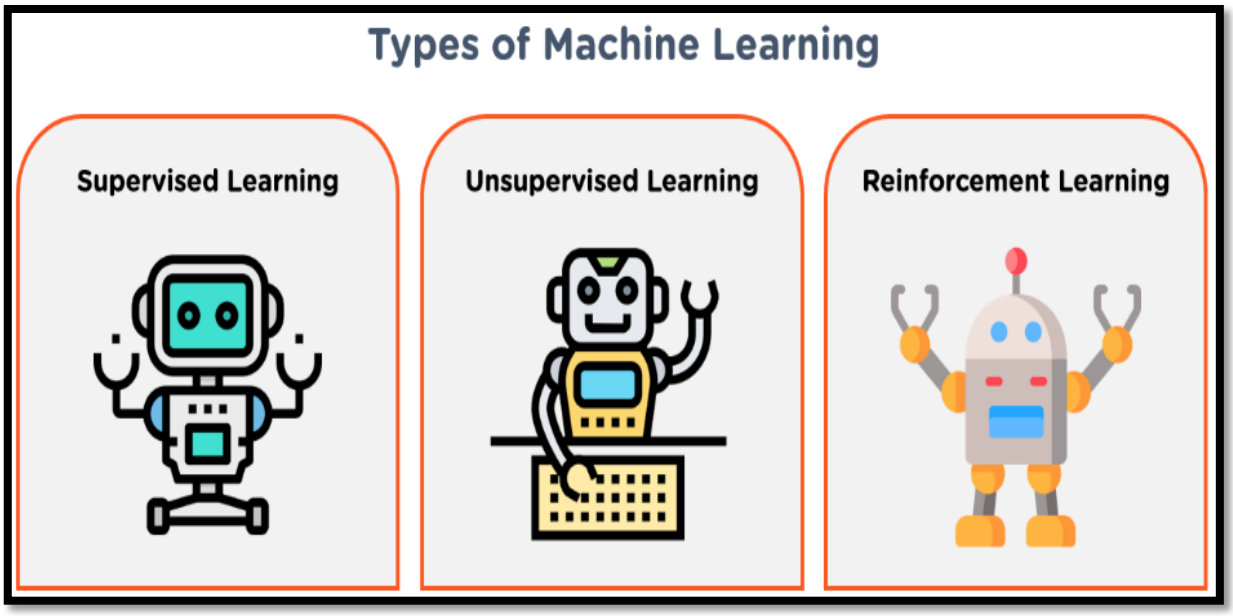
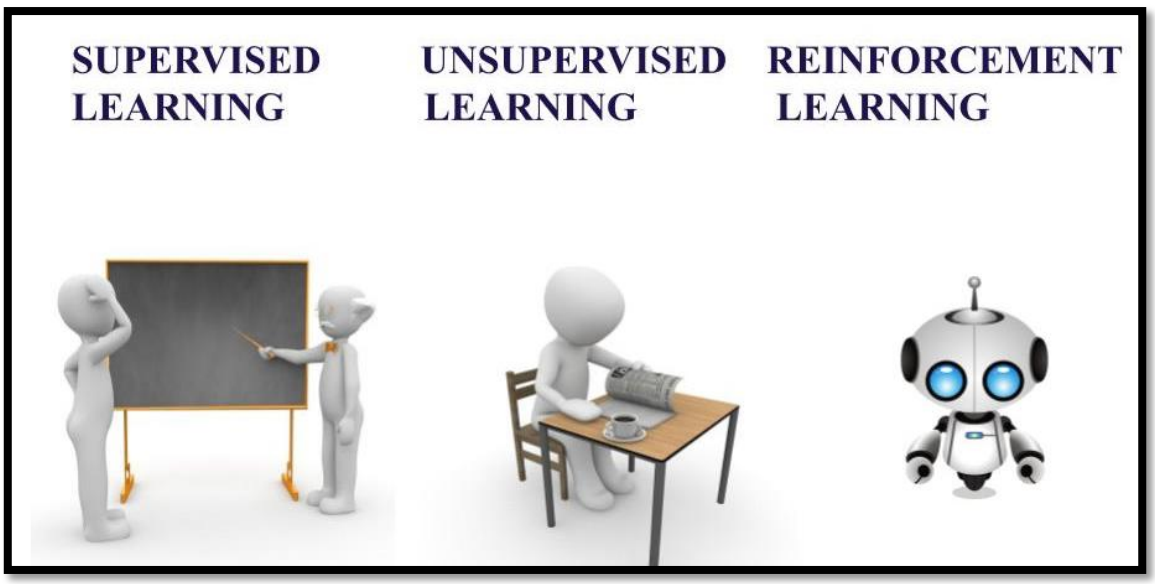
Data Science Application

_

➢ What are algorithms in machine learning?

- A machine learning algorithm **is a set of rules or processes used by an AI system to conduct tasks**—most often to discover new data insights and patterns, or to predict output values from a given set of input variables.
- **Machine learning algorithms** are techniques based on statistical concepts that enable computers to learn from data, discover patterns, make predictions, or complete tasks without the need for explicit programming.
- Algorithms enable machine learning (ML) to learn



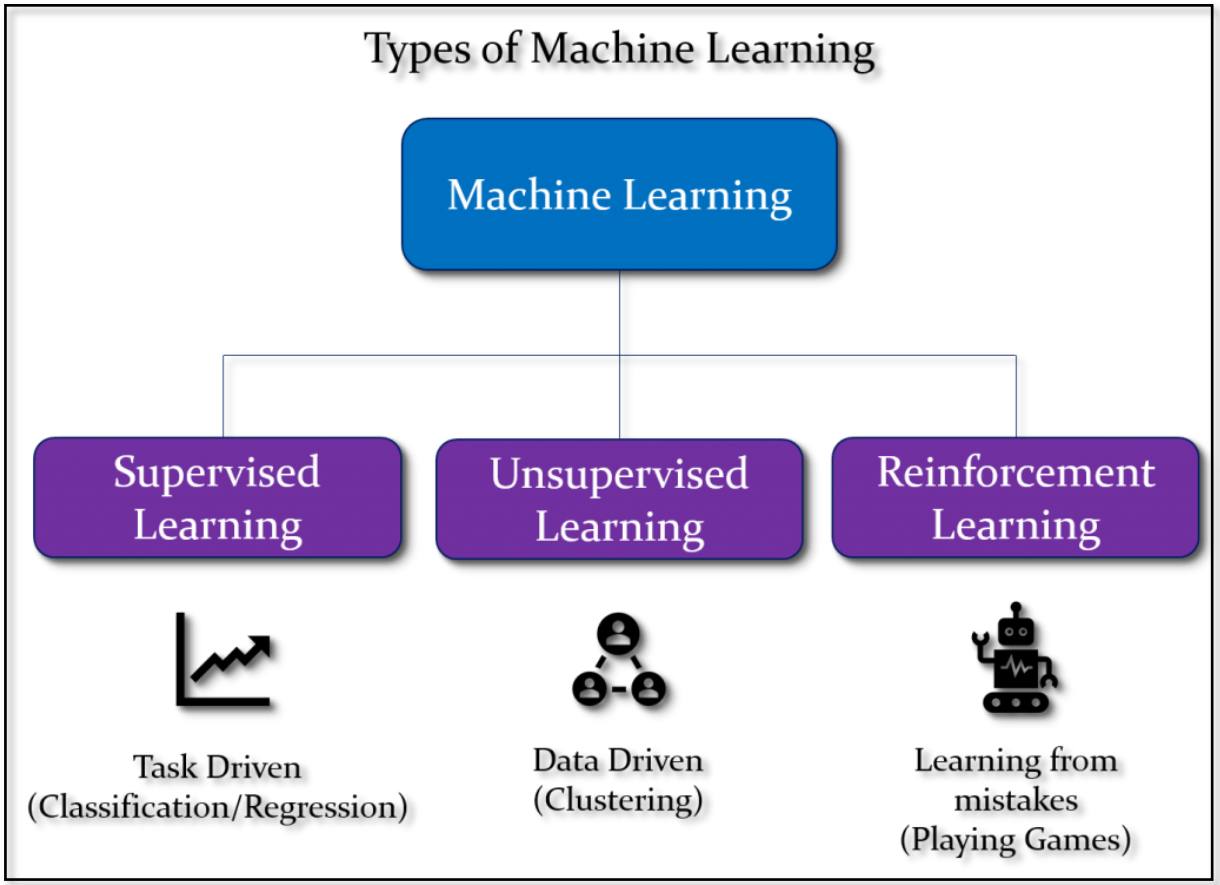


Data Science Application

➢ Types of machine learning algorithms

- Machine learning involves showing a large volume of data to a machine so that it can learn and make predictions, find patterns, or classify data.
- As new data is fed to these algorithms, they learn and optimize their operations to improve performance, developing 'intelligence' over time.
- Depending on your budget, need for speed and precision required, each type and variant has its own advantages.
- There are three types of machine learning algorithms.

1. Supervised Learning
    a. Regression
    b. Classification
2. Unsupervised Learning
    a. Clustering
3. Reinforcement Learning

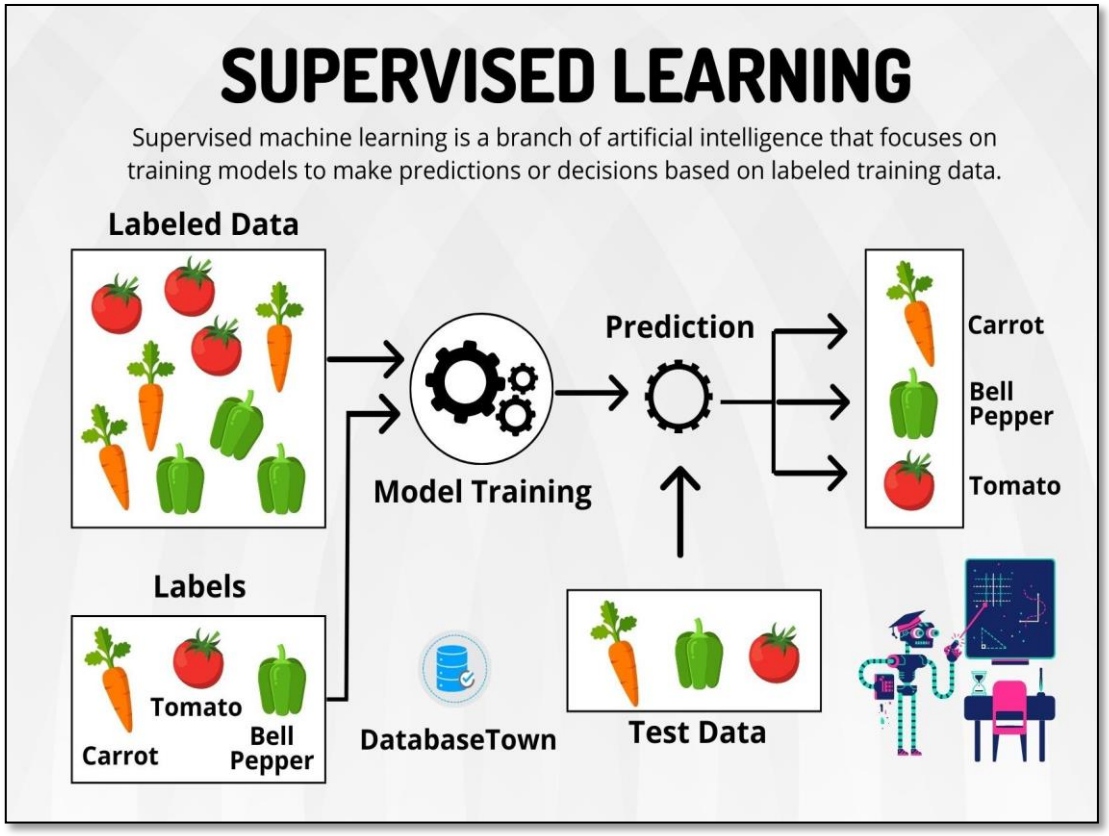


Data Science Application

1. Supervised learning algorithms

- Supervised learning is a machine learning method in which models are trained using labelled data. Supervised learning needs supervision to train the model, which is similar to as a student learns things in the presence of a teacher.

