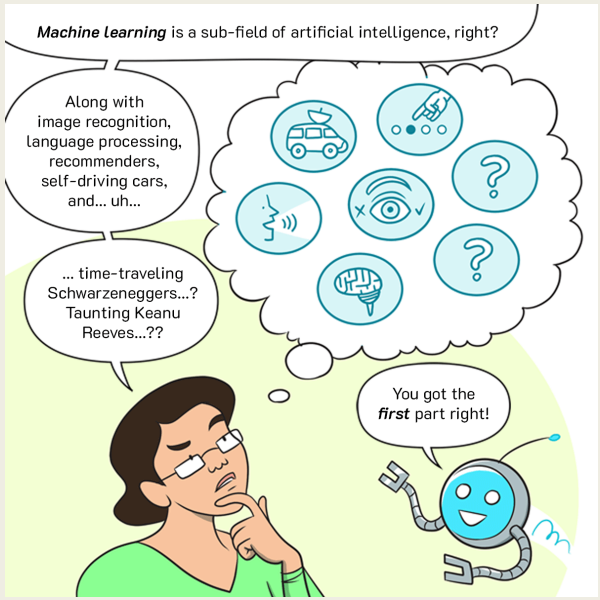
MACHINE LEARNING ALGORITHMS

Machine Learning

- It is a branch of AI that intimates the System to “self-taught” from training data and improve over time.



Focused Algorithms

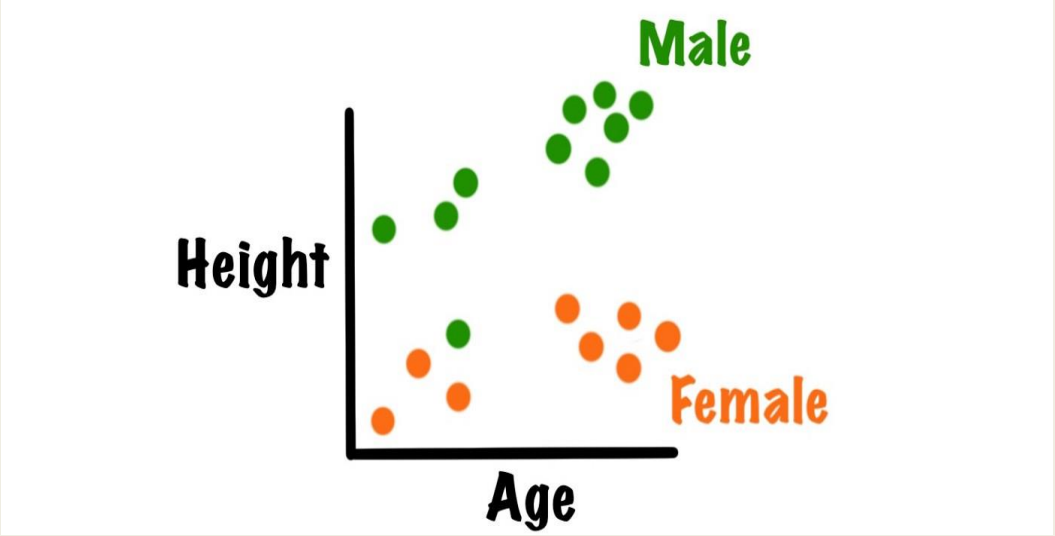
- Support Vector Machine
- Linear Regression
- Logistic Regression
- Naive Bayes
- KNN

- K-means
- Decision Tree
- Artificial Neural Network
- Random Forest
- Apriori Algorithm

Support Vector Machine Algorithm

It is a classification Method, In this Algorithm, we plot each data item as a point in a dimensional space based on features that the dataset has with the value of each feature being the value of a particular coordinate in the plot

- For example, if we had features like Height and Age of male and female groups, we first plot these features in dimensional spaces, then we will find some lines that split data between two classified groups.



