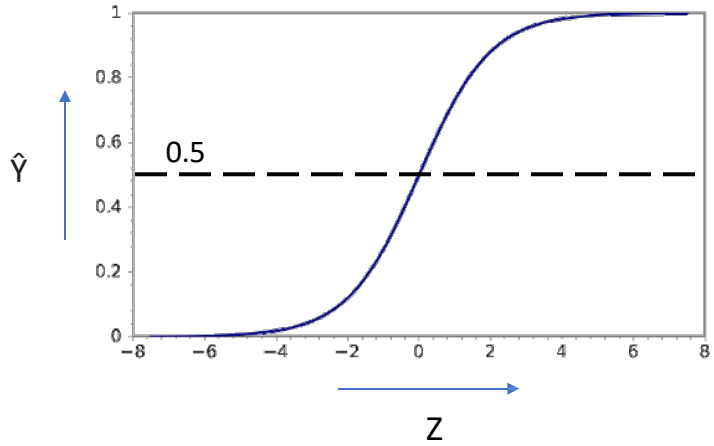
Z = 5(9) + 10

Z = 55

$$\hat{Y} = \frac{1}{1+e^{-55}}$$

$\hat{Y}$ = 1

Logistic Regression



$$\hat{Y} = \frac{1}{1 + e^{-Z}}$$

$$Z = w.X + b$$

Sigmoid Function

Inference:

If Z value is a large positive number,

$$\hat{Y} = \frac{1}{1 + 0}$$

$$\hat{Y} = 1$$

If Z value is a large negative number,

$$\hat{Y} = \frac{1}{1 + (large\ positive\ number)}$$

$$\hat{Y} = 0$$

Logit Function

$$\frac{P}{1-P} = e^{b_0 + b_1 x}$$

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

$$P = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

In the logistic regression the constant (b_0) moves the curve left and right and the slope (b_1) defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio.

Finally, taking the natural log of both sides, we can write the equation in terms of **log-odds (logit)** which is a linear function of the predictors. The coefficient (b_1) is the amount the logit (log-odds) changes with a one unit change in x.

logistic regression can handle any number of numerical and/or categorical variables.

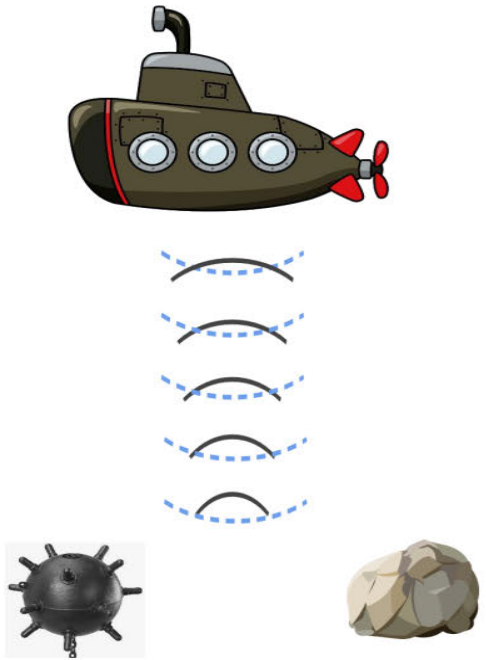
Linear Regression Vs Logistic Regression

	Linear Regression	Logistic Regression
1 Definition	To predict a continuous dependent variable based on values of independent variables	*To predict a categorical dependent variable based on values of independent variables*
2 Variable Type	Continuous dependent variable	Categorical dependent variable
3 Estimation method	Least square estimation	Maximum like-hood estimation
4 Equation	$Y = b_0 + b_1 x + e$	$\log\left(\frac{Y}{1-Y}\right) = C + B1X1 + B2X2 + \ldots.$
5 Best fit line	Straight line	Curve
6 Relationship between DV & IV	Linear relationship between the dependent and independent variable	Linear relationship is not mandatory
7 Output	Predicted integer value	Predicted binary value (0 or 1)
8 Applications	Business domain, forecasting sales	Classification problems, cybersecurity, image processing