- **How it works:** This algorithm consists of a target/outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using this set of variables, we generate a function that maps input data to desired outputs. The training process continues until the model achieves the desired level of accuracy on the training data.

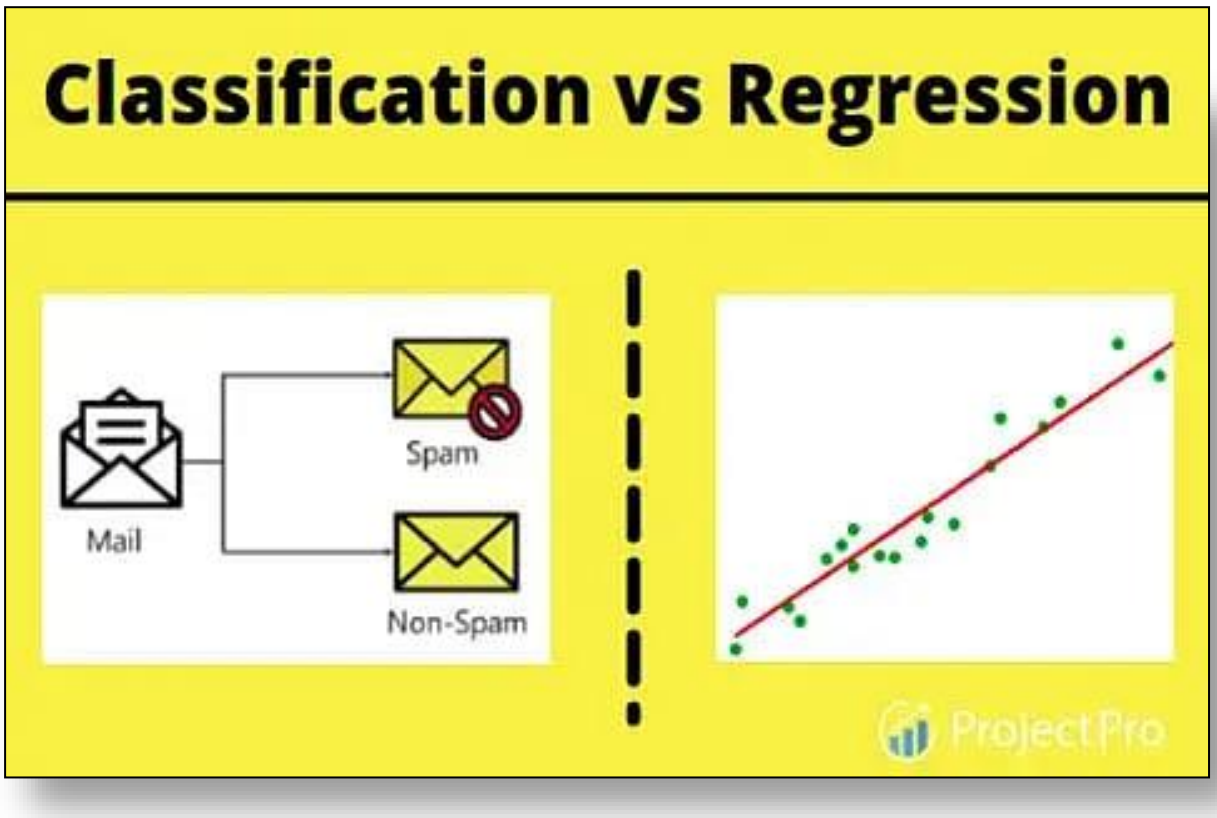


Data Science Application

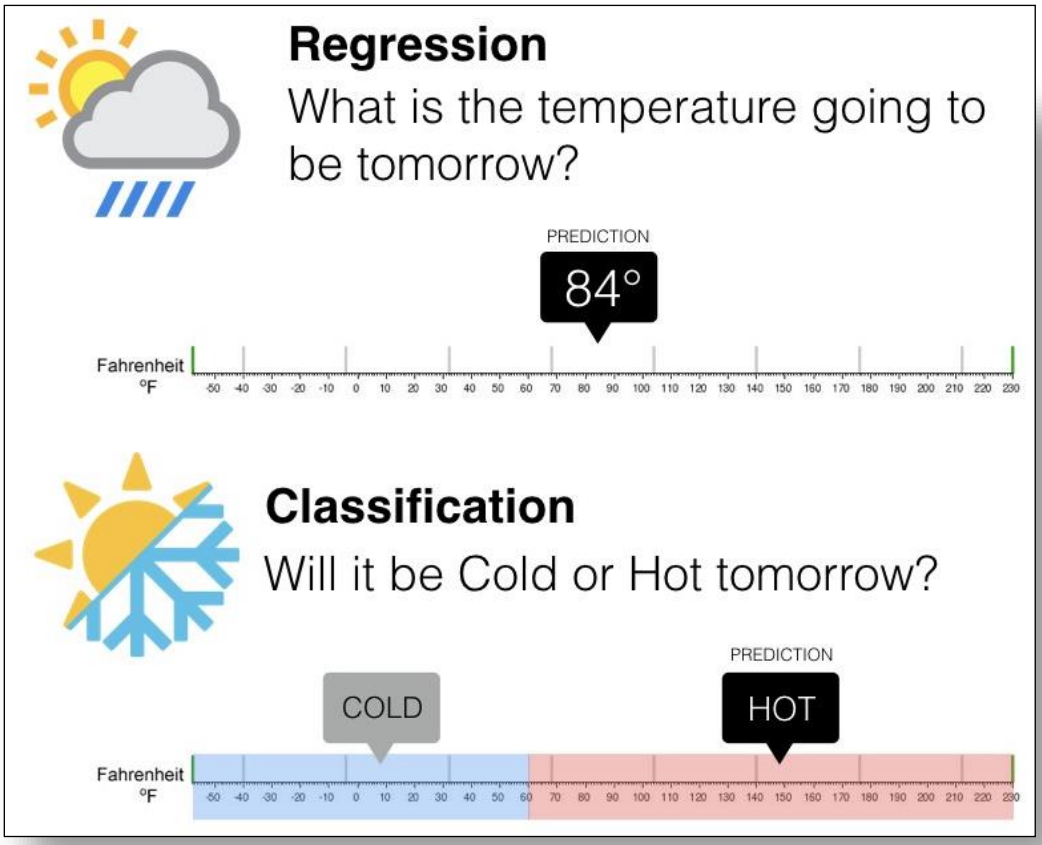
- **Supervised learning can be used for two types of problems:** Classification **and** Regression**.**

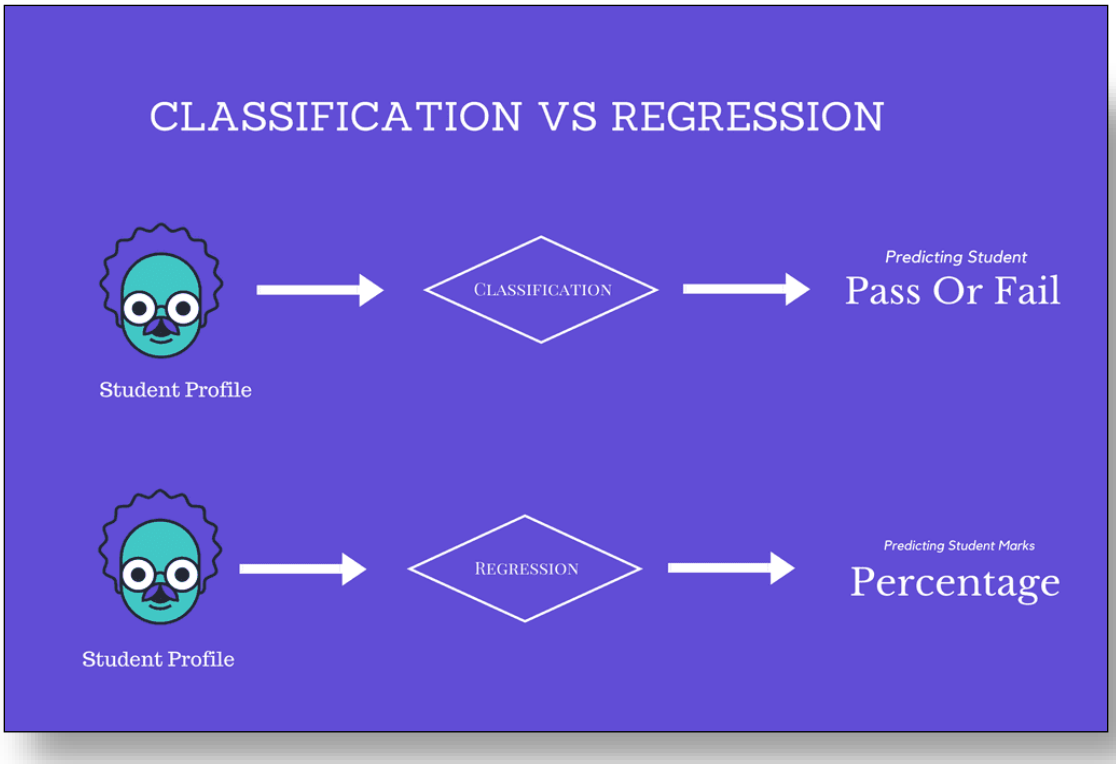
1. **Classification** uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labelled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, K-nearest neighbour and random forest, which are described in more detail below.

2. **Regression** is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.