Linear Regression

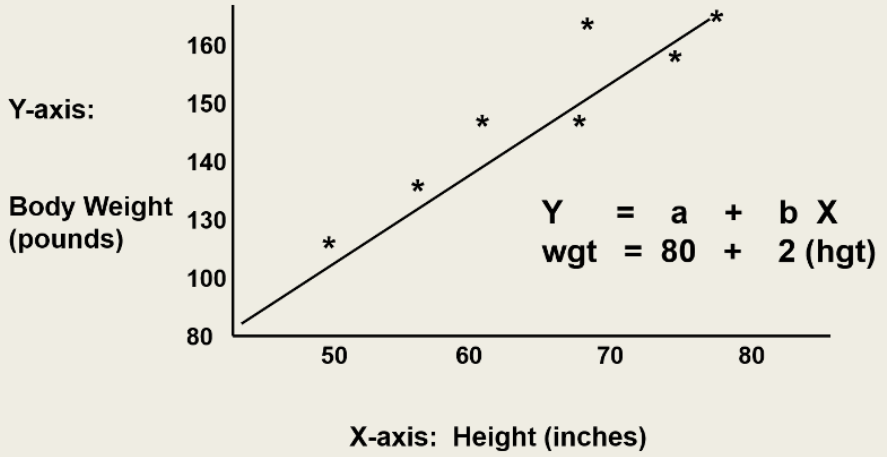
This algorithm shows the relationship between an independent and dependent variable, Here we establish the relationship between the independent and dependent variable by fitting the best line.

This line is represented by the linear equation **Y = a + b X**

Here Y is the Dependent variable, and X is the independent variable.

a is the slope and b is the intercept

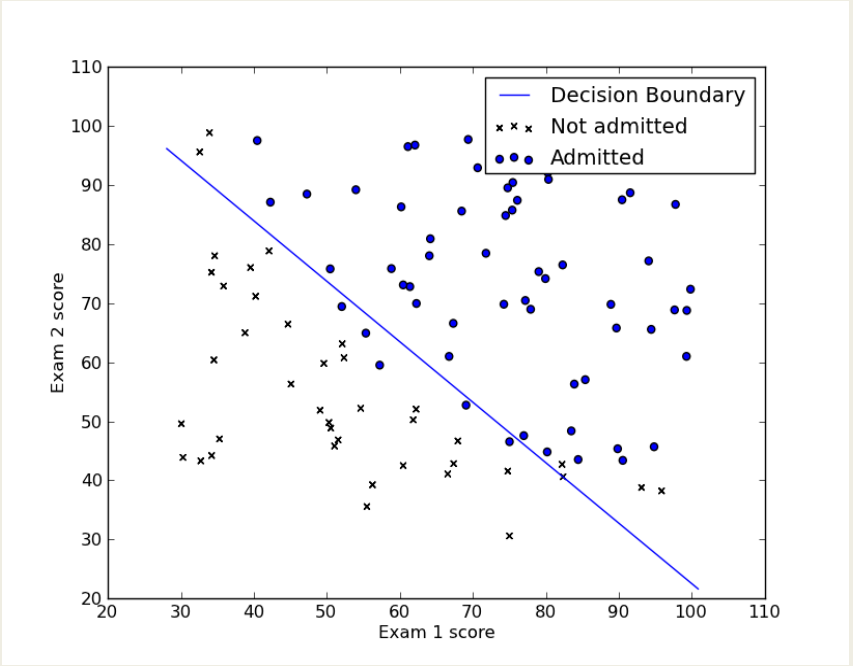
- For example, let's consider the features having Body weight as a dependent variable and Height as the independent variable, here we calculated the best-fit line having linear equation Y= 80+2X, Now we can find the weight of body, knowing height of a person



Logistic Regression

It deals with discrete values, It best suits for the binary classification where if an event occurs, it is classified as 1, if not classified as 0 based on given set of independent features, in simple terms it predicts the probability of occurrence of an event by fitting data to a logit regression.

- Let's consider an example of independent variables containing variables exam 1 and exam 2 scores, based on decision boundary it predicts the students who are admitted and students who are not admitted



Naïve Bayes

This algorithm is based on the Bayes theorem of probability and it allocates the element value to a variable from one of the categories.

