Basic Concepts in Machine Learning

Machine Learning is continuously growing in the IT world and gaining strength in different business sectors.

Although Machine Learning is in the developing phase, it is popular among all technologies.

It is a field of study that makes computers capable of automatically learning and improving from experience.

Hence, Machine Learning focuses on the strength of computer programs with the help of collecting data from various observations.

Machine Learning enables computers to behave like human beings by training them with the help of past experience and predicted data.

There are three key aspects of Machine Learning

Task

Experience

Performance

Task:

A task is defined as the main problem in which we are interested. This task/problem can be related to the predictions and recommendations and estimations, etc.

Experience:

It is defined as learning from historical or past data and used to estimate and resolve future tasks.

Performance:

It is defined as the capacity of any machine to resolve any machine learning task or problem and provide the best outcome for the same. However, performance is dependent on the type of machine learning problems.

Applications of Machine Learning

Machine Learning is widely being used in approximately every sector, including healthcare, marketing, finance, infrastructure, automation, etc.

Difference between machine learning and Artificial Intelligence:

Artificial Intelligence is a technology used to create an intelligent system that enables a machine to simulate human behavior. Whereas, Machine Learning is a branch of AI which helps a machine to learn from experience without being explicitly programmed.

Applications of AI: Siri, customer support using catboats, Expert System, Online game playing, an intelligent humanoid robot, etc.

Applications of ML: Online recommender system, Google search algorithms, Facebook auto friend tagging suggestions, etc.

Traditional Programming

Machine Learning

Machine Learning	Artificial Intelligence
Machine learning is a method of data analysis that automates analytical model building.	Artificial Intelligence is a method of data analysis that makes your model intelligent.
Machine Learning results in data	Machine Learning results in Knowledge or making your system intelligent
The aim is to extend accuracy	The aim is to extend probability of success
ML permits system to be told new things from knowledge.	AI is the higher cognitive process.

Commonly used Machine Learning Algorithms

- Linear Regression
- Logistic Regression
- K Nearest Neighbour
- K-Means Clustering
- Decision Tree
- Random Forest
- Support Vector Machines
- Naïve Bayes

