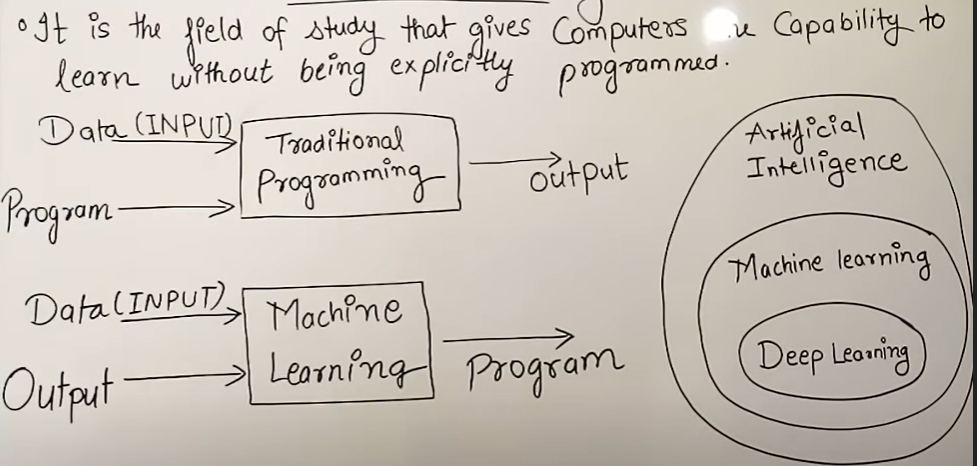
|  |  |
| --- | --- |
|  |  |

**Machine learning**



**Application of ML :-** manufacturing, healthcare, insurance, customer service, transportation, E-commerce, Automobile

Types of ML :- 1)supervised 2)unsupervised

3)reinforcement

1.Supervised :- **Supervised learning**is when the model is getting trained on a labelled dataset. A **labelled** dataset is one that has both input and output parameters. In this type of learning both training and validation, datasets are labelled

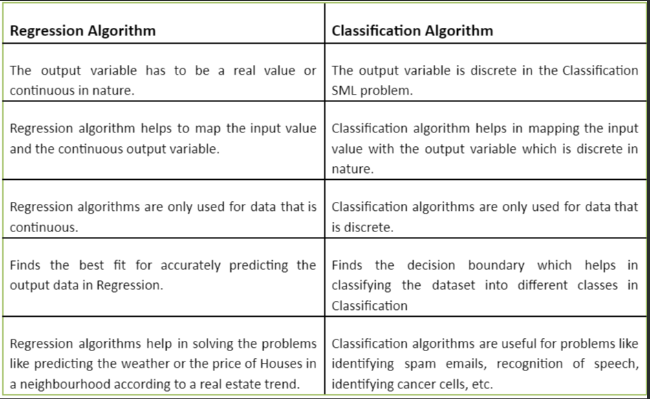
Supervised learning is further divided into sub categories:

*A) Classification B) Regression*

**A) Classification :-**It is a Supervised Learning task where output is having defined labels(discrete value)

**B) Regression :-** it is a Supervised Learning task where output is having continuous value.

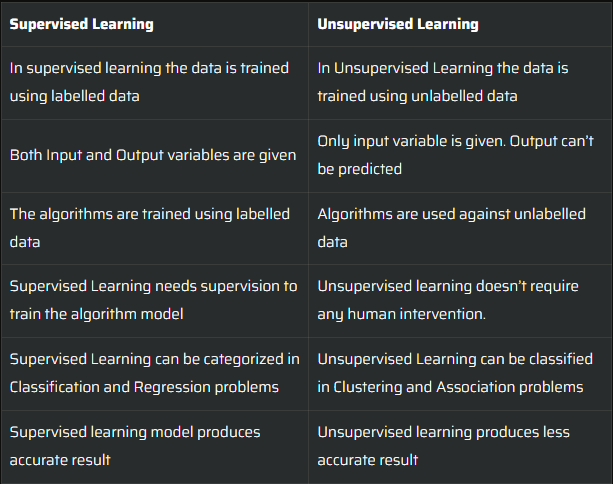
**Example of Supervised Learning Algorithms:**

