DATA SCIENCE

Prajwal S
ENG18CS0211
6th SEM D SECTION

01 Introduction to Machine Learning

02 SL, USL & RL

03 Regression, Classification, Clustering, Association

04 Steps involved in Machine Learning program

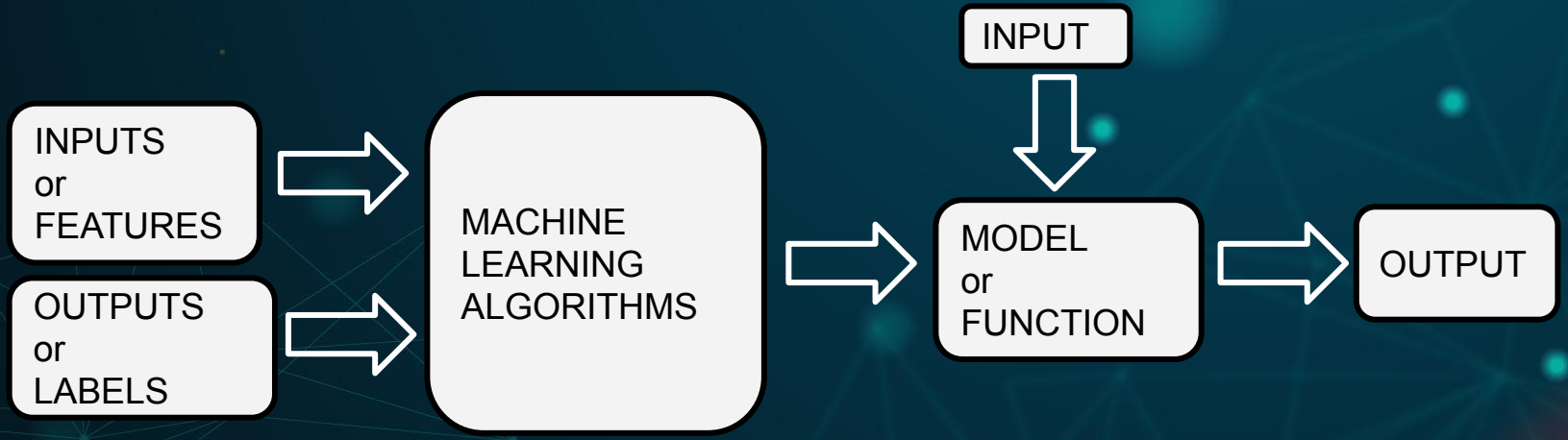
Introduction to Machine Learning

01

- Machine Learning (ML) is a branch of Artificial Intelligence, concerned with design and development of algorithms that allows computers to evolve behaviours based on empirical data.

- Machine Learning uses statistical techniques to give computer system the ability to learn with data without being explicitly programmed.

- In ML, we give input and output data to ML algorithms and a model is built. To this model, when we give a new set of inputs, it predicts the output.