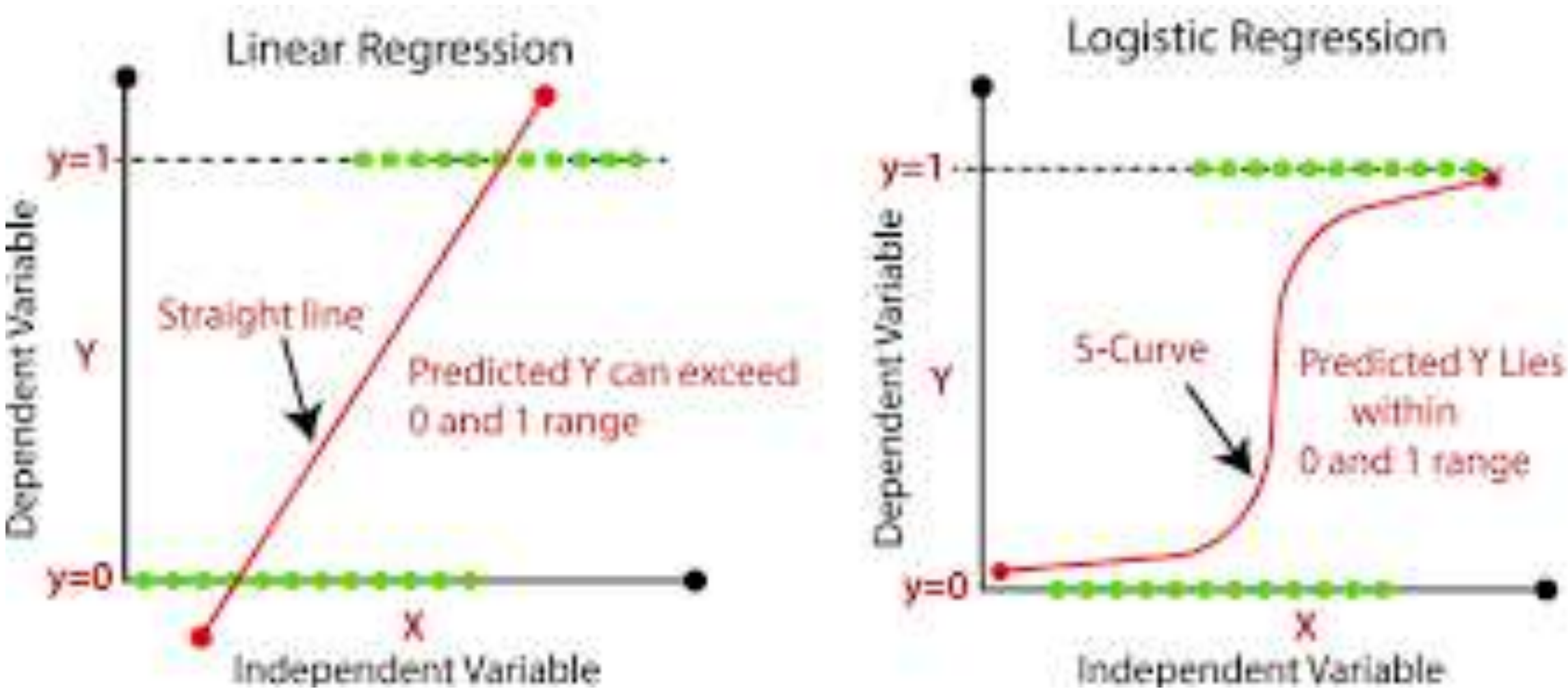
Linear Regression

Linear Regression is one of the simplest and popular machine learning algorithms recommended by a data scientist.

It is used for predictive analysis by making predictions for real variables such as experience, salary, cost, etc.

It is a statistical approach that represents the linear relationship between two or more variables, either dependent or independent, hence called Linear Regression.

It shows the value of the dependent variable changes with respect to the independent variable, and the slope of this graph is called as Line of Regression.


Linear Regression
y=1
y=0
Y
Depnndent Variable
Straight line
Predicted Y can exceed
0 and 1 range
X
Independent Variable
Logistic Regression
y=1
y=0
Y
Dependent Variable
S-Curve
Predicted Y Lies
within
0 and 1 range
X
Independent Variable

Logistic Regression

Logistic Regression is a subset of the Supervised learning technique. It helps us to predict the output of categorical dependent variables using a given set of independent variables.

It can be Binary (0 or 1) as well as Boolean (true/false), but instead of giving an exact value, it gives a probabilistic value between o or 1.

It is much similar to Linear Regression, depending on its use in the machine learning model.

As Linear regression is used for solving regression problems, similarly, Logistic regression is helpful for solving classification problems.

Logistic Regression can be expressed as an 'S-shaped curve called sigmoid functions. It predicts two maximum values (0 or 1).

K Nearest Neighbor (KNN)

It is also one of the simplest machine learning algorithms that come under supervised learning techniques.