Data Science Application



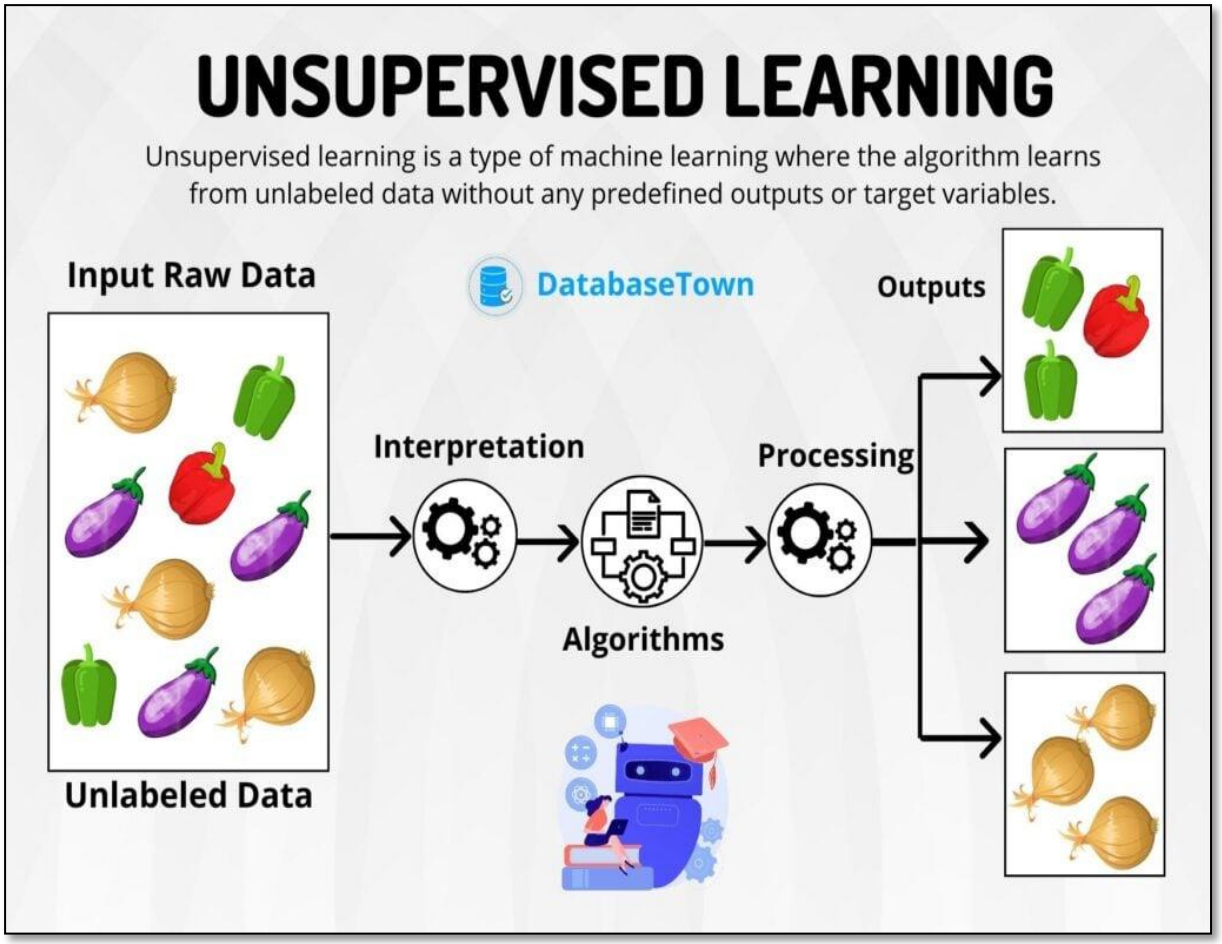


2. Unsupervised learning algorithms

Data Science Application

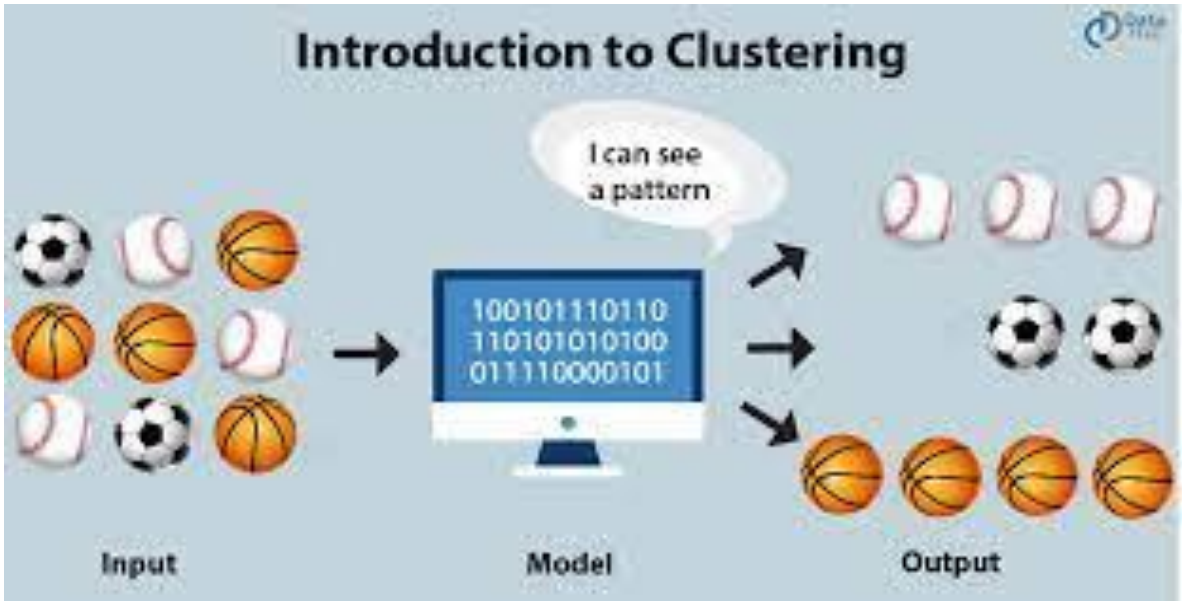
- unsupervised learning uses unlabelled data. From that data, the algorithm discovers patterns that help solve clustering or association problems. This is particularly useful when subject matter experts are unsure of common properties within a data set. Common clustering algorithms are hierarchical, K-means,

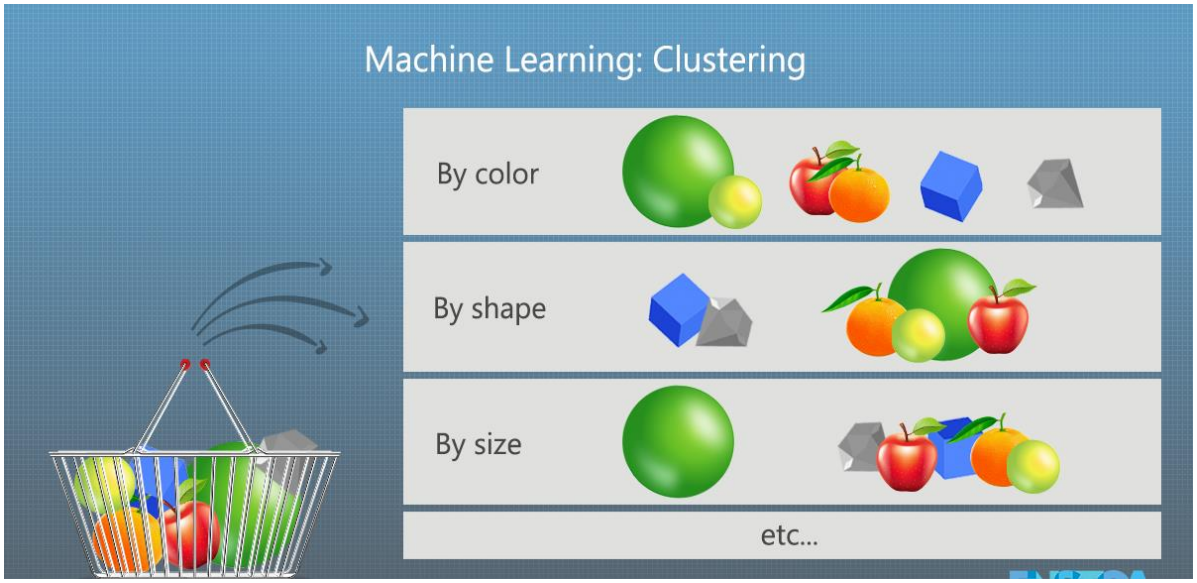
- **How it works:** In this algorithm, we do not have any target or outcome variable to predict / estimate (which is called unlabelled data). It is used for recommendation systems or clustering populations in different groups



Data Science Application

1. **Clustering**: These algorithms can identify patterns in data so that it can be grouped. Algorithms can help data scientists by identifying differences between data items that humans have overlooked.
2. **Hierarchical clustering**: This groups data into a tree of clusters[8]. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes these steps: 1) identify the two clusters which can be closest together, and 2) merge the two maximum comparable clusters. These steps continue until all the clusters are merged together.