Supervised, Unsupervised & Reinforcement Learning

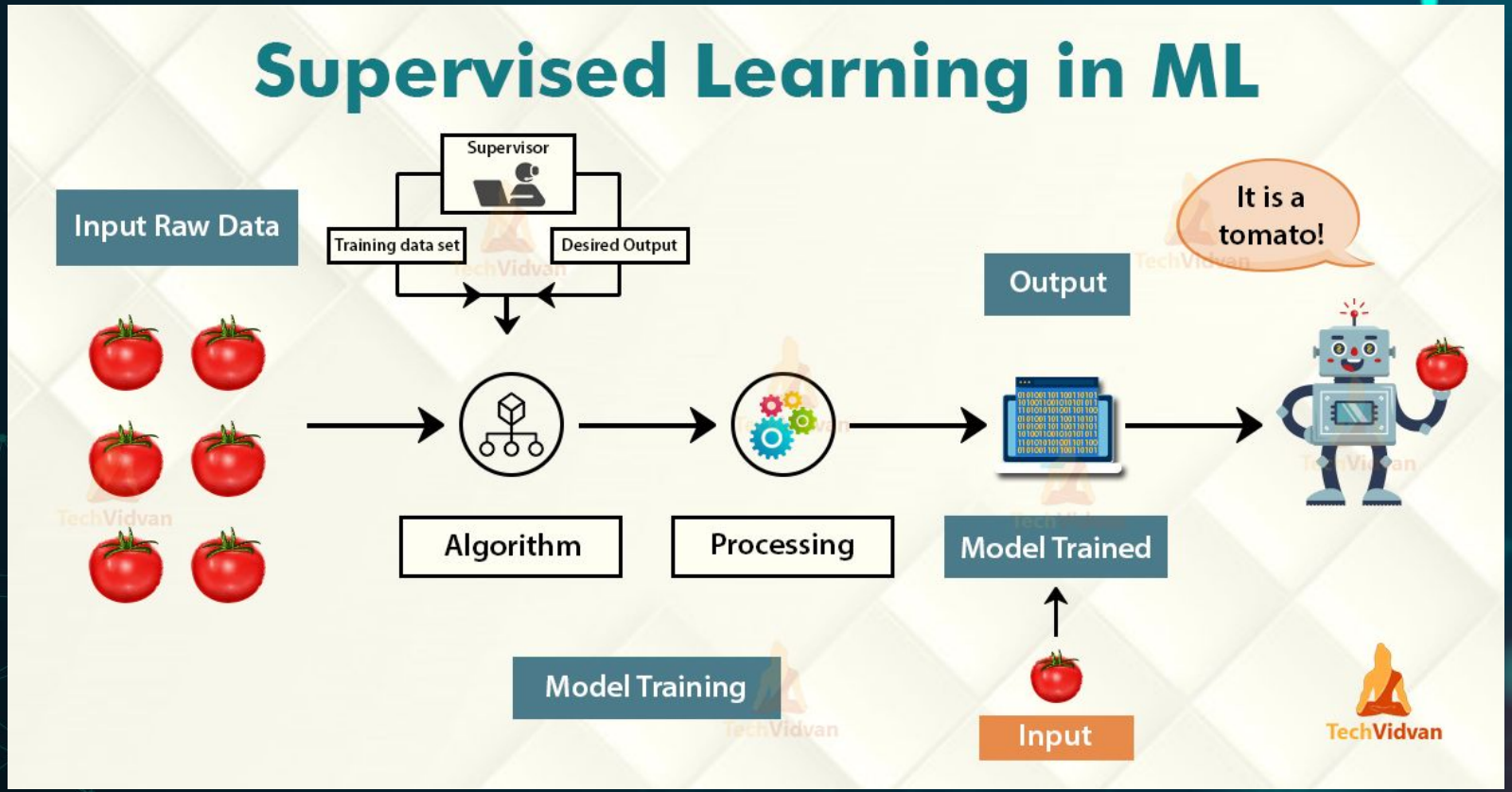
02

SUPERVISED LEARNING

- In supervised learning, we give features and labels to the model. With help of a supervisor, this model will predict and give output.