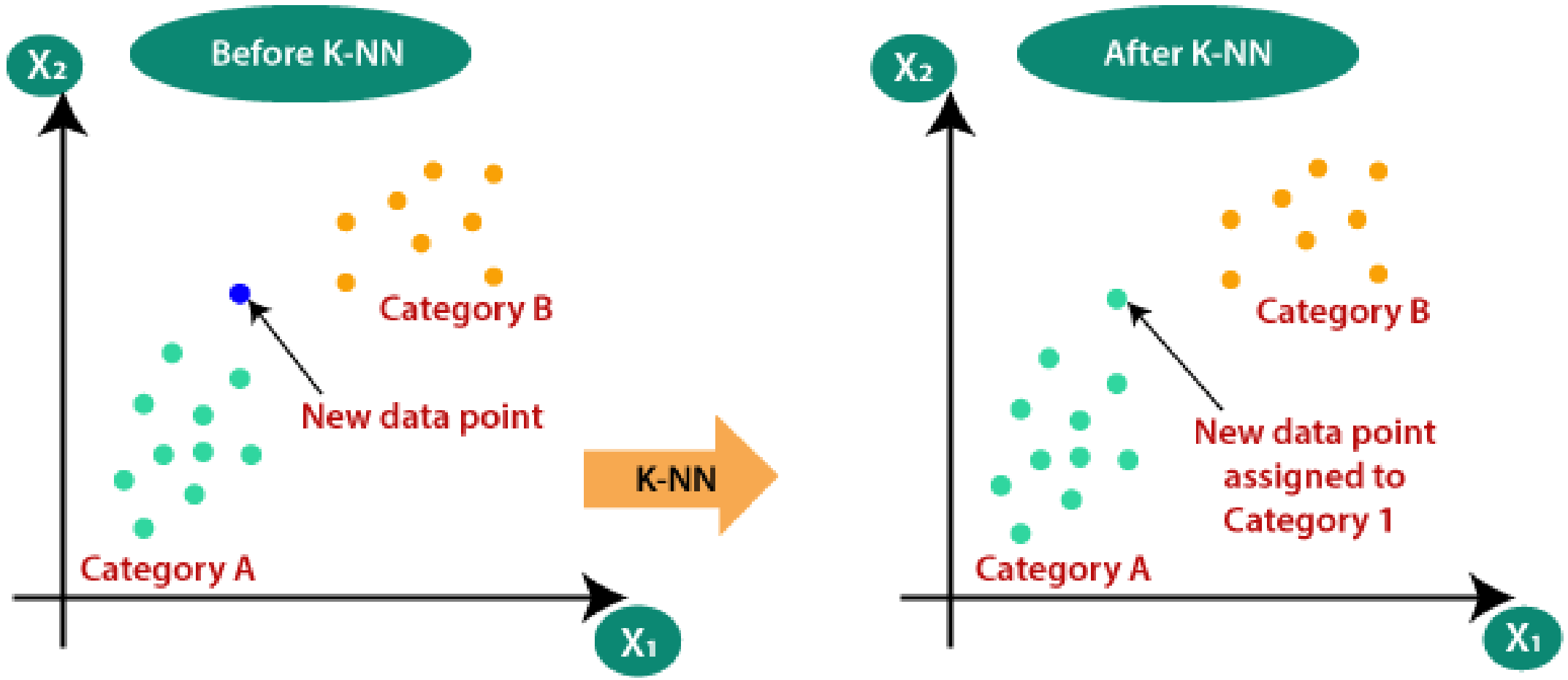
It is helpful for solving regression as well as classification problems. It assumes the similarity between the new data and available data and puts the new data into the category that is most similar to the available categories.

It is also known as Lazy Learner Algorithms because it does not learn from the training set immediately; instead, it stores the dataset, and at the time of classification, it performs an action on the dataset.

Let's suppose we have a few sets of images of cats and dogs and want to identify whether a new image is of a cat or dog.

Then KNN algorithm is the best way to identify the cat from available data sets because it works on similarity measures.