* Linear Regression
* Logistic Regression
* Nearest Neighbour
* Gaussian Naive Bayes
* Decision Trees Etc. ****

**Difference between Regression and Classification**

2.Unsupervised learning:-

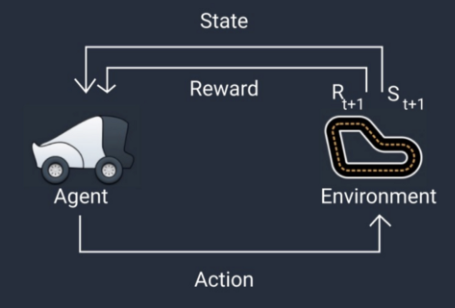
Unsupervised learning is the training of a machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabelled data by itself.

Application of unsupervised learning :-

* learn clusters/groups without any label
* customer segmentation (i.e. grouping)
* image compression
* bioinformatics

*methods to define no. of clusters in UL:*- 1)elbow methods,2) the optimiztn of the silhouette coefficnt etc.

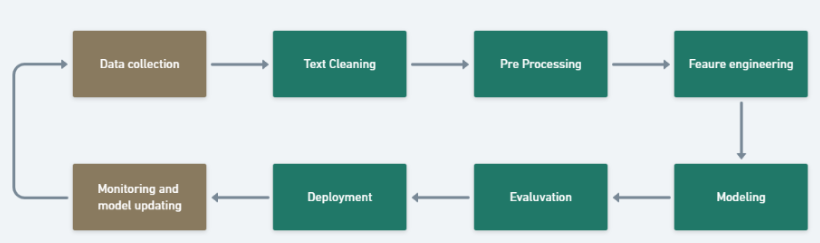
3.Reinforcement Learning:- In this technique, the model keeps on increasing its performance using Reward Feedback to learn the behaviour or pattern. These algorithms are specific to a particular problem e.g. Google Self Driving car, AlphaGo where a bot competes with humans and even itself to get better and better performers in Go Game. Each time we feed in data, they learn and add the data to their knowledge which is training data. So, the more it learns the better it gets trained and hence experienced.

**Q) what is NLP ?**

Natural language processing (NLP) uses machine learning to reveal the structure and meaning of text. With natural language processing applications, organizations can analyse text and extract information about people, places, and events to better understand social media sentiment and customer conversations.

**Q)what is ML pipeline in NLP ?**

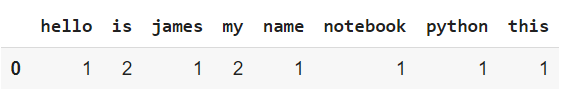
:-A machine learning pipeline is used to help automate machine learning workflows. They operate by enabling a sequence of data to be transformed and correlated together in a model that can be tested and evaluated to achieve an outcome, whether positive or negative.



**Countvectorizer :** Machines cannot understand characters and words. So when dealing with text data we need to represent it in numbers to be understood by the machine. Countvectorizer is a method to convert text to numerical data. To show you how it works let’s take an example:

text = [‘Hello my name is james, this is my python notebook’]

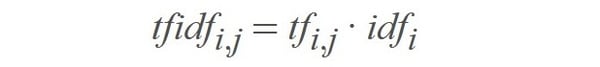
The text is transformed to a sparse matrix as shown below.



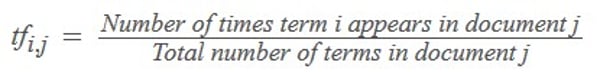
Q) explain TFIDF?

**TF-IDF** stands for “Term Frequency – Inverse Document Frequency.” It reflects how important a word is to a document in a collection or corpus. This technique is often used in information retrieval and text mining as a weighing factor.

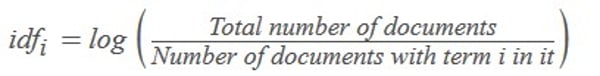
TF-IDF is composed of two terms:



* Term Frequency (TF):  
  The number of times a word appears in a document divided by the total number of words in that document.