- Supervised Learning is of 2 types :
  - Regression
  - Classification

SUPERVISED LEARNING



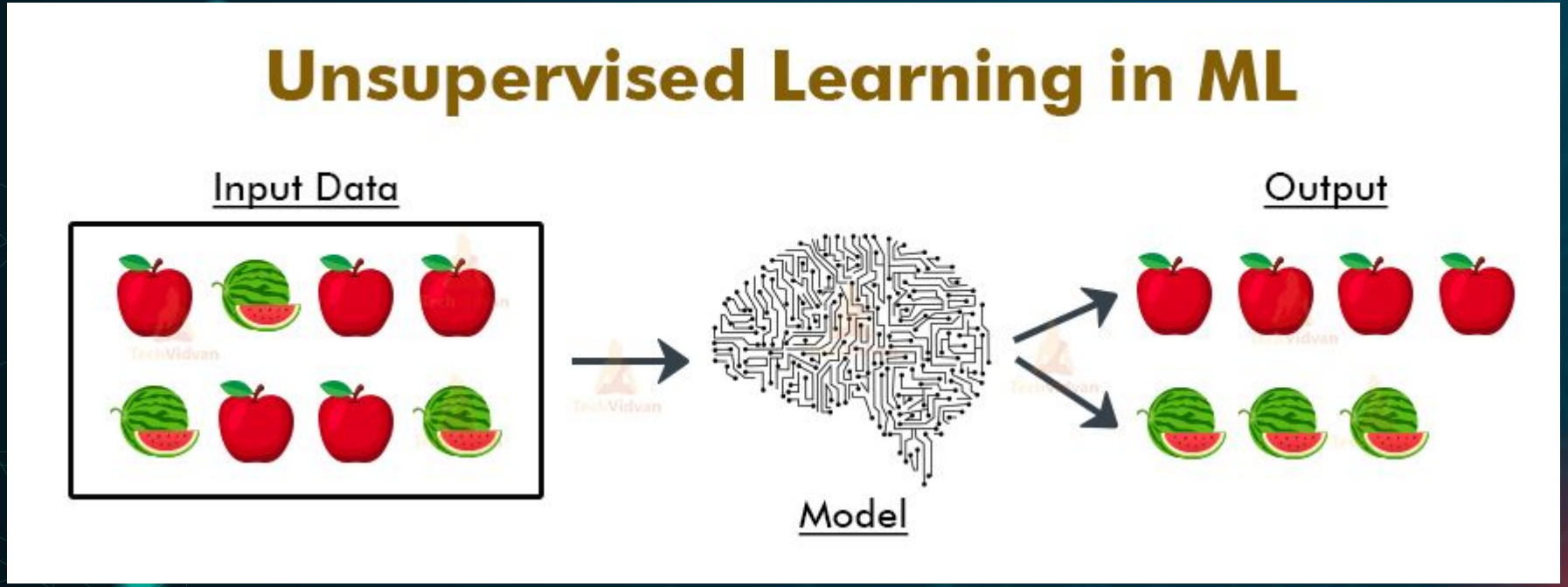
UNSUPERVISED LEARNING

- In unsupervised learning, we give only features to the model. Without the help of a supervisor, this model will predict and give output.

- The model classifies the data according to the features.

- Unsupervised Learning is divided into:
  - Clustering
  - Association

UNSUPERVISED LEARNING



REINFORCEMENT LEARNING

- In reinforcement learning, we give features and labels to the model and let the model to learn for some time.

- If the model could perform the task, we give a reward, else, a penalty is awarded.