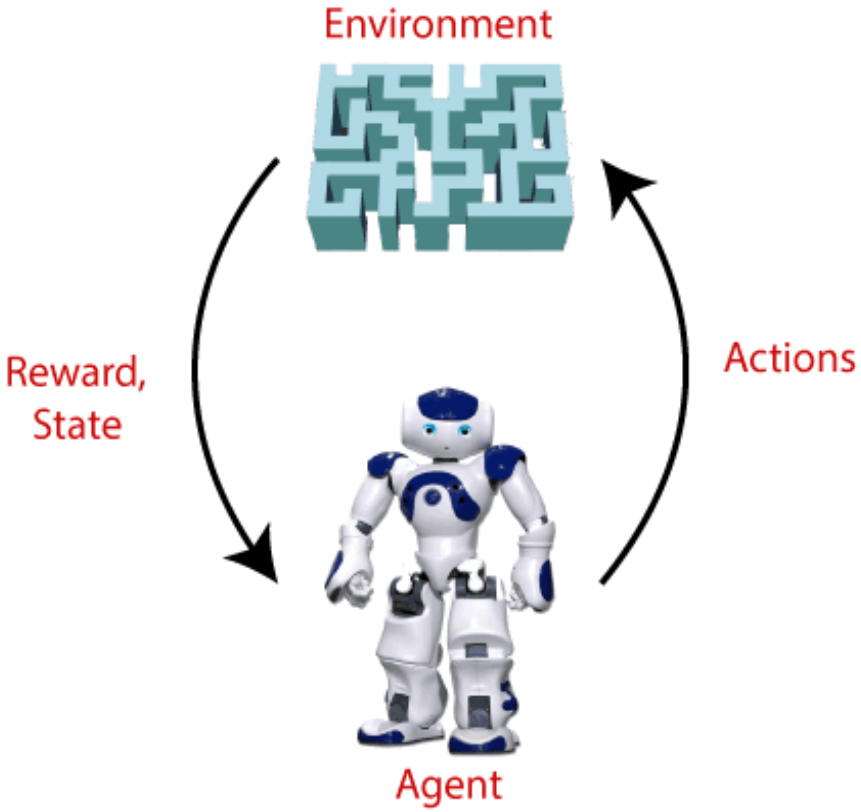


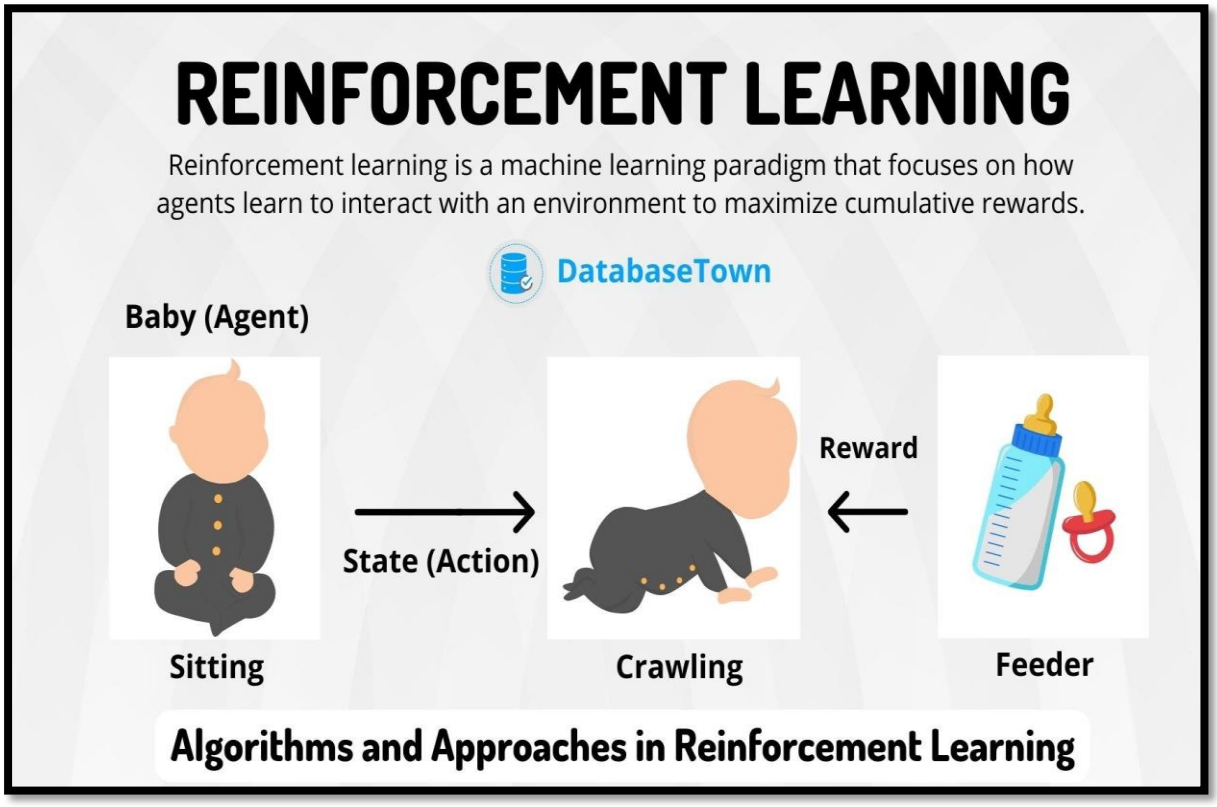


3. Reinforcement learning algorithms

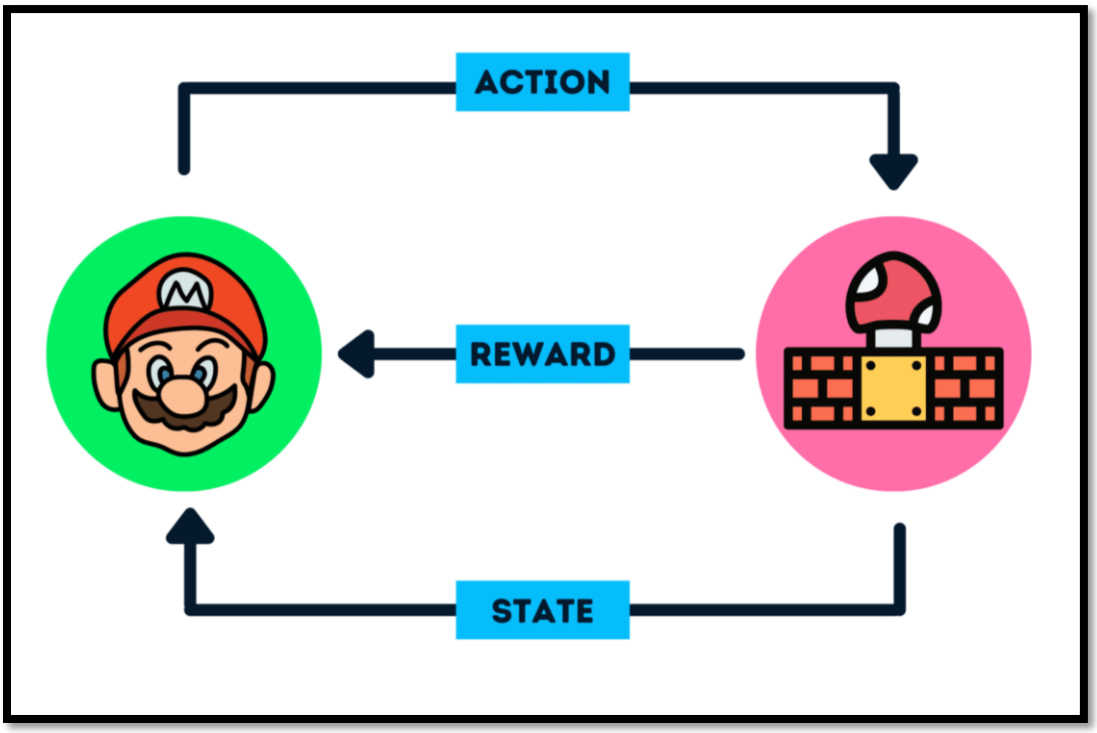
- Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.
- In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.
- Since there is no labeled data, so the agent is bound to learn by its experience only.
- RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as **game-playing, robotics**, etc.
- The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.
- The agent learns with the process of hit and trial, and based on the experience, it learns to perform the task in a better way. Hence, we can say that ***"Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that."*** How a Robotic dog learns the movement of his arms is an example of Reinforcement learning.
- It is a core part of Artificial intelligence, and all AI agent works on the concept of reinforcement learning. Here we do not need to pre-program the agent, as it learns from its own experience without any human intervention.
- **Example:** Suppose there is an AI agent present within a maze environment, and his goal is to find the diamond. The agent interacts with the environment by performing some actions, and based on those actions, the state of the agent gets changed, and it also receives a reward or penalty as feedback.
- The agent continues doing these three things (**take action, change state/remain in the same state, and get feedback**), and by doing these actions, he learns and explores the environment.
- The agent learns that what actions lead to positive feedback or rewards and what actions lead to negative feedback penalty. As a positive reward, the agent gets a positive point, and as a penalty, it gets a negative point.

