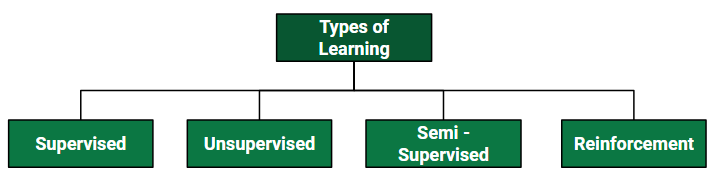
***MACHINE LEARNING***

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data.



Classification of Machine Learning

### A. Supervised learning:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. The given data is labeled. Both *classification*and *regression*problems are supervised learning problems.

* Example —  Consider the following data regarding patients entering a clinic . The data consists of the gender and age of the patients and each patient is labeled as “healthy” or “sick”.

|  |  |  |
| --- | --- | --- |
| Gender | Age | Label |
| M | 48 | sick |
| M | 67 | sick |
| F | 53 | healthy |
| M | 49 | sick |
| F | 32 | healthy |
| M | 34 | healthy |
| M | 21 | healthy |

**B. Unsupervised learning:**

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. In unsupervised learning algorithms, classification or categorization is not included in the observations. Example: Consider the following data regarding patients entering a clinic. The data consists of the gender and age of the patients.

|  |  |
| --- | --- |
| Gender | Age |
| M | 48 |
| M | 67 |
| F | 53 |
| M | 49 |
| F | 34 |
| M | 21 |

As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

### *To* ****C. Reinforcement learning:****

Reinforcement learning is the problem of getting an agent to act in the world so as to maximize its rewards.

A learner is not told what actions to take as in most forms of machine learning but instead must discover which actions yield the most reward by trying them. For example — Consider teaching a dog a new trick: we cannot tell him what to do, what not to do, but we can reward/punish it if it does the right/wrong thing.

 When watching the video, notice how the program is initially clumsy and unskilled but steadily improves with training until it becomes a champion.

### ****D. Semi-supervised learning:****

Where an incomplete training signal is given: a training set with some (often many) of the target outputs missing. There is a special case of this principle known as Transduction where the entire set of problem instances is known at learning time, except that part of the targets are missing. Semi-supervised learning is an approach to machine learning that combines small labeled data with a large amount of unlabeled data during training. Semi-supervised learning falls between unsupervised learning and supervised learning

**Categorizing based on Required Output**

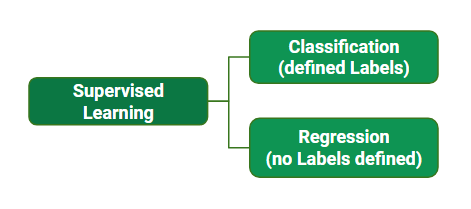
Another categorization of machine-learning tasks arises when one considers the desired output of a machine-learned system:

1. [**Classification**](https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/)**:** When inputs are divided into two or more classes, the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.
2. [**Regression**](https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/)**:** Which is also a supervised problem, A case when the outputs are continuous rather than discrete.
3. [**Clustering**](https://www.geeksforgeeks.org/clustering-in-machine-learning/)**:** When a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