REINFORCEMENT LEARNING

Regression, Classification, Clustering, Association

03

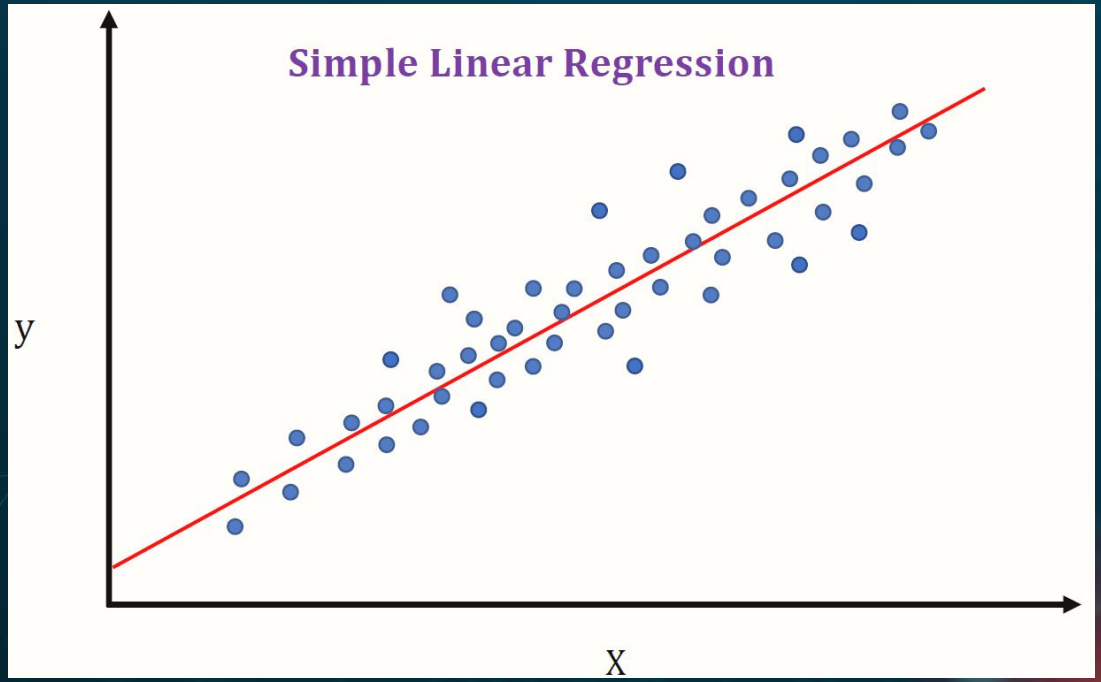
REGRESSION

- Regression is a statistical method used to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variable/s (known as independent variable/s).

- The most commonly used regression algorithm is Linear Regression.

LINEAR REGRESSION

- It is a ML algorithm, which includes modelling with the help of a dependent variable.
- The format of the projection for this model is y= mx+c.
- Linear regression helps us to understand the relationship between two variables.

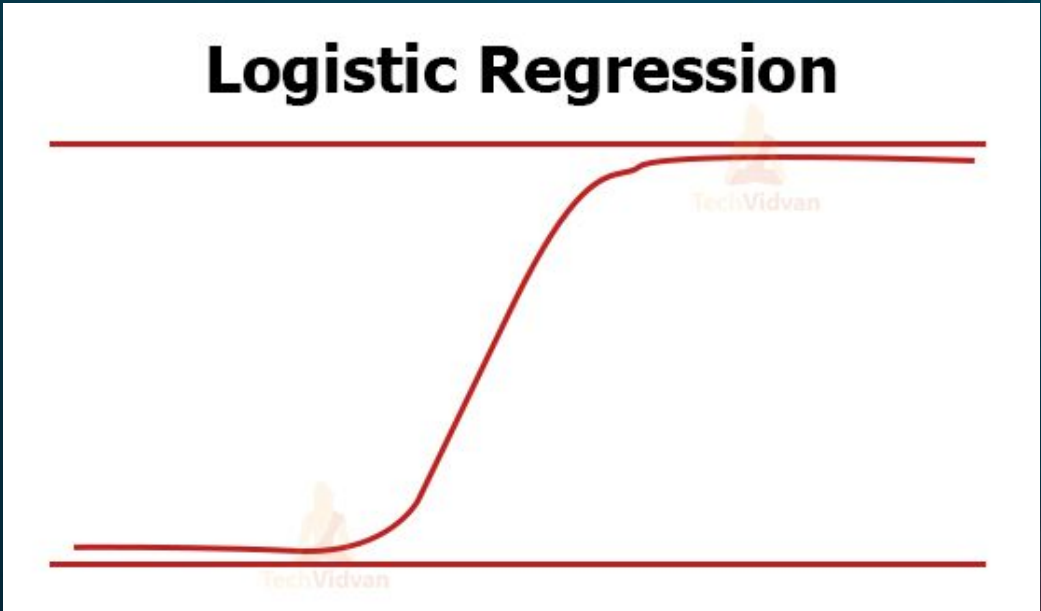


CLASSIFICATION

- Classification is a process of categorizing a given set of data into classes or groups.
- The various classification algorithms available are:
  - Logistic Regression
  - K-Nearest Neighbors
  - Decision Tree
  - Random Forest
  - Naive Bayes
  - Support Vector Machine (SVM)