This Algorithm is noted by using the notation $P(y|X)=P(X|y)\,P(Y)/P(X)$

Here y is a class variable and X is a dependent feature vector of size n

- For example, let's consider an example of having a dataset Frequency and Likelihood table of **"color"**

Frequency Table

		Stolen?	
		Yes	No
Color	Red	3	2
	Yellow	2	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

- For this table, the following conversion was made
- Convert the given dataset into frequency tables
- Generate a Likelihood table by finding the probabilities of given features
- Now, Use the algorithm to calculate the probability

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

KNN Algorithm

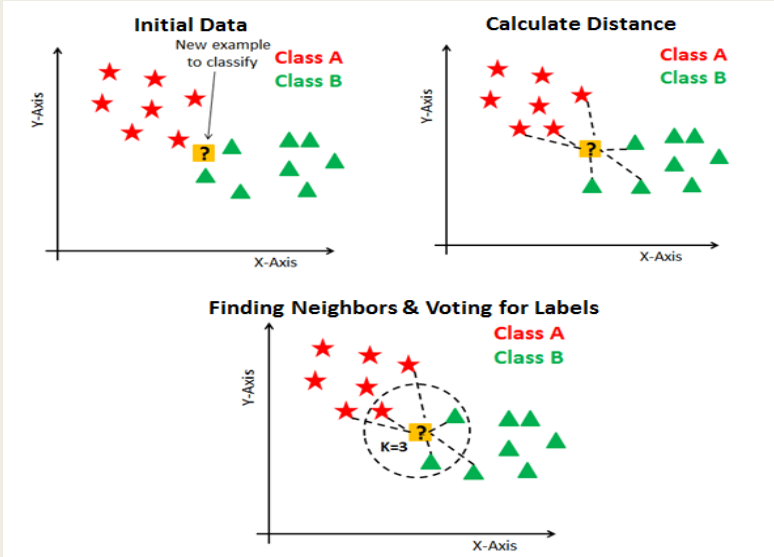
This Algorithm can be used for both classification and regression problems, K-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by majority

It is based on simple distance function, a prediction is made for new point by searching through the entire data set for the K most similar neighbors and summarizing the output variable for K.

- This Algorithm contains three basic steps:
  Calculate the distance

  Find the k-nearest neighbors

  Vote for classes



K-Means

This algorithm is used for solving clustering problems, Its procedure is very simple and easy to classify the data set through a certain number of clusters.

In K-means, we have clusters and each cluster has its own centroid, the sum of differences between centroids and the data points within a cluster constitutes the sum of the square value for that cluster

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in an unlabeled dataset on their own without training

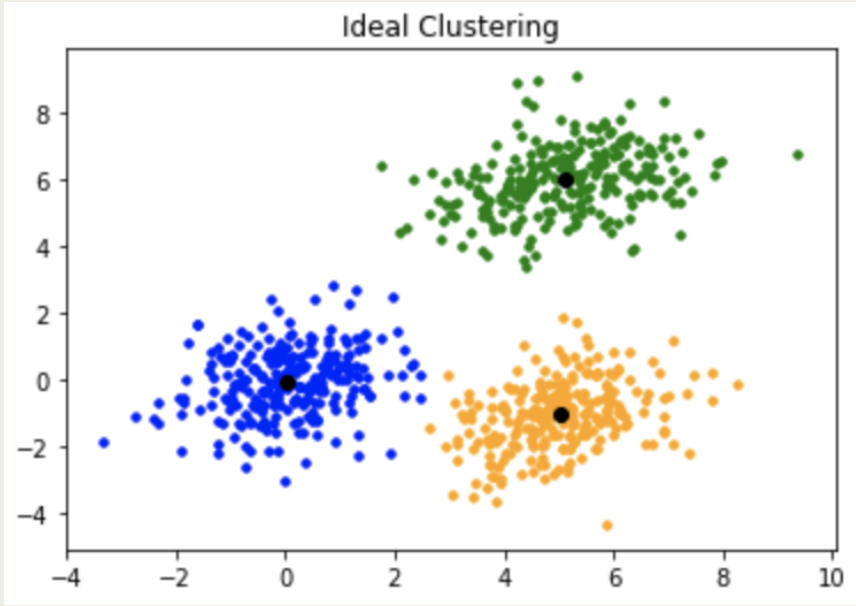
- Steps involved in K-Means :

✓ K-means picks k number of points for each cluster

✓ Each data point forms a cluster with the closest clusters

✓ Find the centroid of each cluster based on an existing cluster, here we have new centroids