***know***

### ****Types of Supervised Learning:****

***more about supervised and unsupe***

*KN**K*

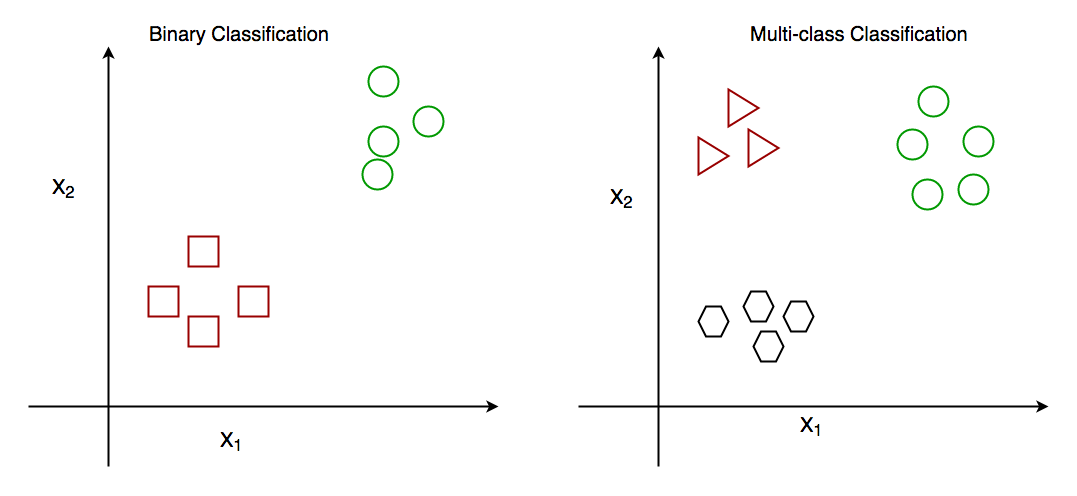
**Regression:** In regression, the target variable is a continuous value. The goal of regression is to predict the value of the target variable based on the input variables. Linear regression, polynomial regression, and decision trees are some of the examples of regression algorithms.

**Classification**: In classification, the target variable is a categorical value. The goal of classification is to predict the class or category of the target variable based on the input variables. Some examples of classification algorithms include logistic regression, decision trees, support vector machines, and neural networks.

## Classification Algorithms

[**Classification**](https://www.geeksforgeeks.org/getting-started-with-classification/) is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in the input and then the labels are predicted for the data.

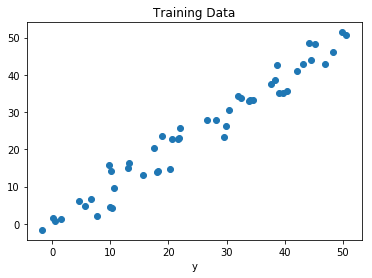
* In a classification task, we are supposed to predict discrete target variables(class labels) using independent features.
* In the classification task, we are supposed to find a [decision boundary](https://www.geeksforgeeks.org/ml-decision-function/) that can separate the different classes in the target variable.



## Regression Algorithms

[**Regression**](https://www.geeksforgeeks.org/ml-linear-regression/) is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions.

* In a regression task, we are supposed to predict a continuous target variable using independent features.
* In the regression tasks, we are faced with generally two types of problems linear and non-linear regression.



## Types of Regression Techniques

1. [Linear Regression](https://www.geeksforgeeks.org/ml-linear-regression/)

2.[Decision Tree Regression](https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/)

3.[Random Forest Regression](https://www.geeksforgeeks.org/random-forest-regression-in-python/)

4.[Support Vector Regression](https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/)

5. Logistic Regression

**Linear Regression**

Linear regression is used for predictive analysis. [Linear regression](https://www.geeksforgeeks.org/ml-linear-regression/) is a linear approach for modeling the relationship between the criterion or the scalar response and the multiple predictors or explanatory variables. Linear regression focuses on the conditional probability distribution of the response given the values of the predictors. For linear regression, there is a danger of [overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/). The formula for linear regression is:

*Syntax:*

***y = θx + b***

*where,*

* ***θ****– It is the model weights or parameters*
* ***b****– It is known as the bias.*

### Decision Tree Regression

A Decision Tree is the most powerful and popular tool for classification and prediction. A [Decision tree](https://www.geeksforgeeks.org/decision-tree/) is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. There is a non-parametric method used to model a decision tree to predict a continuous outcome.

**Decision Tree Terminologies**

#### Gini Impurity or index:

Gini Impurity is a score that evaluates how accurate a split is among the classified groups. The Gini Impurity evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes. In this case, we want to have a Gini index score as low as possible. Gini Index is the evaluation metric we shall use to evaluate our Decision Tree Model.

The Gini Impurity formula is: 1 – (p₁)² – (p₂)²,Here,

pi is the proportion of elements in the set that belongs to the ith category

#### Information Gain:

Information gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable. The higher the information gain, the more valuable the feature is in predicting the target variable.