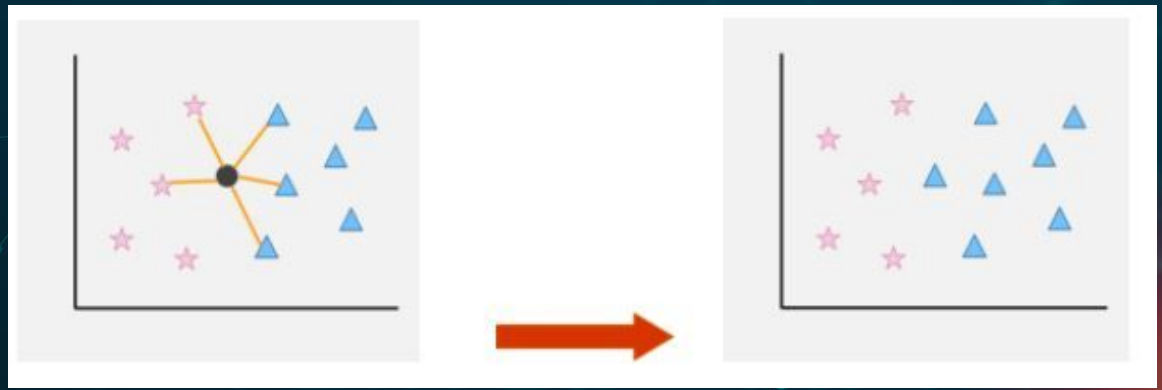
LOGISTIC REGRESSION

- Logistic regression helps us to understand the relationship between one binary dependent variable and an independent variable.
- It is shown as $y = \ln(P/(1-P))$.
- The above function is a sigmoid function.
- The graph for this is S-shaped.



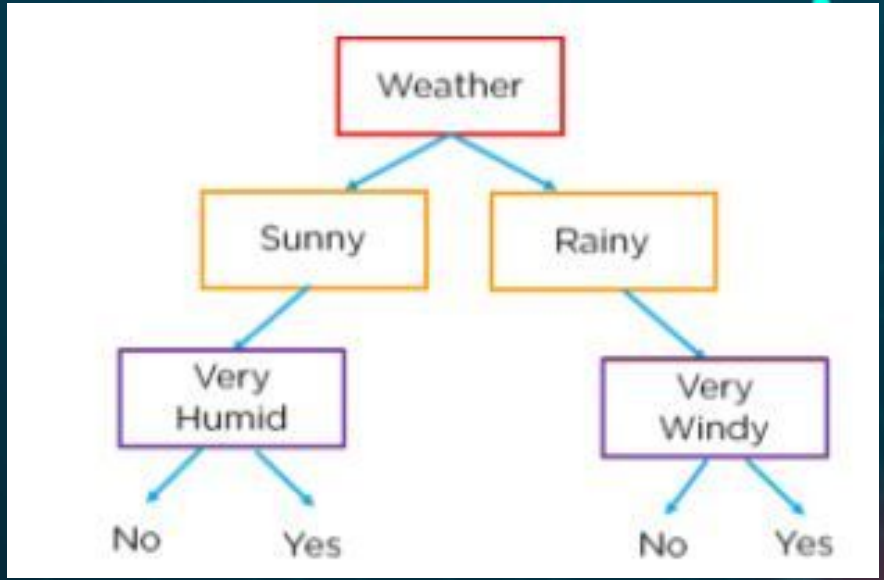
K - NEAREST NEIGHBORS

- K-Nearest Neighbor is a classification and prediction algorithm which is used to divide data into classes based on the distance between the data points.
- K-Nearest Neighbor assumes that data points which are close to one another must be similar and hence, the data point to be classified will be grouped with the closest cluster.