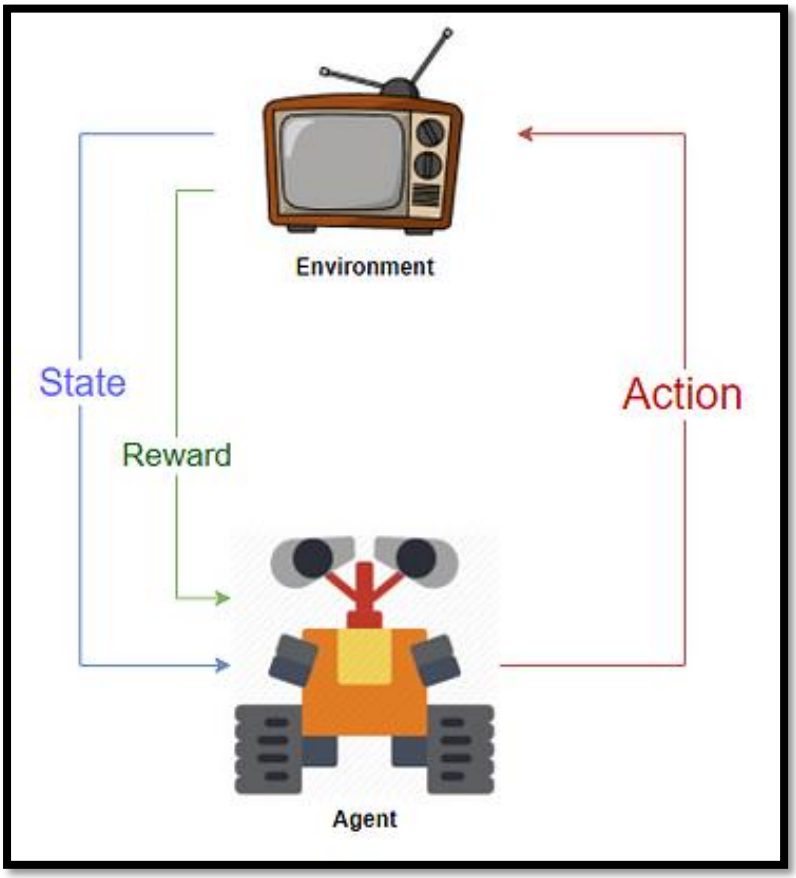
Data Science Application





Data Science Application

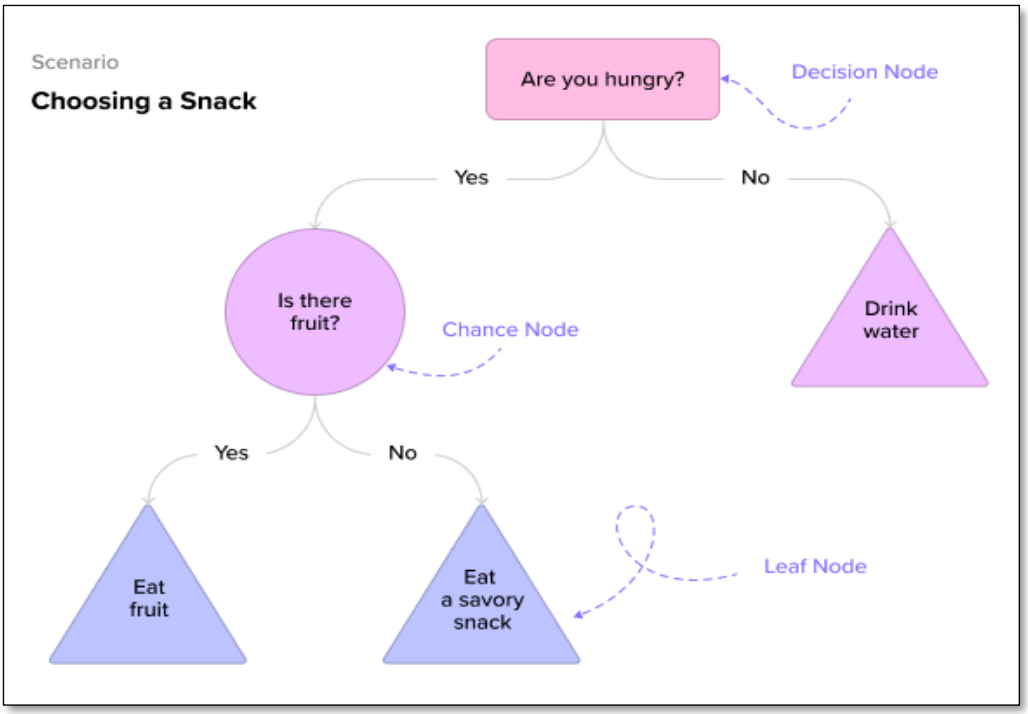


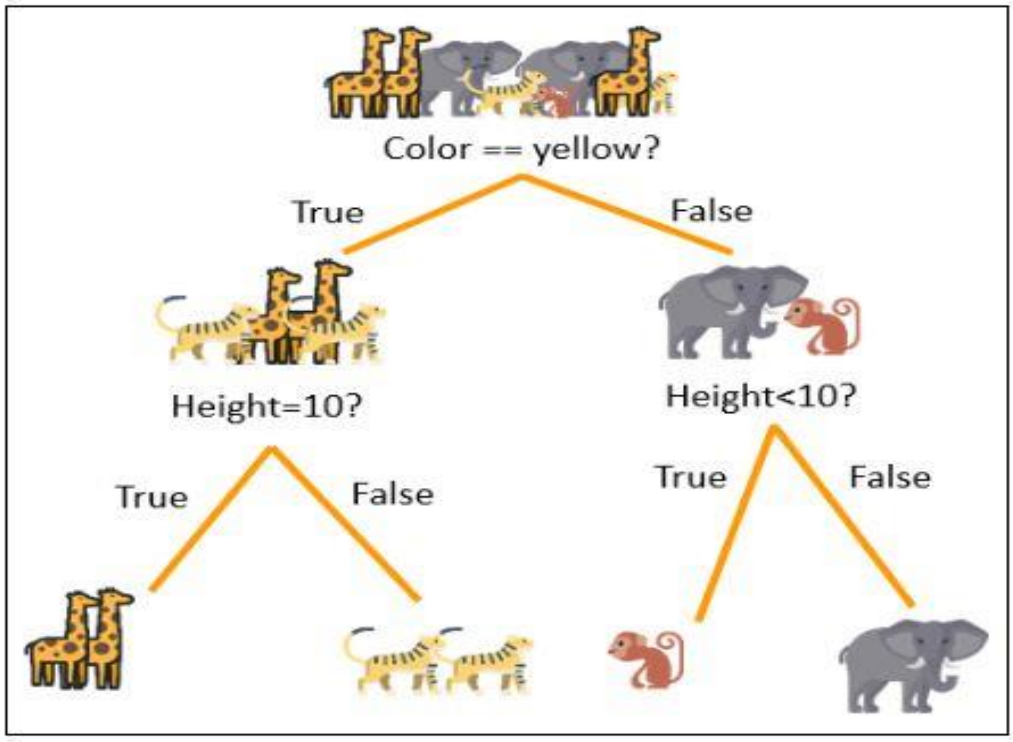


1. **Decision tree algorithms**

Data Science Application

Used for both predicting numerical values (regression problems) and classifying data into categories, decision trees use a branching sequence of linked decisions that may be represented with a tree diagram. One of the advantages of decision trees is that they are easy to validate and audit, unlike the black box of a neural network.

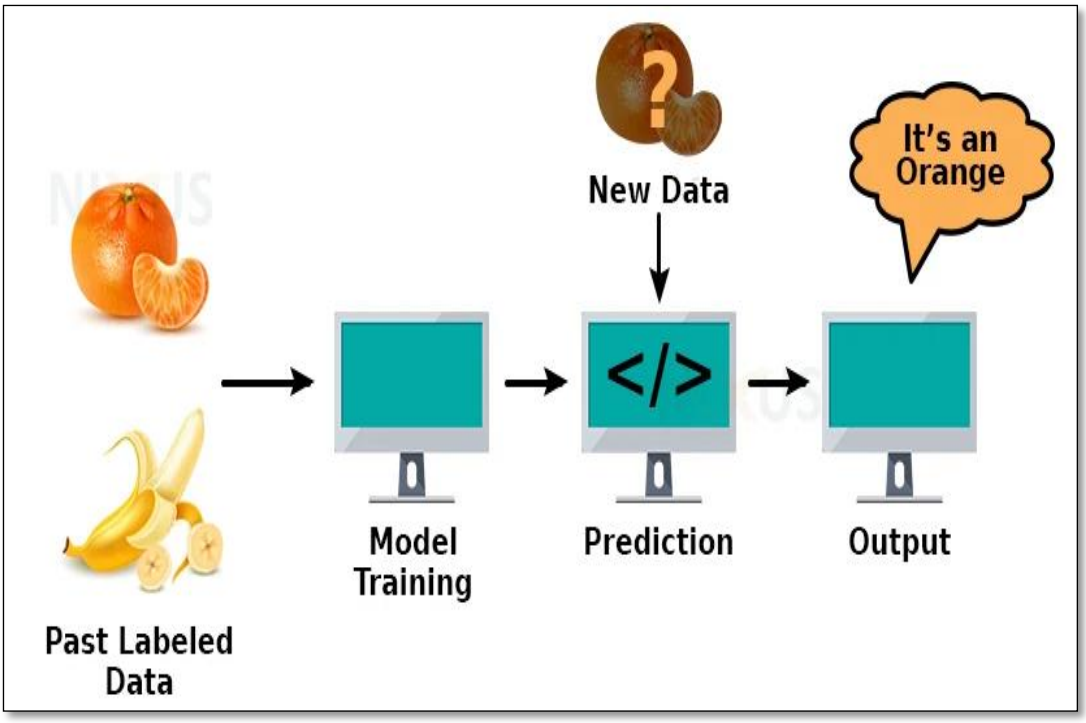


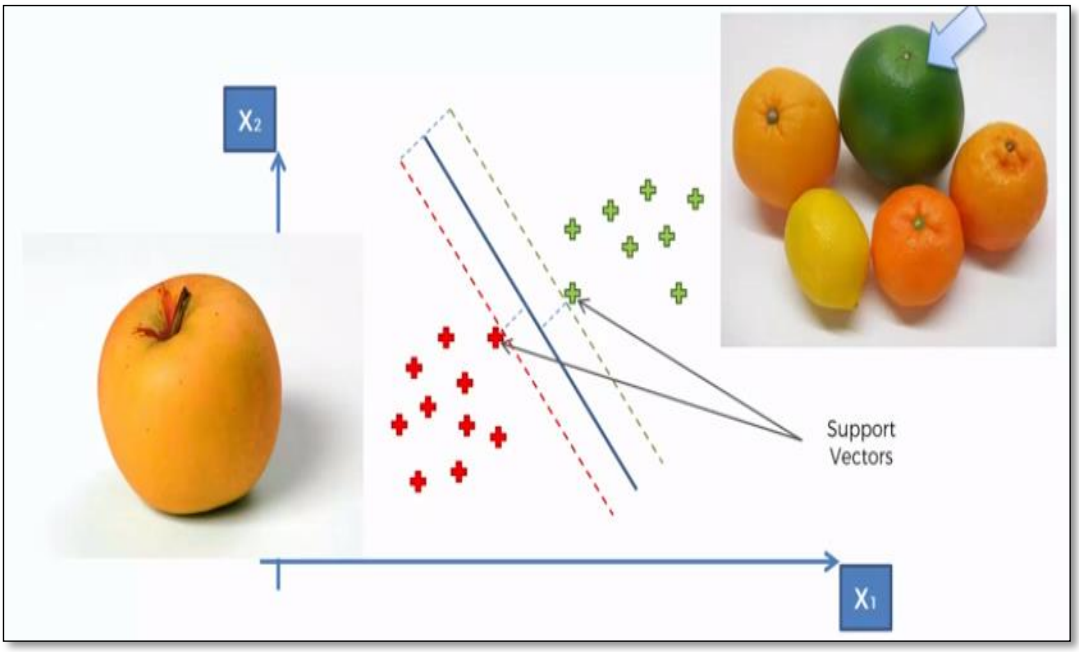


2. **SVM (Support vector machines)**
   SVMs are algorithms that can perform classification and regression tasks. It finds a hyperplane that best separates classes in feature space

Data Science Application

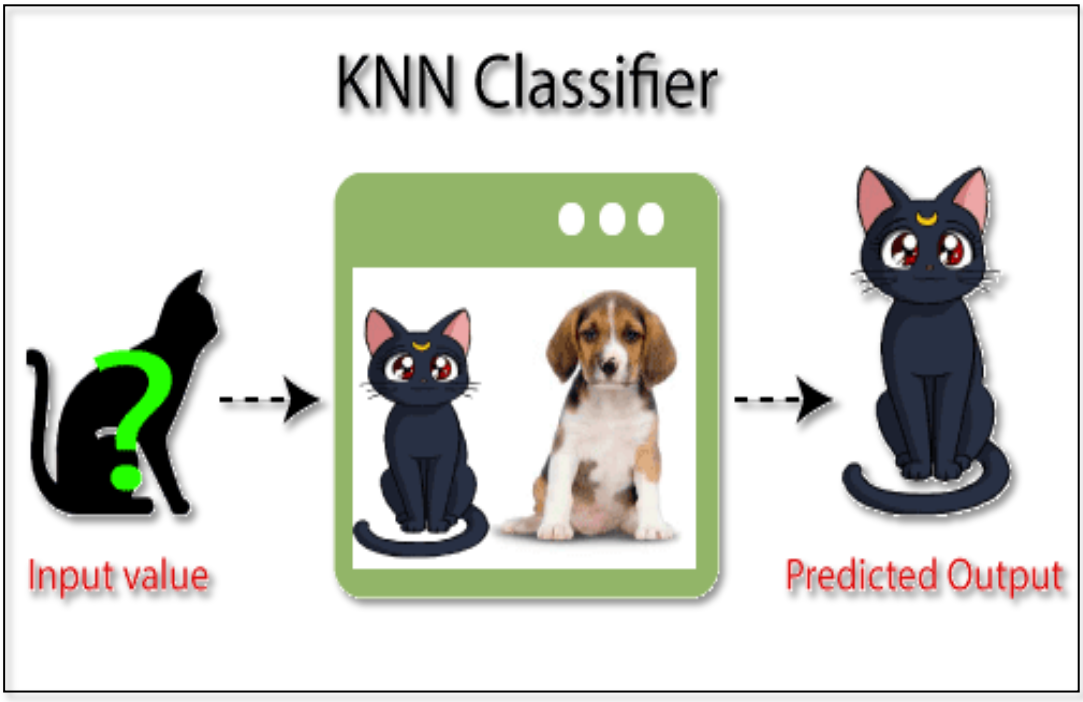
This algorithm may be used for both data classification and regression, but typically for classification problems, constructing a hyperplane where the distance between two classes of data points is at its maximum. This hyperplane is known as the decision boundary, separating the classes of data points (such as oranges vs. apples) on either side of the plane.