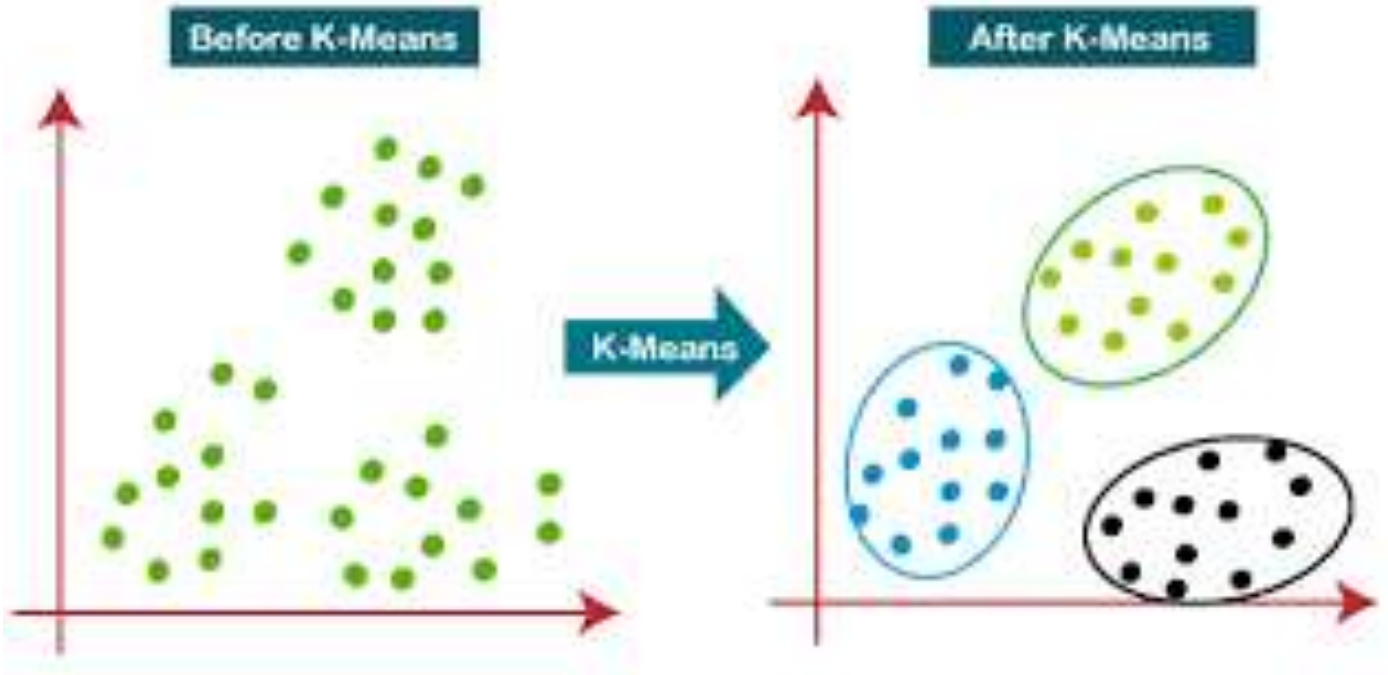
Hence, the KNN model will compare the new image with available images and put the output in the cat's category.


Before K-NN
X2
Category B
New data point
Category A
X1
K-NN
After K-NN
X2
Category B
New data point
assigned to
Category 1
Category A
X1

K-Means Clustering:

K-Means Clustering is a subset of unsupervised learning techniques. It helps us to solve clustering problems by means of grouping the unlabeled datasets into different clusters.

Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



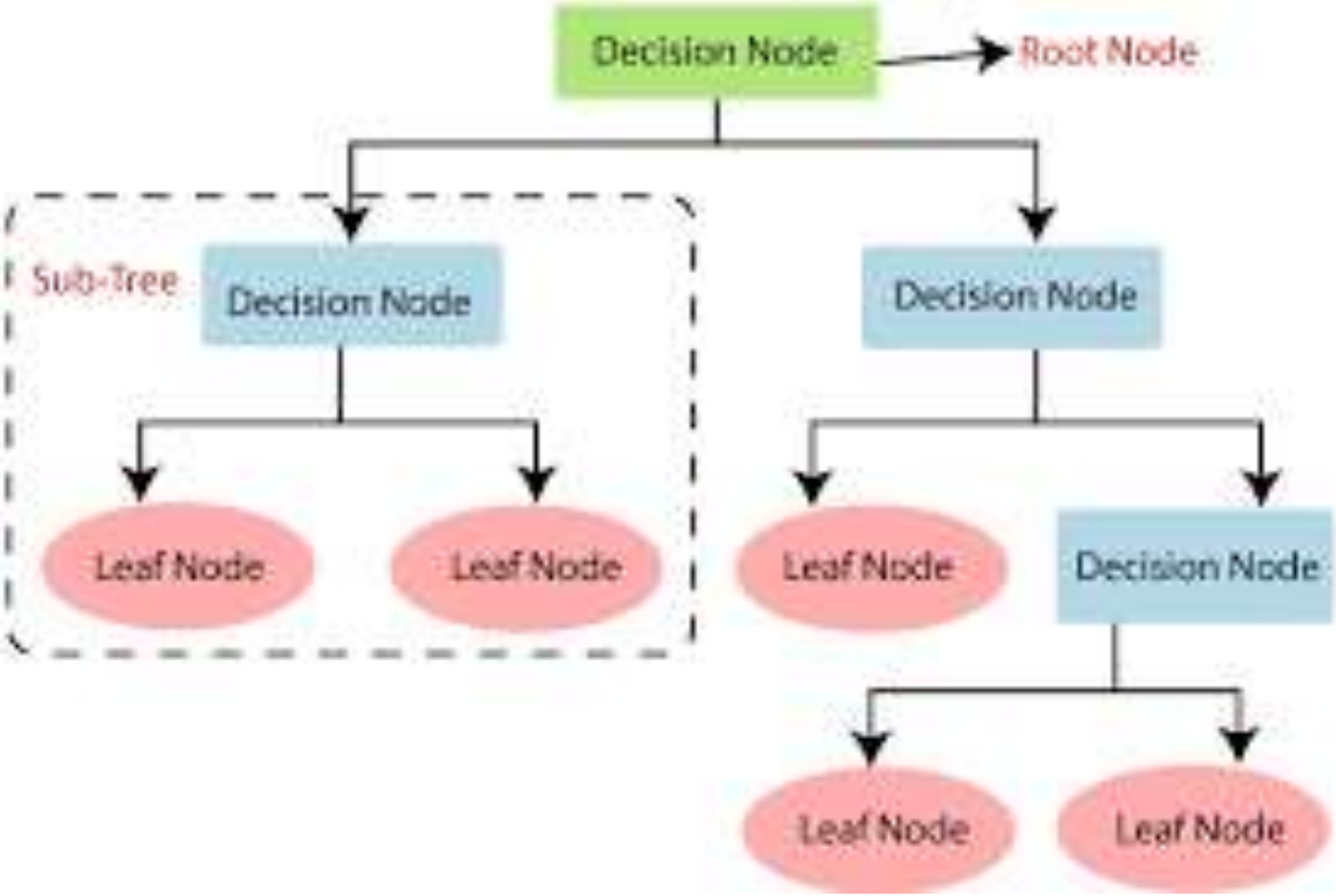
Decision Tree

Decision Tree is also another type of Machine Learning technique that comes under Supervised Learning.

Similar to KNN, the decision tree also helps us to solve classification as well as regression problems, but it is mostly preferred to solve classification problems.

The name decision tree is because it consists of a tree-structured classifier in which attributes are represented by internal nodes, decision rules are represented by branches, and the outcome of the model is represented by each leaf of a tree.

The tree starts from the decision node, also known as the root node, and ends with the leaf node.


Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node

Random Forest

Random Forest is also one of the most preferred machine learning algorithms that come under the Supervised Learning technique.