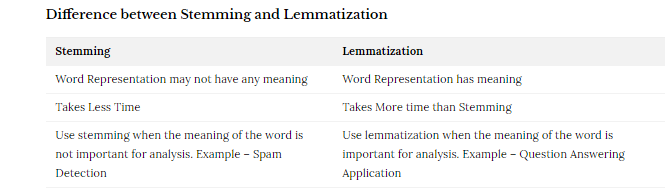
* Inverse Document Frequency (IDF):  
  The logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.



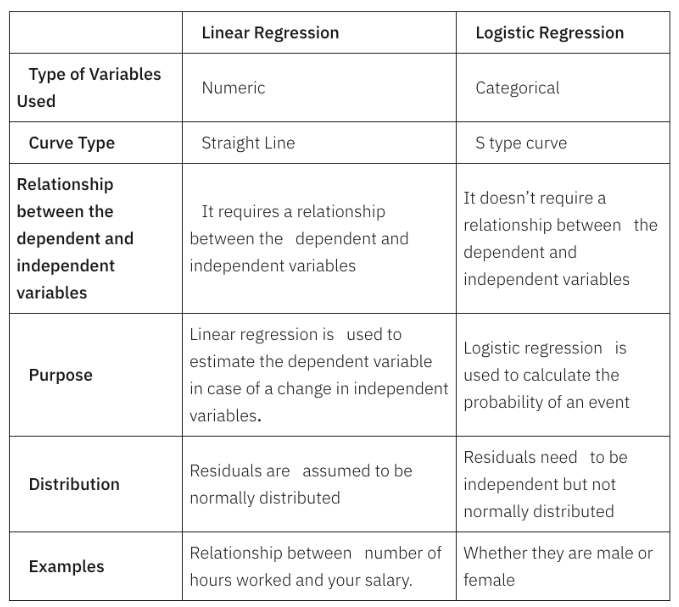
So, essentially, the TF-IDF value increases as the word’s frequency in a document (TF) increases. However, this is offset by the number of times the word appears in the entire collection of documents or corpus (IDF).

**Q) what is histogram?**

A **histogram** is a graphical representation of a grouped frequency distribution with continuous classes. It is an area diagram and can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes. In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries. The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.



Q. Difference between Linear Regression and Logistic Regression



What is Data?

Data is nothing but facts and statistics stored or free flowing over a network, generally it's raw and unprocessed. For example: When you visit any website, they might store you IP address, that is data, in return they might add a cookie in your browser, marking you that you visited the website, that is data, your name, it's data, your age, it's data.

Data becomes information when it is processed.

**Types of data :**

Categorical , discrete and continuous

Here's a clear explanation of the difference between discrete and continuous data, along with examples:

**Discrete data** consists of distinct, countable values that can't be subdivided into smaller units. It often involves integers (whole numbers) and represents items that can be counted.

Examples of discrete data:

* Number of students in a class (e.g., 25)
* Number of cars sold in a month (e.g., 120)
* Shoe size (e.g., 8, 9, 10)
* Number of pets owned (e.g., 0, 1, 2, 3)
* Number of goals scored in a game (e.g., 3)
* Number of items in a shopping cart (e.g., 5)

**Continuous data,** on the other hand, can take on any value within a given range, even values that include decimals or fractions. It represents measurements or quantities that can be divided into smaller and smaller units.

Examples of continuous data:

* Height (e.g., 1.75 meters)
* Weight (e.g., 65.3 kilograms)
* Time (e.g., 2.5 hours)
* Temperature (e.g., 23.7 degrees Celsius)
* Distance (e.g., 100.2 kilometers)
* Speed (e.g., 60.5 kilometers per hour)
* Age (e.g., 25.8 years)

Key differences:

* Values: Discrete data has distinct, separate values, while continuous data can take on any value within a range.
* Countability: Discrete data can be counted, while continuous data cannot be counted but can be measured.
* Measurement: Discrete data is often measured using whole numbers, while continuous data is often measured using decimals or fractions.
* Visualization: Discrete data is often visualized using bar charts or pie charts, while continuous data is often visualized using line charts, scatter plots, or histograms.