### Random Forest Regression

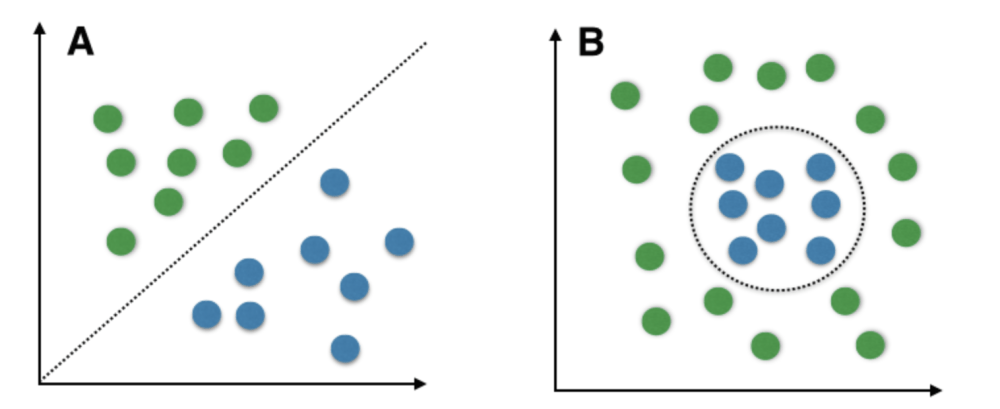
Random Forest is an [ensemble](https://www.geeksforgeeks.org/ensemble-methods-in-python/) technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as [bagging](https://www.geeksforgeeks.org/ml-bagging-classifier/). The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

[Random Forest](https://www.geeksforgeeks.org/random-forest-regression-in-python/) has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

### Support Vector Regression (SVR)

[Support vector regression (SVR)](https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/)is a type of [support vector machine (SVM)](https://www.geeksforgeeks.org/support-vector-machine-algorithm/) that is used for regression tasks. It tries to find a function that best predicts the continuous output value for a given input value.

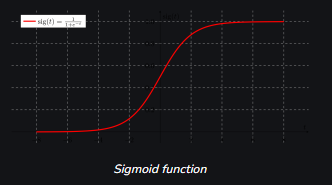
SVR can use both linear and non-linear kernels. A linear kernel is a simple dot product between two input vectors, while a non-linear kernel is a more complex function that can capture more intricate patterns in the data. The choice of kernel depends on the data’s characteristics and the task’s complexity.



# Logistic Regression

Logistic regression is a [supervised machine learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) algorithm mainly used for [classification](https://www.geeksforgeeks.org/getting-started-with-classification/) tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it’s referred to as regression because it takes the output of the [linear regression](https://www.geeksforgeeks.org/ml-linear-regression/)function as input and uses a sigmoid function to estimate the probability for the given class. The [difference between linear regression and logistic regression](https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/) is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

* Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.
* It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
* In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1).