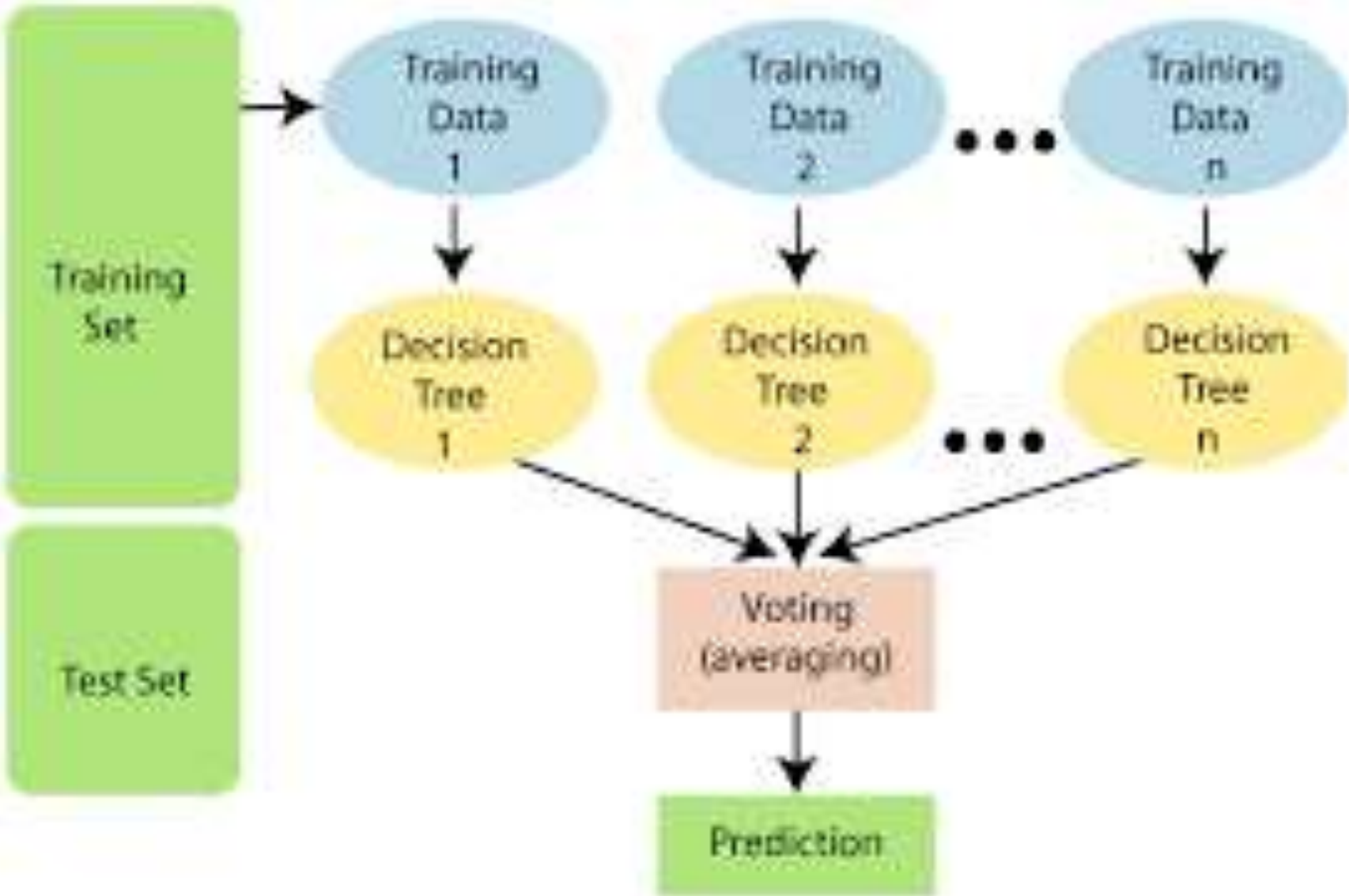
Similar to KNN and Decision Tree, It also allows us to solve classification as well as regression problems, but it is preferred whenever we have a requirement to solve a complex problem and to improve the performance of the model.

A random forest algorithm is based on the concept of ensemble learning, which is a process of combining multiple classifiers.

Random forest classifier is made from a combination of a number of decision trees as well as various subsets of the given dataset.

This combination takes input as an average prediction from all trees and improves the accuracy of the model. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Further, It also takes less training time as compared to other algorithms.