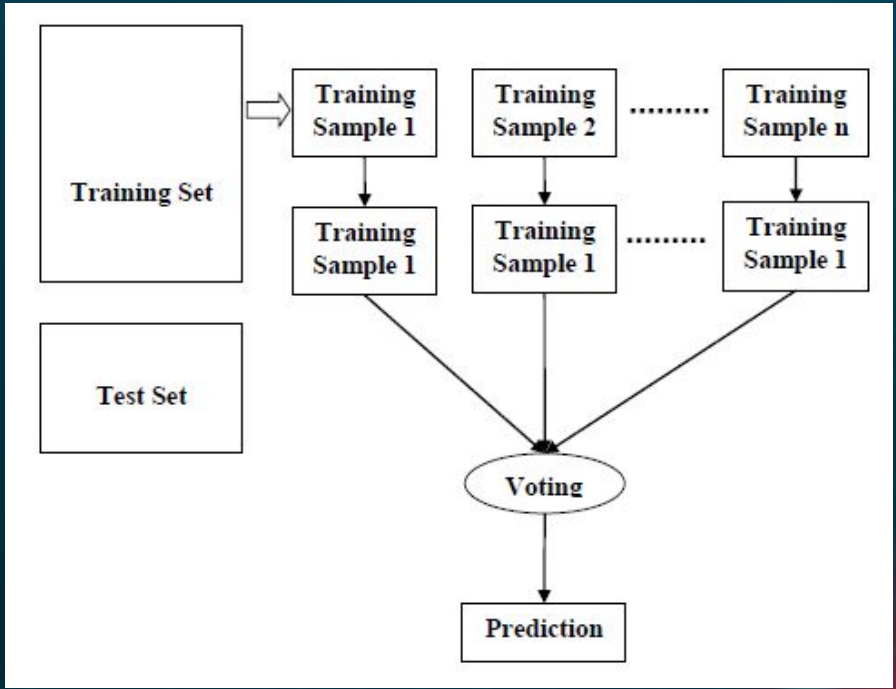
DECISION TREE

- A Decision Tree is an algorithm that is used to visually represent decision making.
- A Decision Tree can be made by asking a yes/no question and splitting the answer to lead to another decision.
- The question is at the node and it places the resulting decisions below at the leaves.
- Let’s consider an example of playing a tennis match.