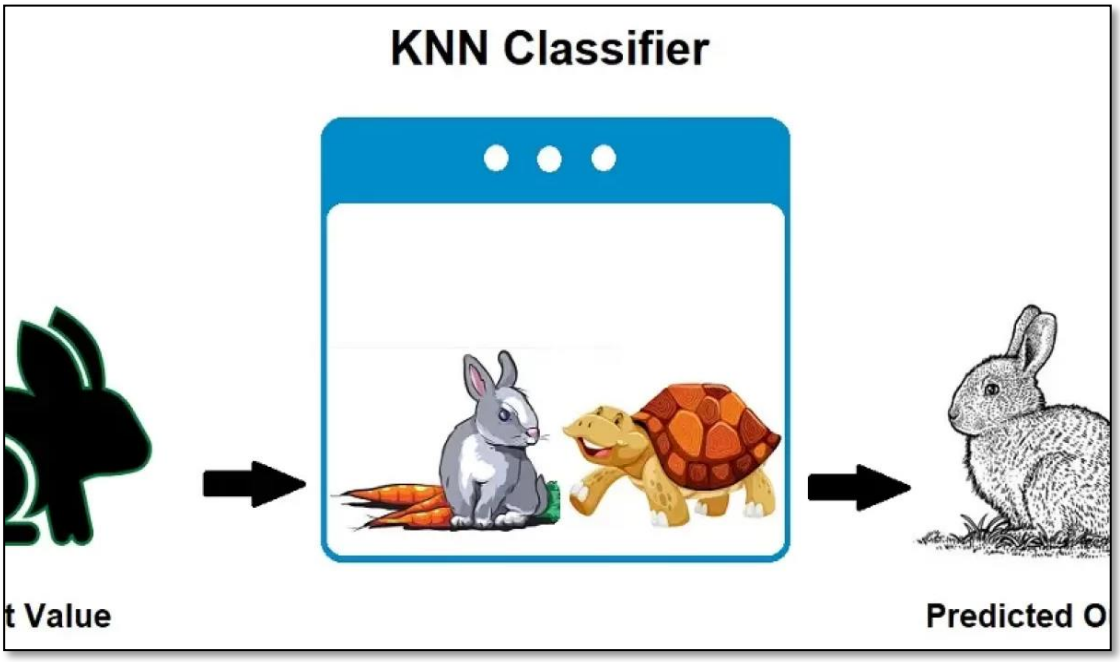




3. **KNN (K-nearest Neighbour)**
   KNN is a non-parametric technique that can be used for classification as well as regression. It works by identifying the k most similar data points to a new data point and then predicting the label of the new data point using the labels of those data points. It assumes that similar data points can be found near each other. As a result, it

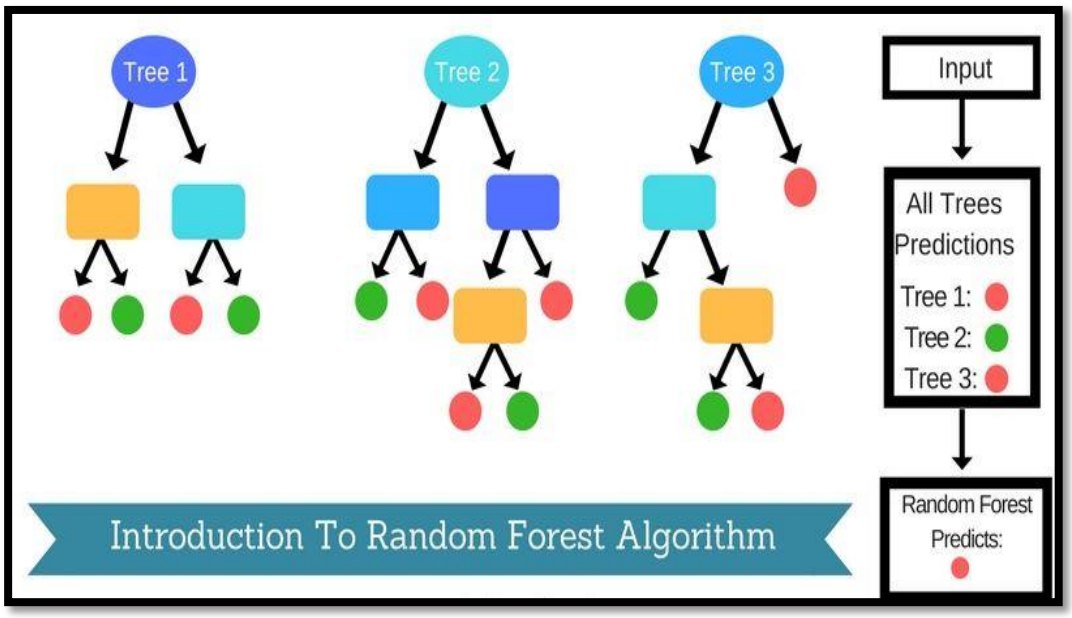
seeks to calculate the distance between data points, usually through Euclidean distance, and then it assigns a category based on the most frequent category or average.

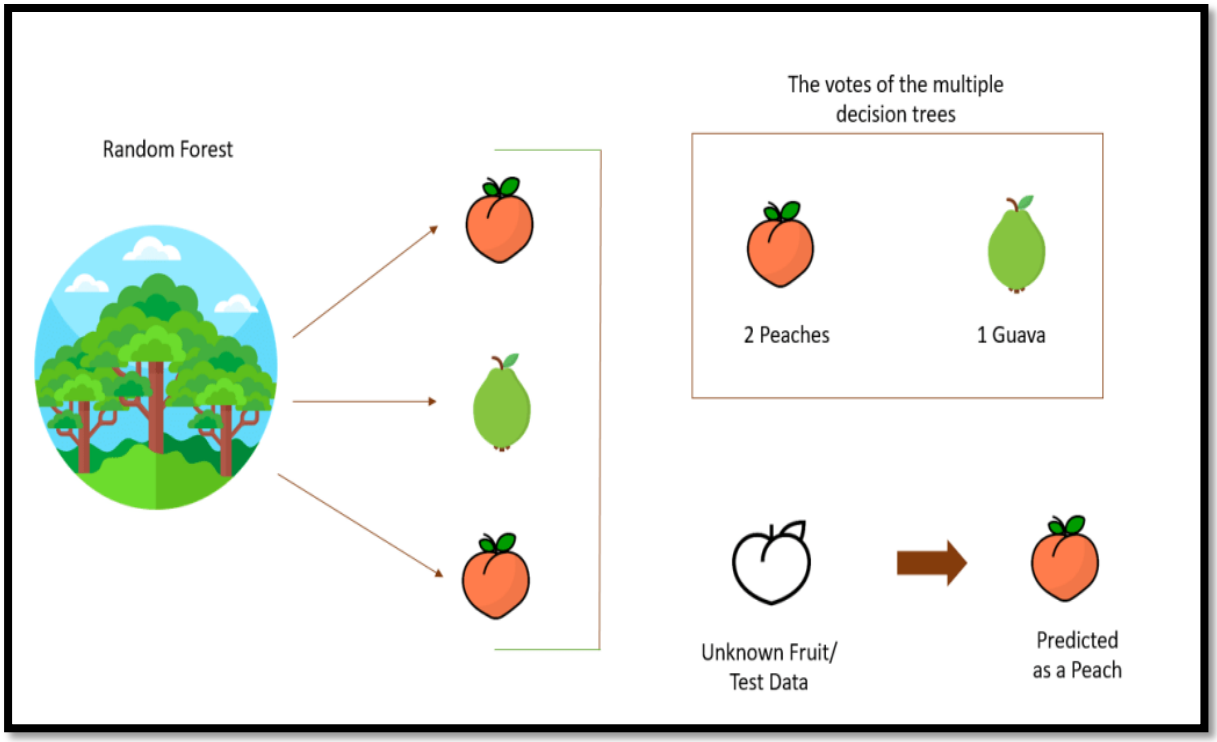




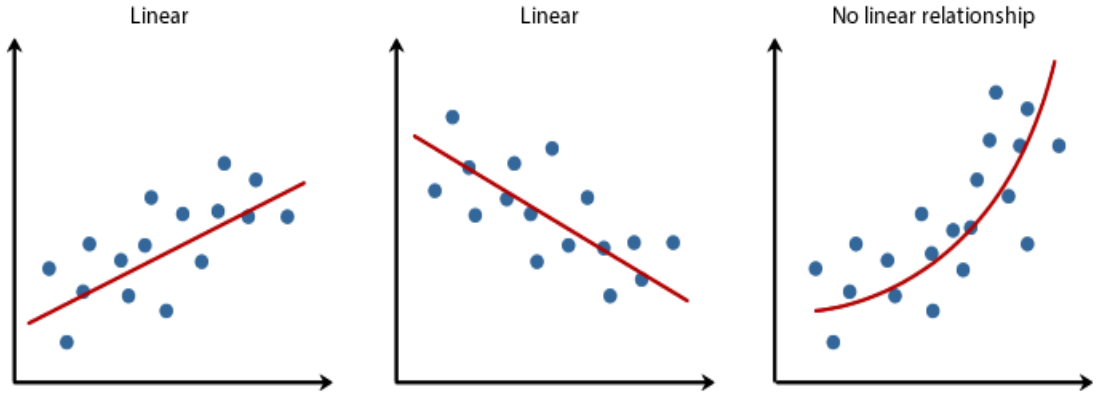
Data Science Application

4. **Random forests**
   In a random forest, the machine learning algorithm predicts a value or category by combining the results from a number of decision trees. The "forest" refers to uncorrelated decision trees, which are assembled to reduce variance and enable more accurate predictions.