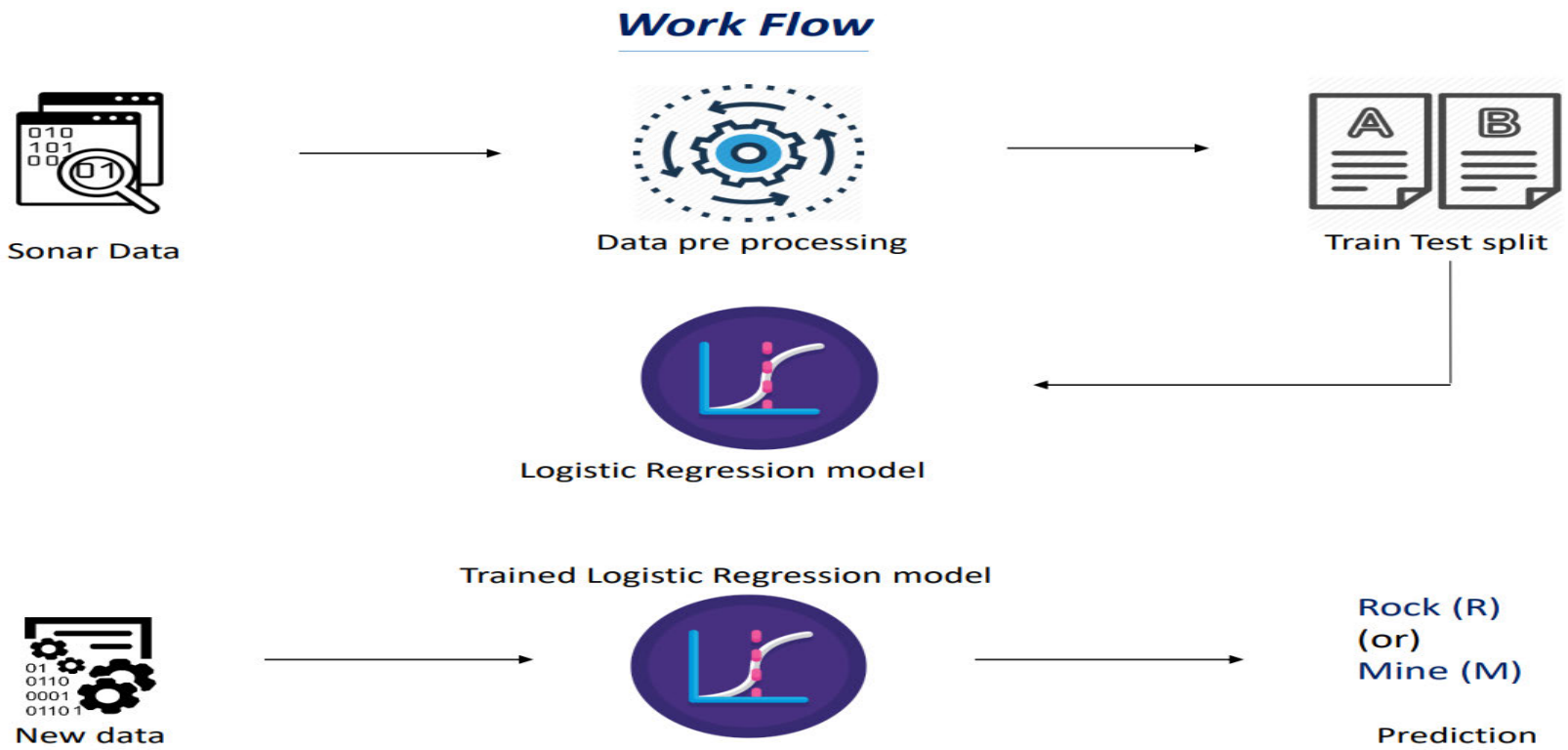
Implementation of Logistic Regression

Submarine Rock vs Mine Prediction

Machine Learning Project



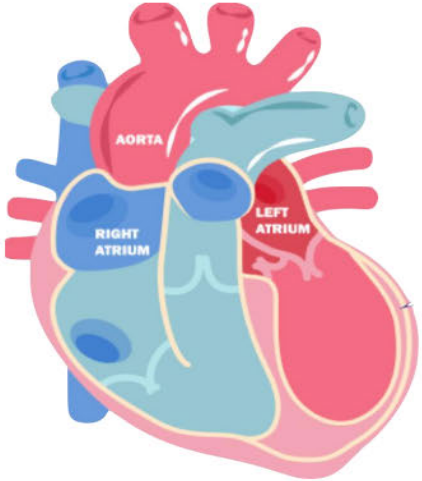
Implementation of Logistic Regression



