- **Concepts of Supervised, Unsupervised, Semi-Supervised Learning**

Learning is the process of converting experience into expertise or knowledge.

Learning can be broadly classified into three categories, as mentioned below, based on the nature of the learning data and interaction between the learner and the environment.

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning

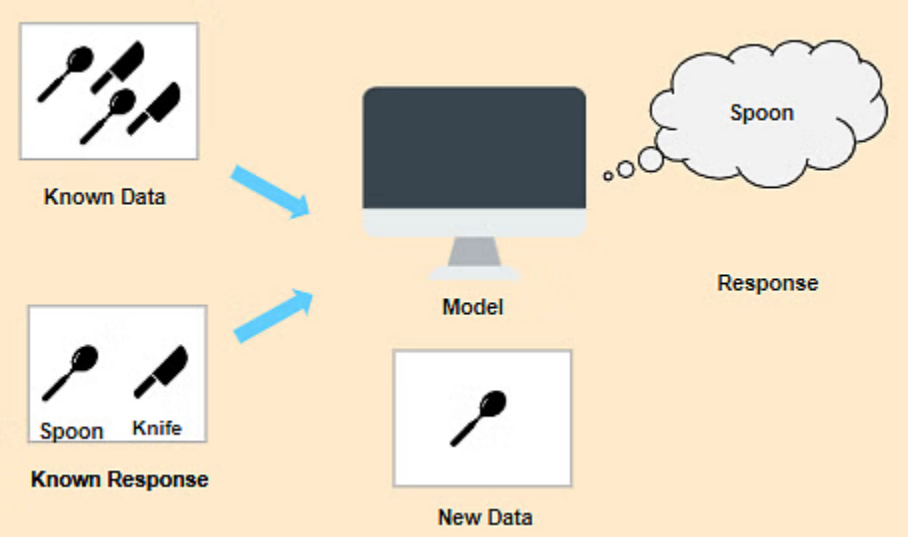
Similarly, there are four categories of machine learning algorithms as shown below −

- Supervised learning algorithm
- Unsupervised learning algorithm
- Semi-supervised learning algorithm
- Reinforcement learning algorithm

However, the most commonly used ones are supervised and unsupervised learning.

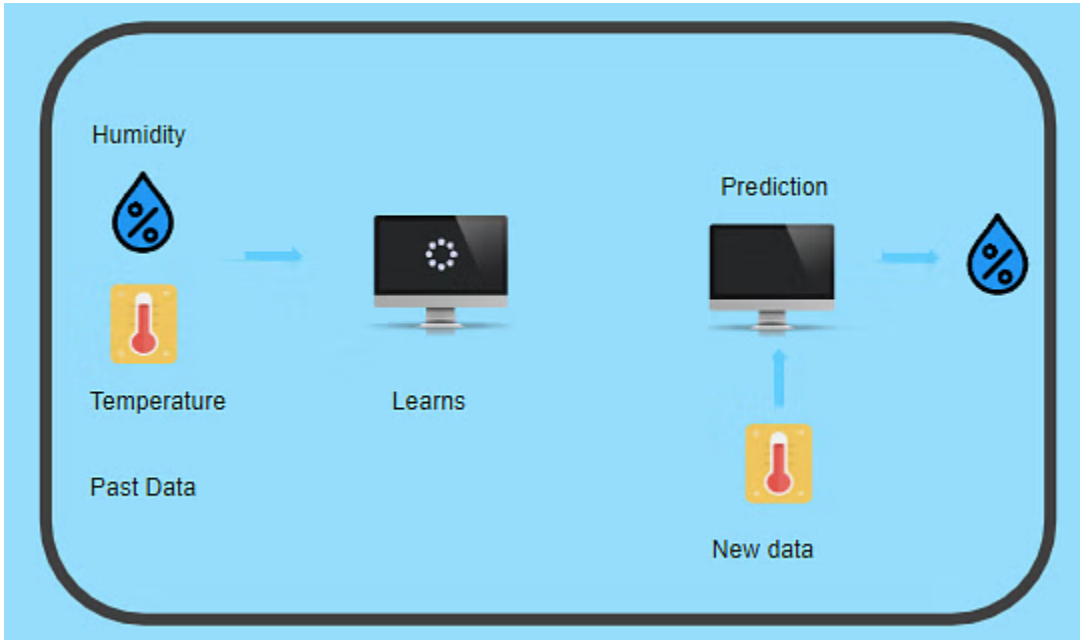
Supervised Learning

Supervised learning is commonly used in real world applications, such as face and speech recognition, products or movie recommendations, and sales forecasting. In Supervised Learning, the machine learns under supervision. It contains a model that is able to predict with the help of a labeled dataset. A labeled dataset is one where you already know the target answer.