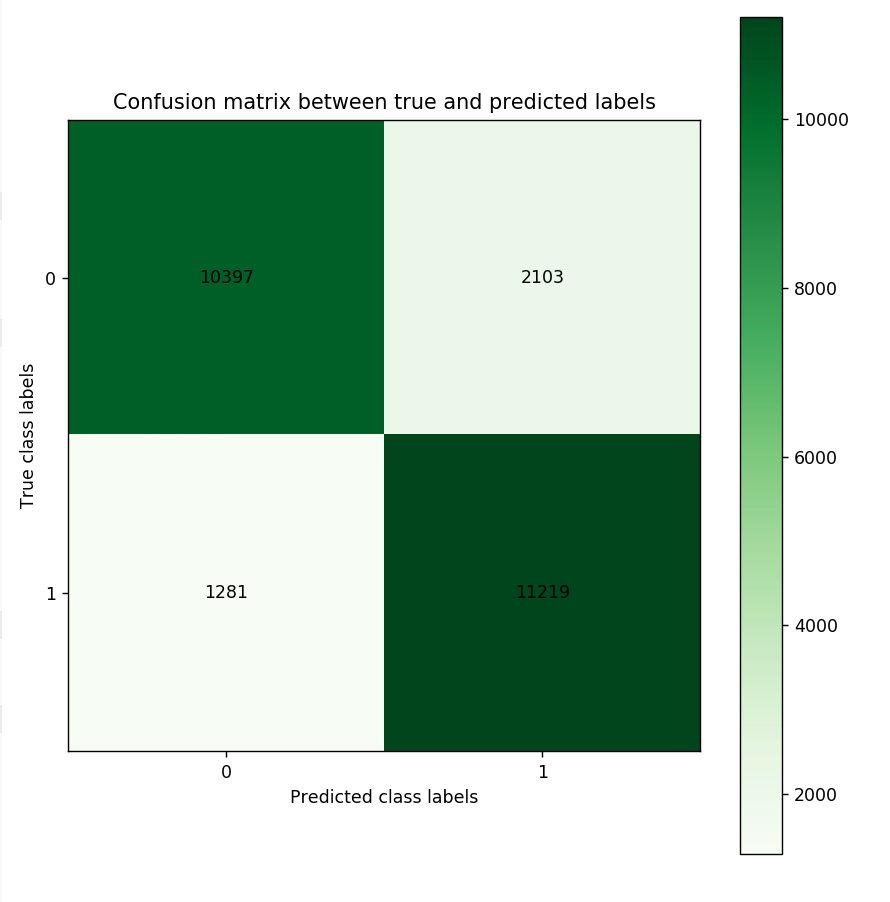


Confusion Matrix

Confusion Matrix is a matrix which is used to measure the accuracy of a classification model. It consists of variables such as:

1. TP(True Positive): The predicted value is true and the real value is positive.
2. FP(False Positive): The predicted value is false and the real value is positive.
3. TN(True Negative): The predicted value is true and the real value is negative.
4. FN(False Negative): The predicted value is false and the real value is negative.

The matrix looks like this



**Accuracy score**: How often the model’s predicted value was equal to actual value

Formula:

**Precision**: How often the model is right when predicted is positive

Formula:

**Recall/sensitivity/true-positive-rate**: When it is actually positive, how often the model predicted yes

Formula:

**False-Positive rate/fallout**: When it is actually positive, how often the model predicted no

Formula:

**Specificity/True-Negative rate:** When actual value is no and model predicts how much false from it

Formula:

**F-score**: The **F1** (or balanced f-score) score is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean) of the precision and recall. It thus symmetrically represents both precision and recall in one metric.

Formula:

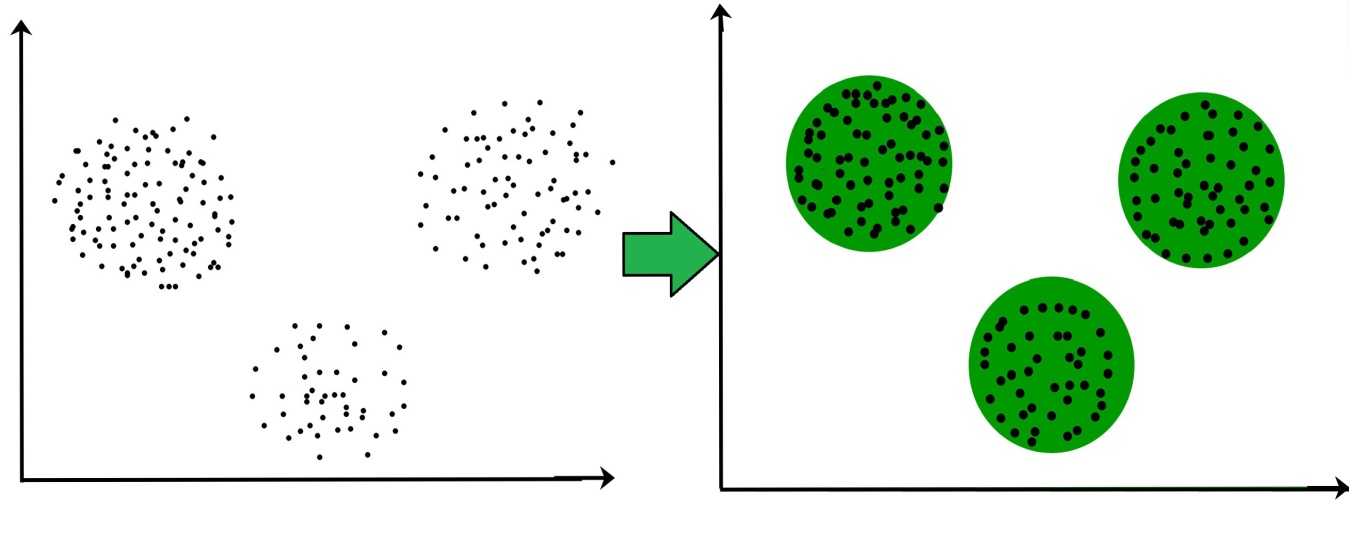
# K-Nearest Neighbor(KNN) Algorithm

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

# Clustering in Machine Learning

**Clustering** is the task of dividing the population or data points 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.

**For example** The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



# Association Rule Learning

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.

The association rule learning is one of the very important concepts of [machine learning](https://www.javatpoint.com/machine-learning), and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.** Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

| **TID** | **Items** |
| --- | --- |
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |
|  |  |

# K means Clustering

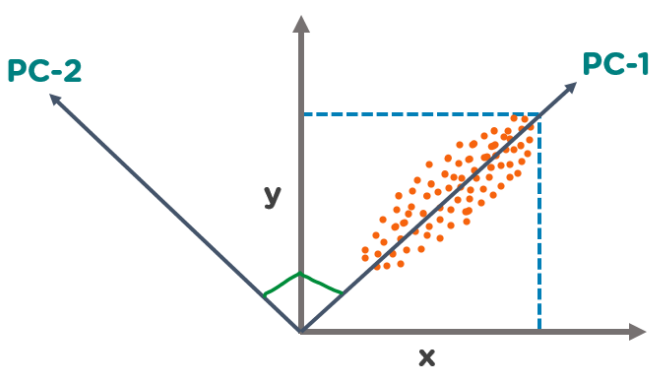
K-Means Clustering is an Unsupervised Machine Learning algorithm, which groups the unlabeled dataset into different clusters.

The goal of [clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/) is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another.

We are given a data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the K-means algorithm; an unsupervised learning algorithm. ‘K’ in the name of the algorithm represents the number of groups/clusters we want to classify our items into.

# Principal Component Analysis

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.

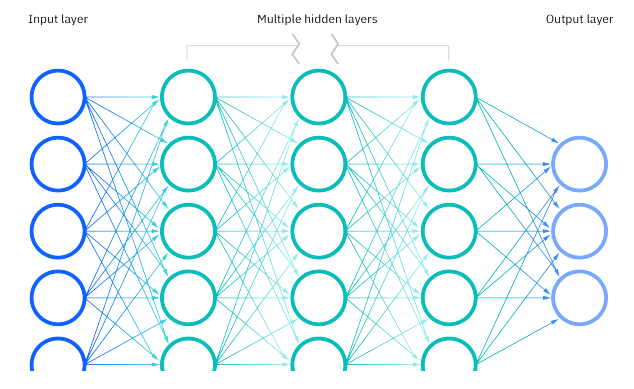


In the above figure, we have several points plotted on a 2-D plane. There are two principal components. PC1 is the primary principal component that explains the maximum variance in the data. PC2 is another principal component that is orthogonal to PC1.

Neural networks

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of [machine learning](https://www.ibm.com/topics/machine-learning) and are at the heart of [deep learning](https://www.ibm.com/topics/deep-learning) algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and [artificial intelligence](https://www.ibm.com/topics/artificial-intelligence), allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google’s search algorithm.



# Reinforcement learning

Reinforcement learning is an area of Machine Learning. It is 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. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

Reinforcement Learning (RL) is the science of decision making. It is about learning the optimal behavior in an environment to obtain maximum reward. In RL, the data is accumulated from machine learning systems that use a trial-and-error method. Data is not part of the input that we would find in supervised or unsupervised machine learning.

Reinforcement learning uses algorithms that learn from outcomes and decide which action to take next. After each action, the algorithm receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect. It is a good technique to use for automated systems that have to make a lot of small decisions without human guidance.