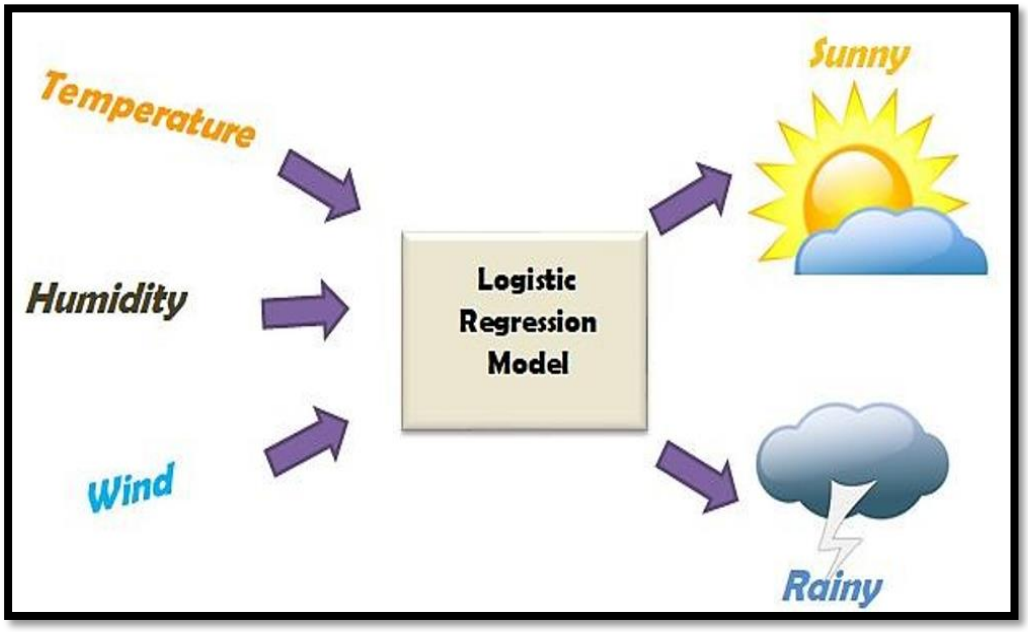


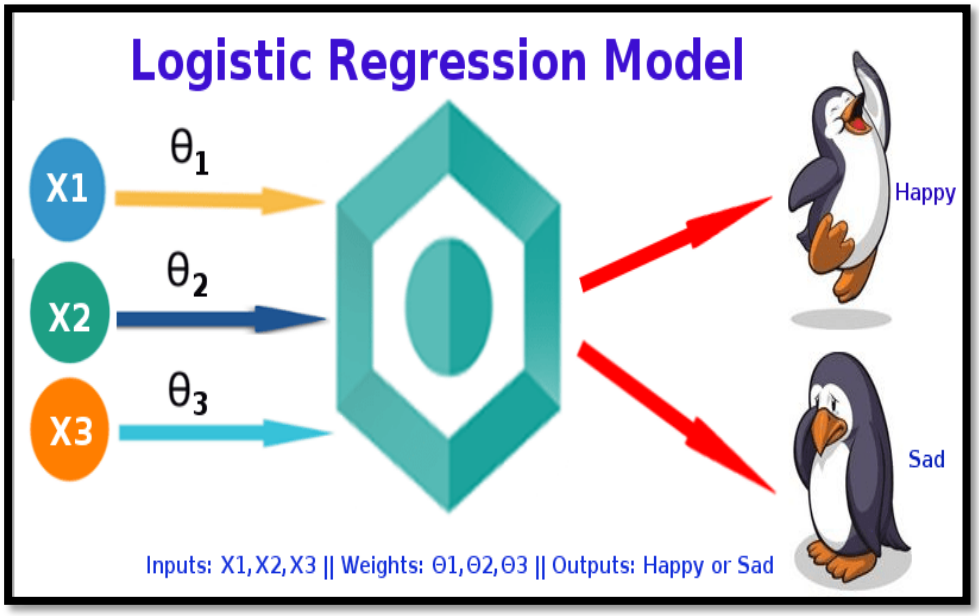


5. **Linear regression**:
   Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes. When there is only one independent variable and one dependent variable, it is known as simple linear regression.
   Linear regression is a simple algorithm used to map the linear relationship between input features and a continuous target variable. It works by fitting a line to the data and then using the line to predict new values.



6. **Logistic regression**:
   Logistic regression is an extension of linear regression that is used for classification tasks to estimate the likelihood that an instance belongs to a specific class.
   While linear regression is leveraged when dependent variables are continuous, logistic regression is selected when the dependent variable is categorical, meaning there are binary outputs, such as "true" and "false" or "yes" and "no." While both regression models seek to understand relationships between data inputs, logistic regression is mainly used to solve binary classification problems, such as spam identification.

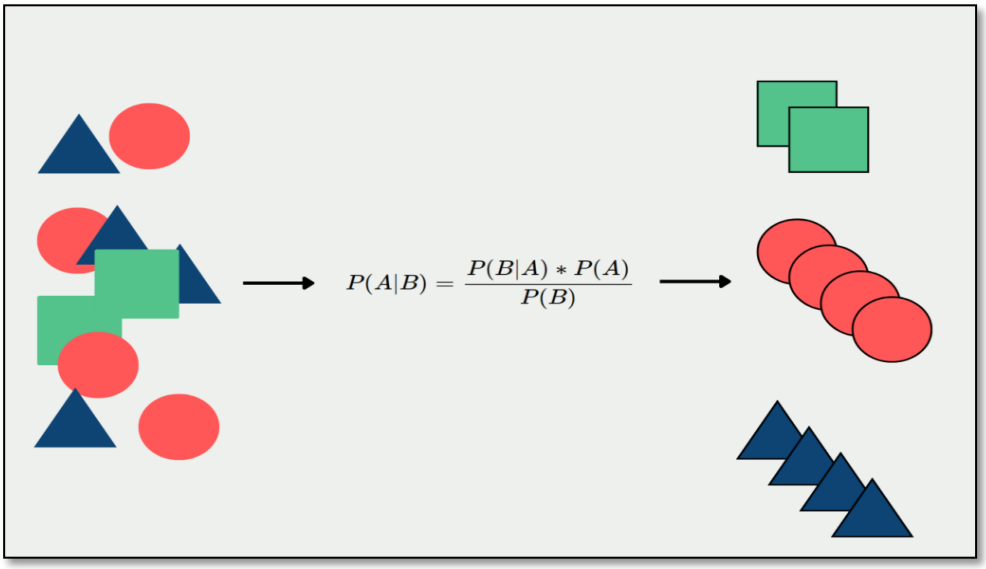


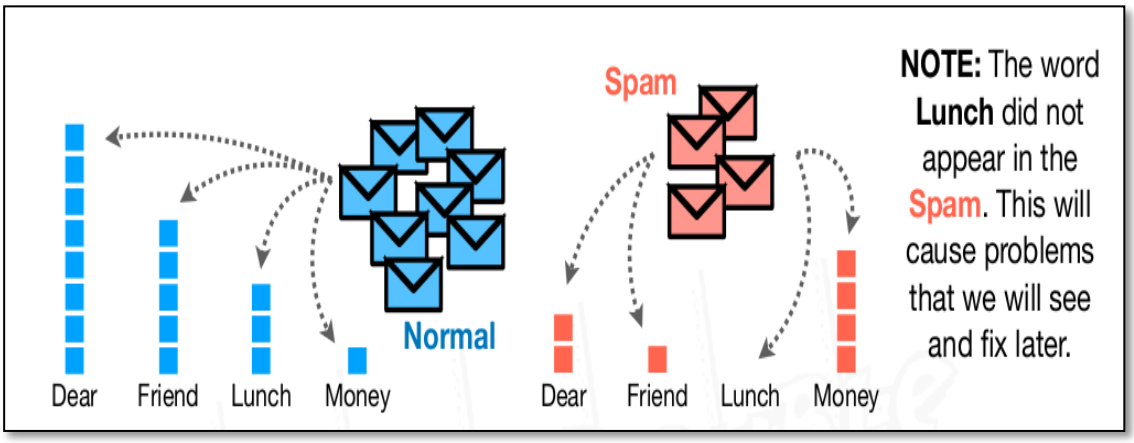


7. **Naïve Bayes**: This approach adopts the principle of class conditional independence from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result. There are three types of Naïve Bayes

Data Science Application

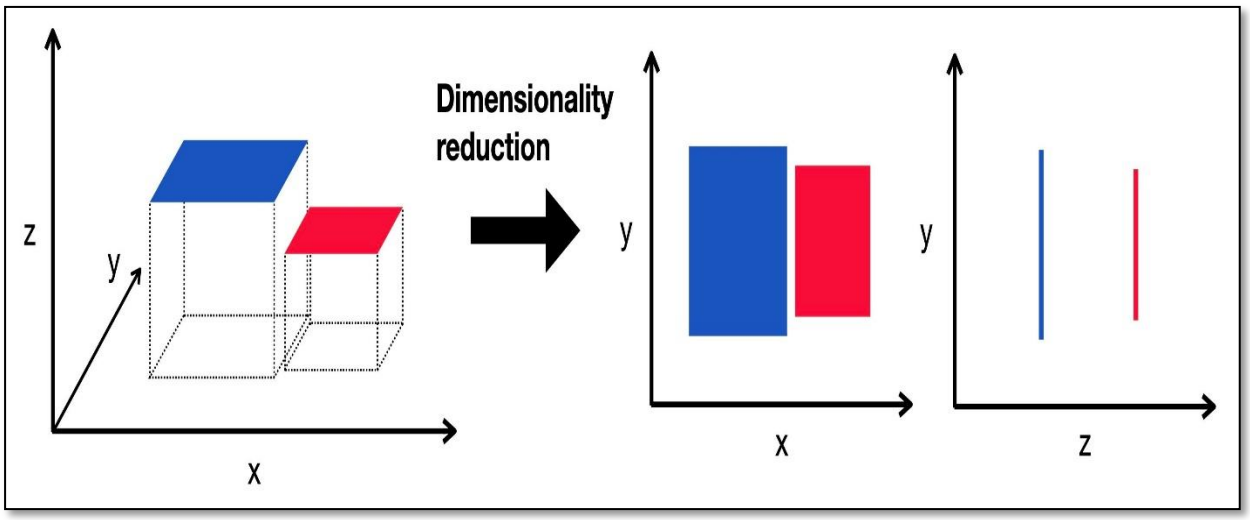
classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes and Gaussian Naïve Bayes. This technique is primarily used in text classification, spam identification and recommendation systems.

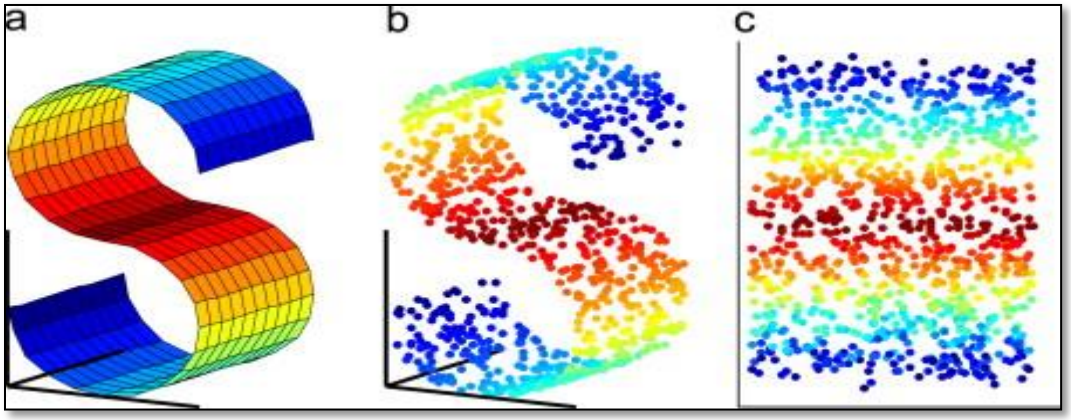




Data Science Application

8. **Dimensionality reduction**: When a selected data set has a high number of features[7] , it has high dimensionality. Dimensionality reduction then cuts down the number of features, leaving only the most meaningful insights or information. An example is principal component analysis.