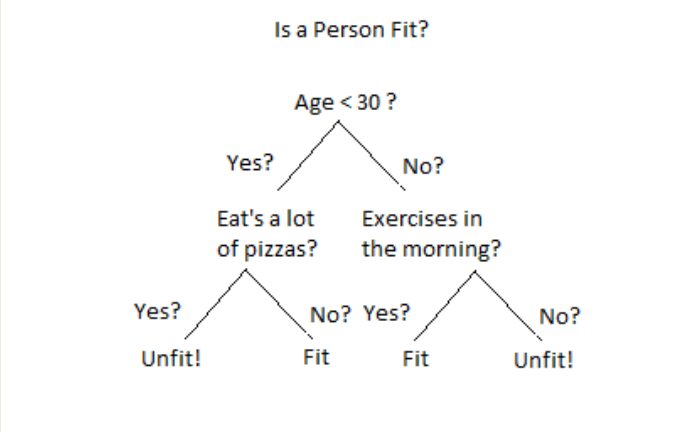
✓ As we have new centroids, find the closest distance for each data point from new centroids and get associated with new k-clusters



Decision Tree

This algorithm is mostly used for classification problems, in this algorithm, we split the data into two or more sets, this is done based on the most significant attributes/ independent variables

In decision analysis, a decision tree can be used to visually and explicitly represent the decision-making, it uses a tree-like model of decisions



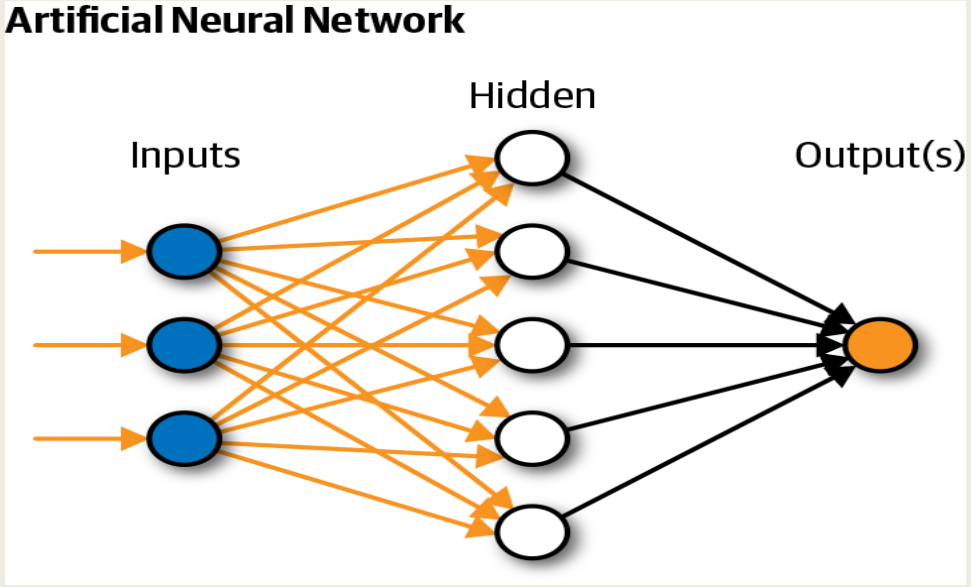
- Let's discuss an example to predict whether a person is fit or not where age is > 30
- The decision nodes contain the questions what's the age? does he exercise? Does he eat a lot of pizzas?
- This is a binary classification that mainly contains YES or NO types.
- Firstly it asks the age of the person, if the answer is in parameters then it classifies the person's exercise level if the person’s answers were not on parameters then it classifies the person as unfit

Artificial Neural Network

Artificial neural networks are comprised of a node layer, containing an input layer, one or more hidden layers, and an output layer

Each node connects to another and has an associated weight and threshold, if the output of an individual node is above the specified threshold value, the node is activated otherwise, no data is passed along to the next layer

- Once an input layer is determined, weights are assigned, these weights help to determine the importance of giving variables
- All the inputs were multiplied by their respective weights and then summed, the output is passed through an activation function which determines the output
- If the output exceeds the threshold, it fires the nodes passing to the next layer