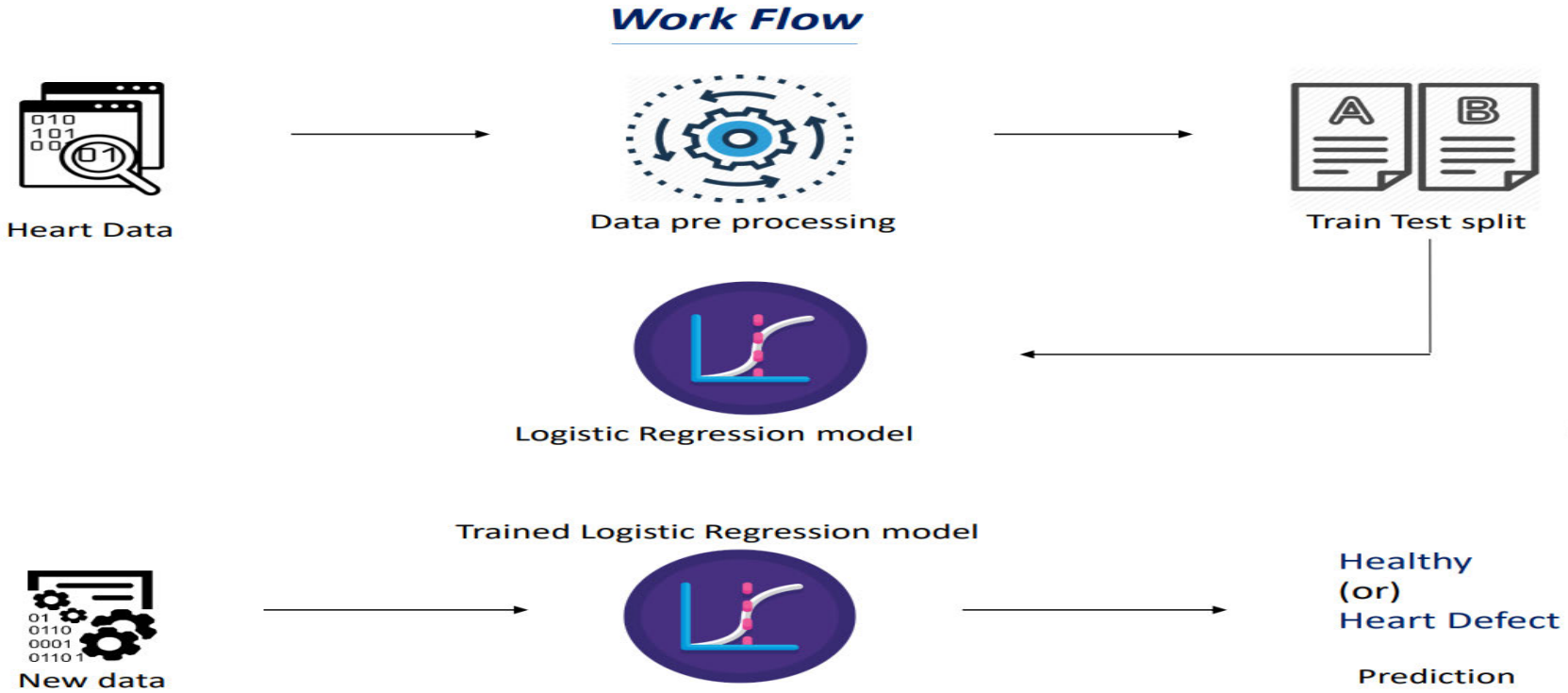
Implementation of Logistic Regression

Heart Disease Prediction With Python

Machine Learning Project



Implementation of Logistic Regression



THANK YOU !!!