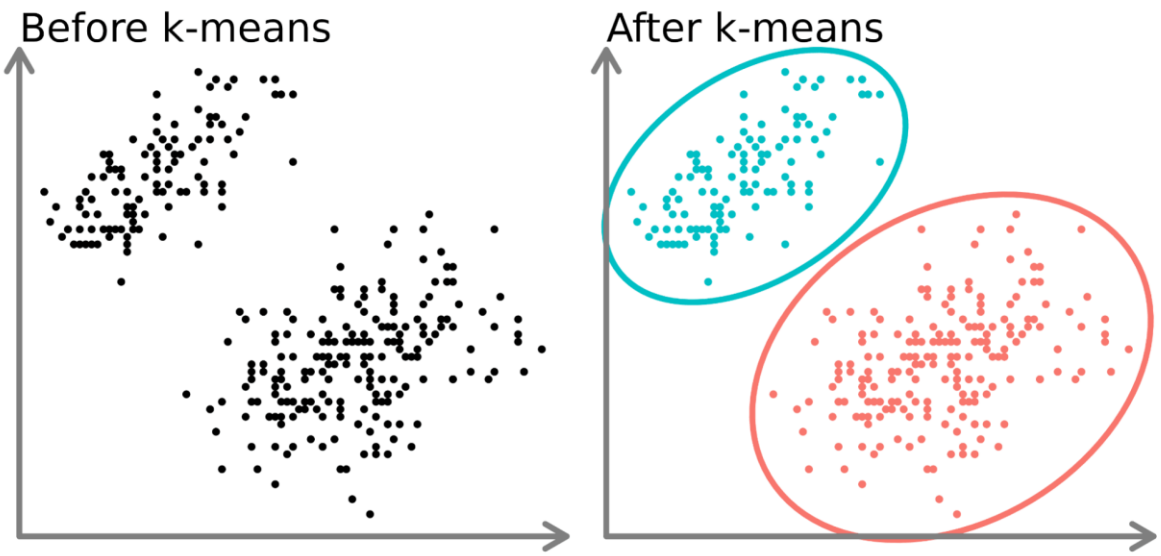


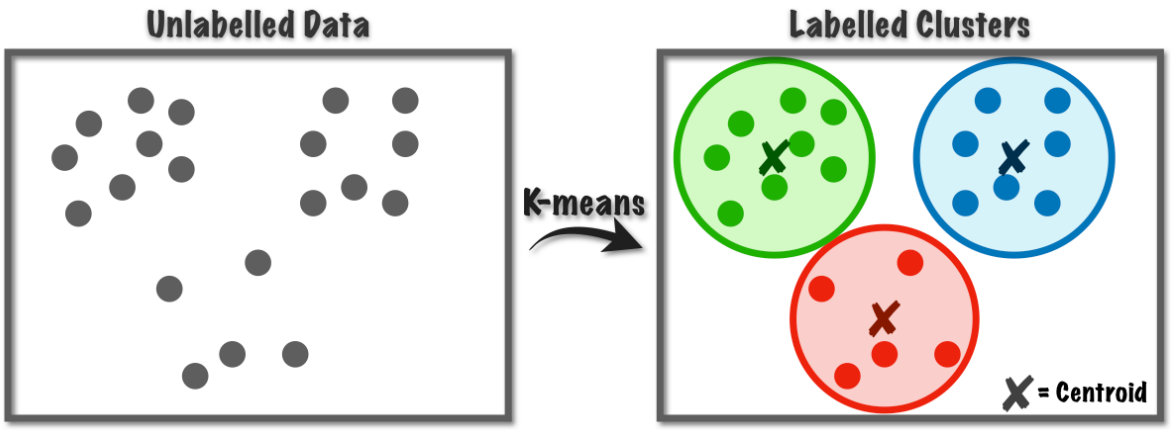


9. **AdaBoost or gradient boosting**:
   - Also called adaptive boosting[7], this technique boosts an underperforming regression algorithm by combining it with weaker ones to create a stronger algorithm that results in fewer errors. Boosting combines the forecasting power of several base estimators.
   - Machine learning is one of the most popular technologies to build predictive models for various complex regression and classification tasks. **Gradient Boosting Machine** (GBM) is considered one of the most powerful boosting algorithms.

10. **K-means clustering**: This identifies groups within data without labels[9] into different clusters by finding groups of data which are similar to one another. The name "K-

means" come from the k centroids that it uses to define clusters. A point is assigned to a particular cluster if it is closer to that cluster's centroid than any other centroid.





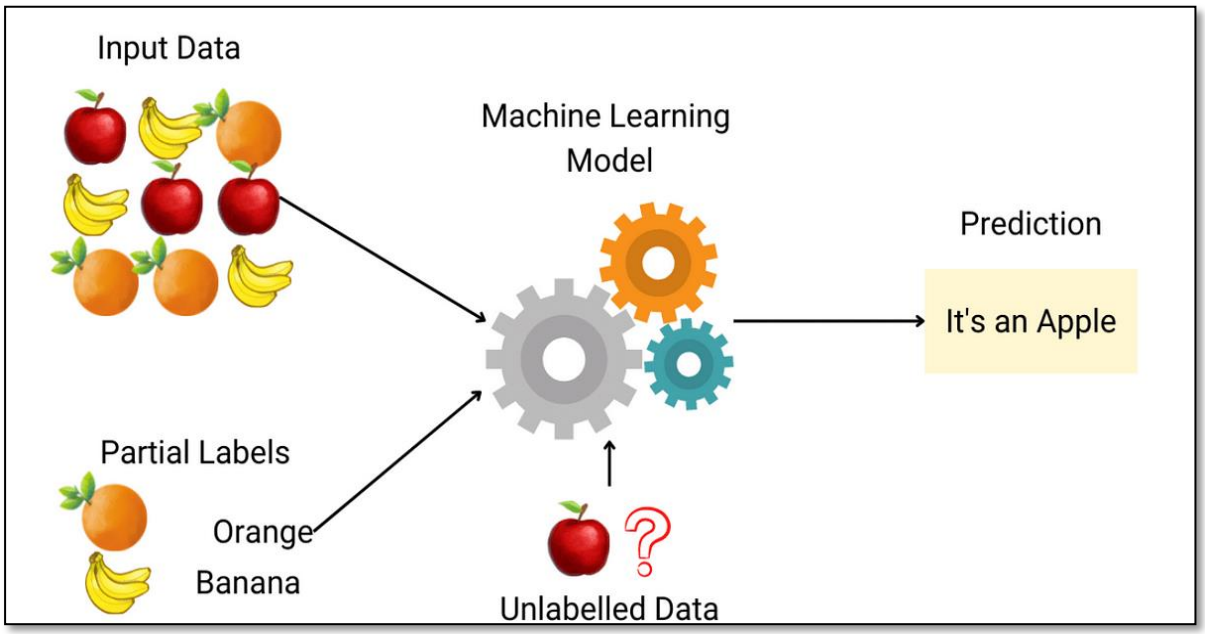
Data Science Application

- **Artificial neural networks**: Also known as ANNs, neural networks or simulated neural networks (SNNs), are a subset of machine learning techniques and are at the heart of deep learning algorithms. The learner algorithm recognizes patterns in input data using building blocks called neurons, approximating the neurons in the human brain, which are trained and modified over time. (More in "neural networks.")

- **Neural networks**: Primarily leveraged for deep learning algorithms, neural networks process the input training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (threshold) and an output. If that output value exceeds a given threshold, it "fires" or activates the node, passing data to the next layer in the network. Neural networks learn from adjustments based on the loss function through the process of gradient descent. When the cost function is at or near zero, you can be confident in the model's accuracy.

4. Semi supervised

- Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms. It uses the combination of labeled and unlabeled datasets during the training period.
- the basic disadvantage of supervised learning is that it requires hand-labeling by ML specialists or data scientists, and it also requires a high cost to process. Further unsupervised learning also has a limited spectrum for its applications. **To overcome these drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced**.



Data Science Application

How machine learning algorithms work

A paper from UC Berkeley breaks out the learning system of a machine learning algorithm into three main parts.

1. **A decision process**:

   In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabelled,
   your algorithm will produce an estimate about a pattern in the data.

2. **An error function**:

   An error function evaluates the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

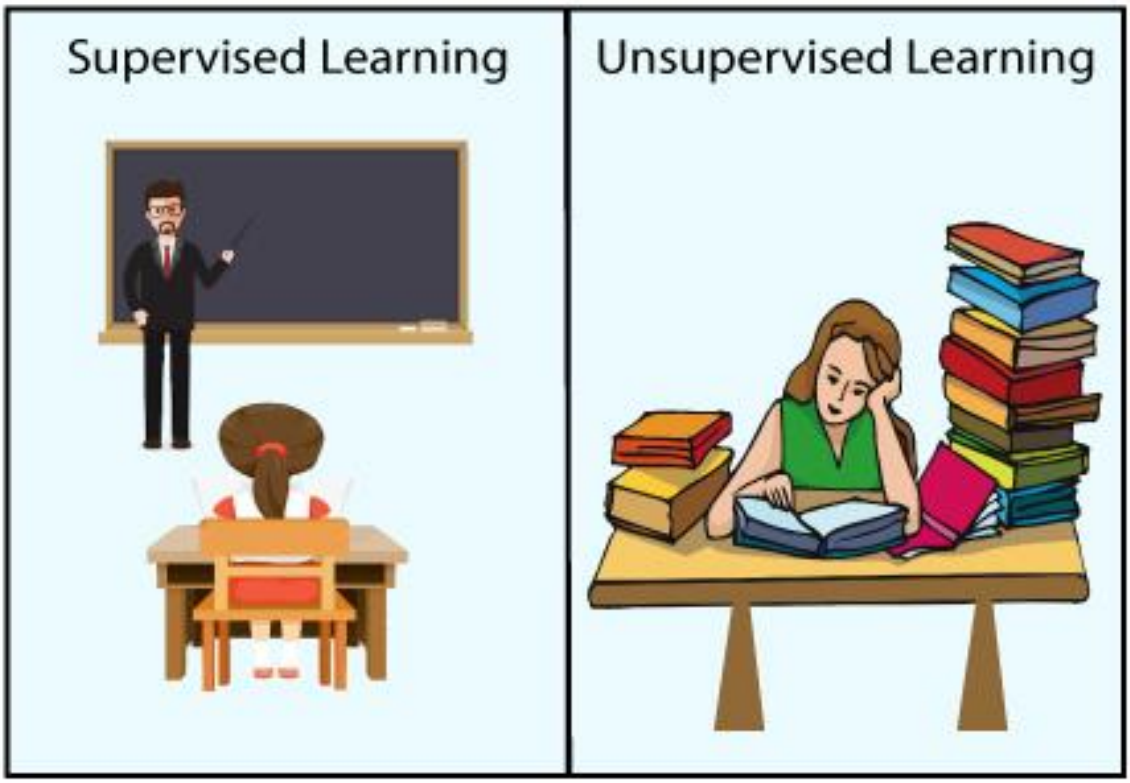
3. **A model optimization process**:

   If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this "evaluate and optimize" process, updating weights autonomously until a threshold of accuracy has been met.

Data Science Application

➢ Difference between Supervised and Unsupervised Learning

- Supervised and Unsupervised learning are the two techniques of machine learning. But both the techniques are used in different scenarios and with different datasets. Below the explanation of both learning methods along with their difference table is given.



Data Science Application

The main differences between Supervised and Unsupervised learning are given below:

Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
Supervised learning can be categorized in **Classification** and **Regression** problems.	Unsupervised Learning can be classified in **Clustering** and **Associations** problems.
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is closer to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.

Data Science Application

It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and