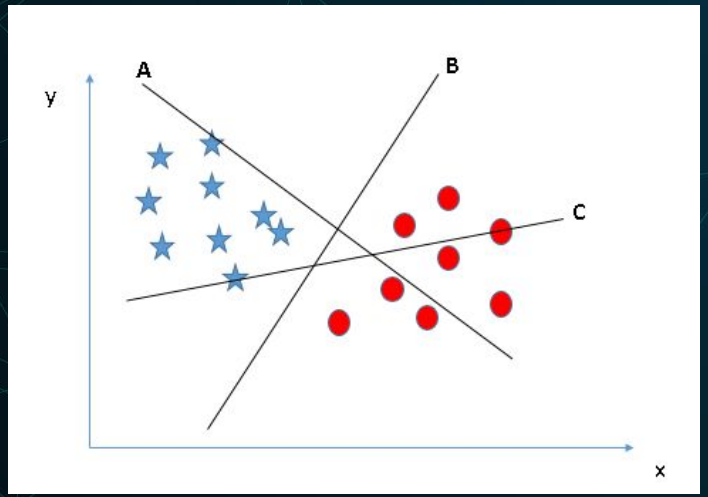
RANDOM FOREST

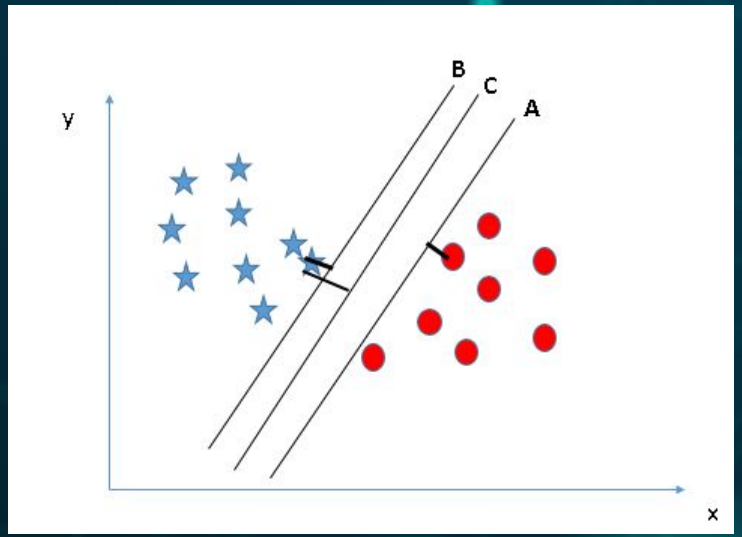
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



SUPPORT VECTOR MACHINE - SVM

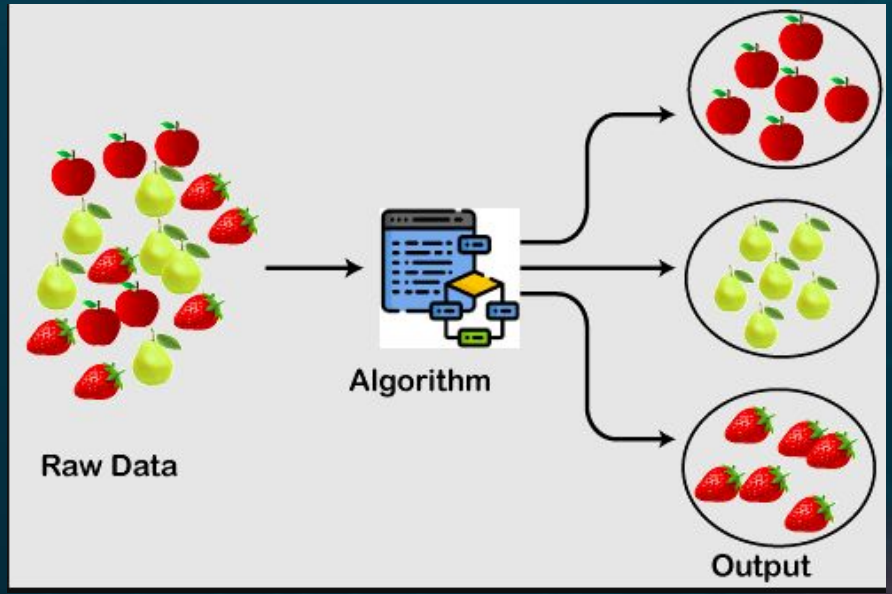
- In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.
- Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.





CLUSTERING

- Clustering is the task of dividing the population into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- It is basically a collection of objects on the basis of similarity and dissimilarity between them.
- Clustering is somewhere similar to the classification algorithm, but the difference is the type of dataset that we are using.
- In classification, we work with the labeled data set, whereas in clustering, we work with the unlabelled dataset.