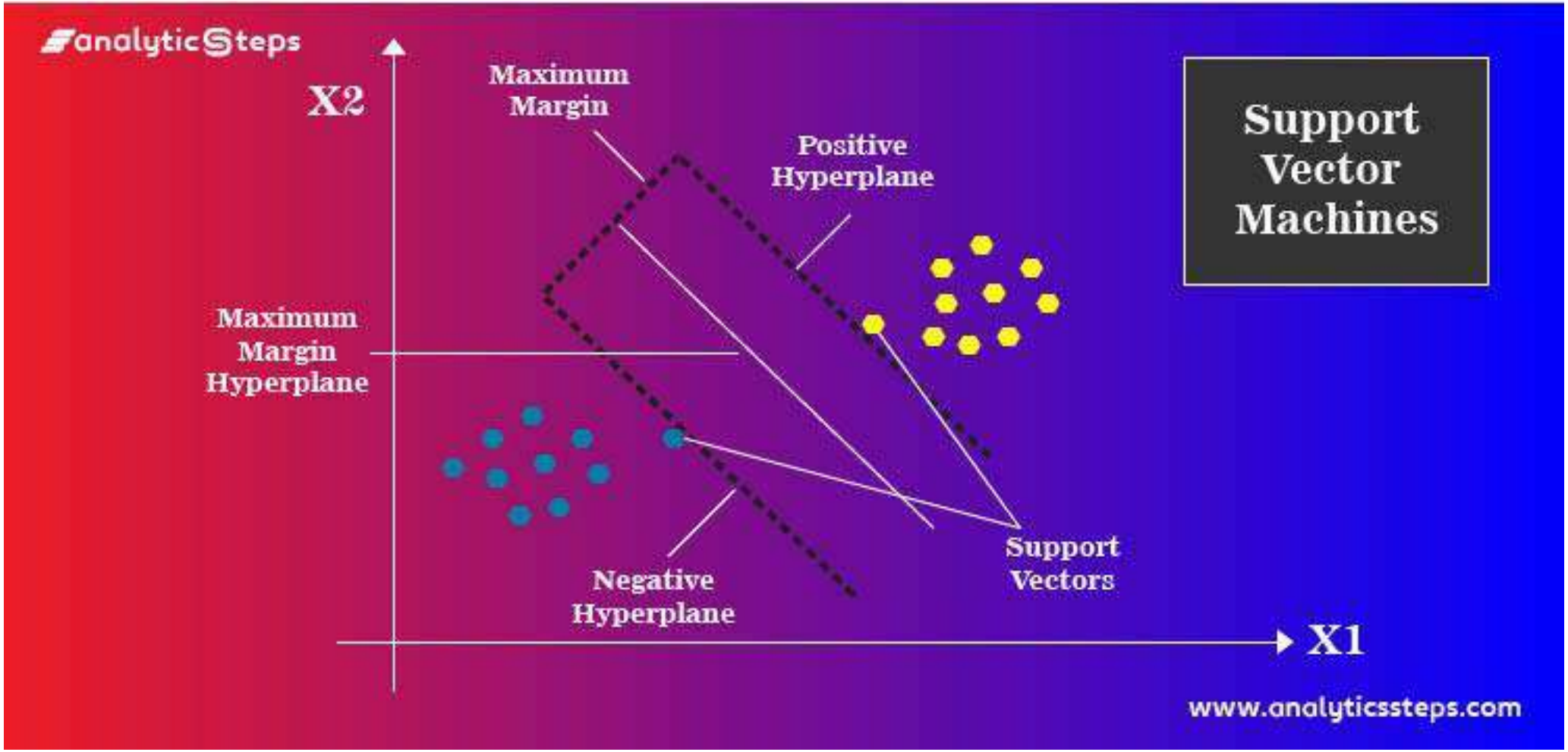

Training Set
Test Set
Training Data 1
Training Data 2
Training Data n
Decision Tree 1
Decision Tree 2
Decision Tree n
Voting (averaging)
Prediction

Support Vector Machines (SVM)

It is also one of the most popular machine learning algorithms that come as a subset of the Supervised Learning technique in machine learning.

The goal of the support vector machine algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. It is also used to solve classification as well as regression problems.

It is used for Face detection, image classification, text categorization, etc.


analyticSteps
X2
Maximum Margin
Positive Hyperplane
Support Vector Machines
Maximum Margin Hyperplane
Negative Hyperplane
Support Vectors
X1
www.analyticssteps.com

Naïve Bayes

The naïve Bayes algorithm is one of the simplest and most effective machine learning algorithms that come under the supervised learning technique.

It is based on the concept of the Bayes Theorem, used to solve classification-related problems.

It helps to build fast machine learning models that can make quick predictions with greater accuracy and performance.

It is mostly preferred for text classification having high-dimensional training datasets.

It is used as a probabilistic classifier which means it predicts on the basis of the probability of an object.

Spam filtration, Sentimental analysis, and classifying articles are some important applications of the Naïve Bayes algorithm.

GAUSSIAN NAIVE BAYES CLASSIFIER

"Gaussian" because this is a normal distribution

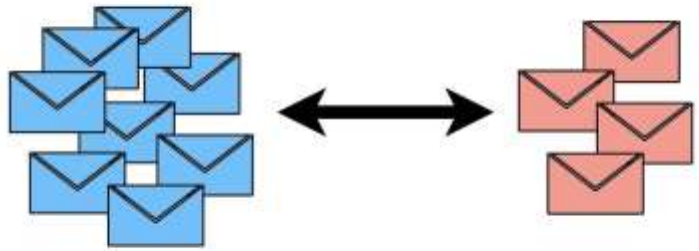
This is our prior belief

$$P(\text{class} \mid \text{data}) = \frac{P(\text{data} \mid \text{class}) \times P(\text{class})}{P(\text{data})}$$

We don't calculate this in naive bayes classifiers

ChrisAlbon

Naive Bayes....



...Clearly Explained!!!