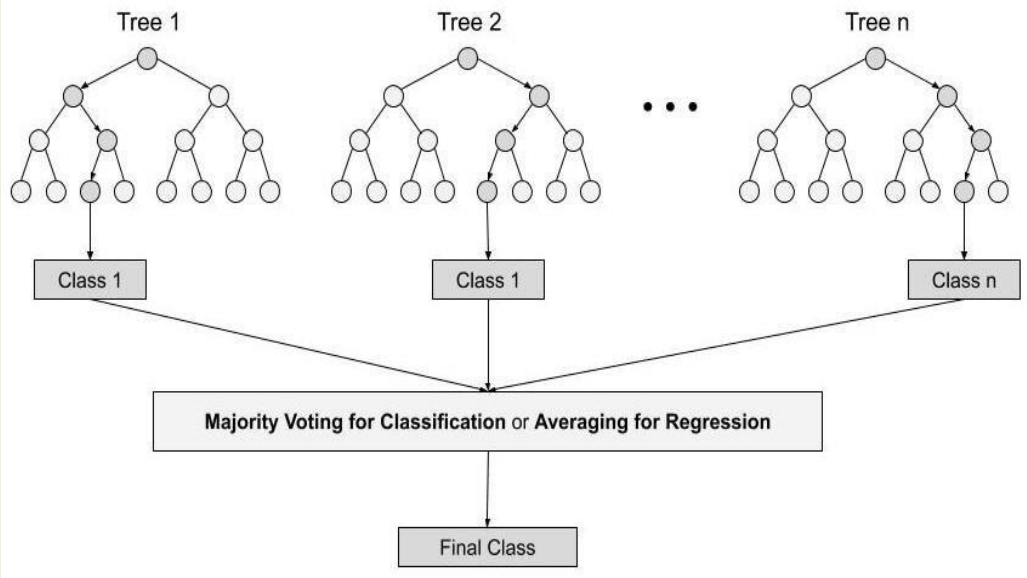
Random Forest

This Algorithm is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the accuracy of the dataset

The Greater number of trees lead to higher accuracy

Random forest combines multiple trees to predict the class of the dataset, it is possible that some decision tree produce valid output and others may not.

- Process flow chart:
- Select the random K data points from training set
- Build the decision tree associated with the selected data point
- Choose the N for decision trees that you want to build
- For new data points, find the prediction of each decision tree, and assign the new data points to category that wins the majority votes.



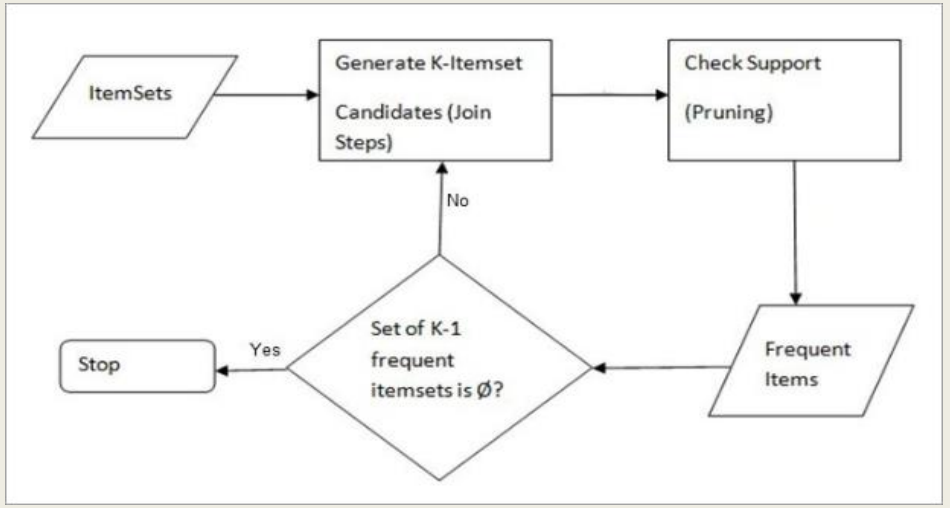
Apriori Algorithm

This algorithm is used to mine frequent itemset and subsequently construct association rules

It uses the IF-THEN format to build rules, i.e. if event A occurs, then event B is likely to occur with a certain probability

- Steps for the Algorithm
- Computing the support for each itemset
- Deciding on the support threshold
- Select the frequent items
- Finding the support of a frequent itemset
- Generate association rules and compute confidence
- Compute lift

$$\text{Lift} = P(X \cap Y)/P(X) * P(Y)$$



THANK YOU