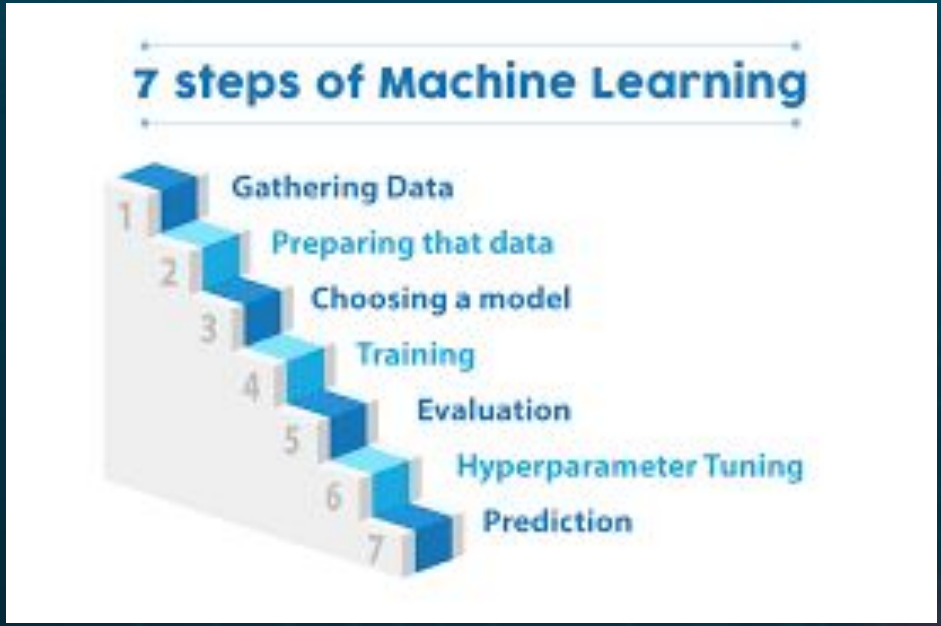
ASSOCIATION

- Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.
- It tries to find some interesting relations or associations among the variables of dataset.
- It is based on different rules to discover the interesting relations between variables in the database

Steps Involved In Machine Learning

04

- Machine Learning involves 7 steps:
  - Collecting the data
  - Preparing the data
  - Choosing the algorithm
  - Training the model
  - Evaluating the model
  - Hyperparameter Tuning
  - Prediction



- COLLECTING THE DATA:
  - Be it the raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the machine learning.
  - The better the variety, density and volume of relevant data, better the learning prospects for the machine becomes.

- PREPARING THE DATA:
  - Now, we visualize the data and check if there are correlations between the different characteristics that we obtained.
  - Exploratory analysis is perhaps one method to study the nuances of the data in details thereby burgeoning the nutritional content of the data.

- CHOOSING THE ALGORITHM:
  - There are several models that you can choose according to the objective that you might have: you will use algorithms of classification, prediction, linear regression, clustering, i.e. k-means or K-Nearest Neighbor, etc.
  - There are various models to be used depending on the data you are going to process such as images, sound, text, and numerical values.

- TRAINING THE MODEL:
  - The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.
  - The test data and train data can be divided approximately in a ratio of 80/20.
  - We need need to train the datasets to run smoothly and see an incremental improvement in the prediction rate.

- EVALUATING THE MODEL:
  - We will have to check the model created against our evaluation data set that contains inputs that the model does not know and verify the precision of the already trained model.
  - If the accuracy is less than or equal to 50%, that model will not be useful since it would be like tossing a coin to make decisions.
  - If we reach 90% or more accuracy, we can have good confidence in the results that the model gives.

- HYPERPARAMETER TUNING:
  - If during the evaluation we did not obtain good predictions and our precision is not the minimum desired, it is possible that we have overfitting - or underfitting problems and must return to the training step before making a new configuration of parameters in the model.
  - We can increase the number of times you iterate your training data- termed epochs.

- PREDICTION:
  - Now, we ready to use the Machine Learning model inferring results in real-life scenarios.
  - This is the stage where we consider the model to be ready for practical applications.
  - The model gains independence from human interference and draws its own conclusion on the basis of its data sets and training.
  - The prediction step is what the end-user sees when they use the machine learning model within their respective industry.

THANK YOU