Supervised learning can be further classified into two types - **Regression and Classification.**

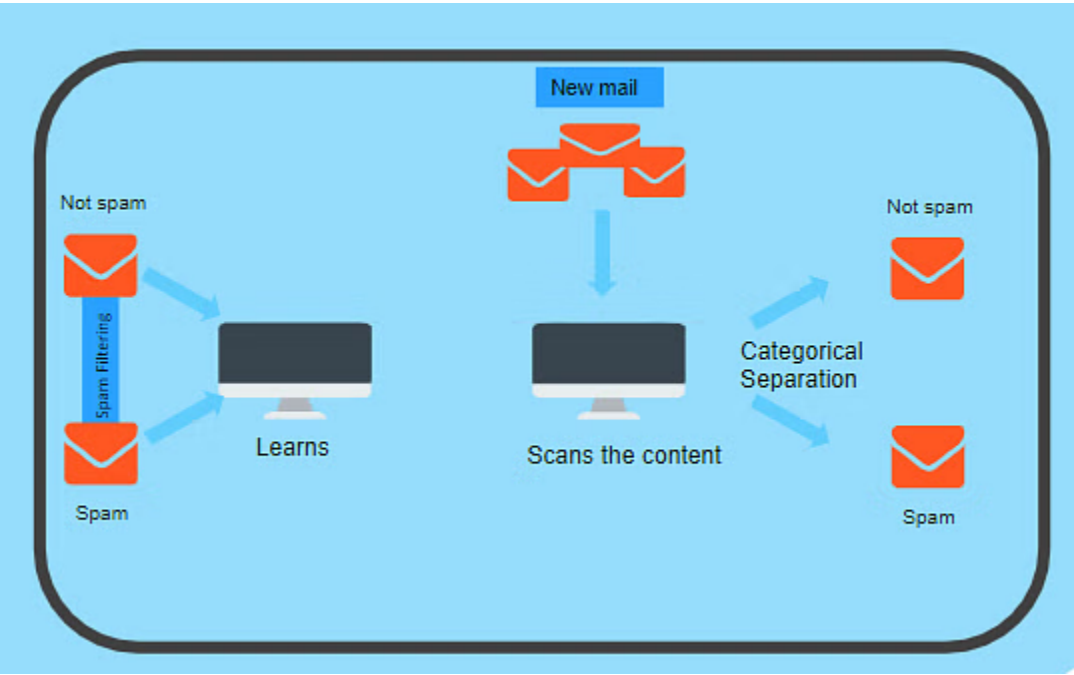
Regression trains on and predicts a continuous-valued response, for example predicting real estate prices. Regression is used when the output variable is a real or continuous value. In this case, there is a relationship between two or more variables i.e., a change in one variable is associated with a change in the other variable. For example, salary based on work experience or weight based on height, etc.



Let's consider two variables - humidity and temperature. Here, 'temperature' is the independent variable and 'humidity' is the dependent variable. If the temperature increases, then the humidity decreases.

These two variables are fed to the model and the machine learns the relationship between them. After the machine is trained, it can easily predict the humidity based on the given temperature.

Classification attempts to find the appropriate class label, such as analyzing positive/negative sentiment, male and female persons, benign and malignant tumors, secure and unsecure loans etc. Classification is used when the output variable is categorical i.e. with 2 or more classes. For example, yes or no, male or female, true or false, etc.



In order to predict whether a mail is spam or not, we need to first teach the machine what a spam mail is. This is done based on a lot of spam filters - reviewing the content of the mail, reviewing the mail header, and then searching if it contains any false information. Certain keywords and blacklist filters that blackmails are used from already blacklisted spammers.

All of these features are used to score the mail and give it a spam score. The lower the total spam score of the email, the more likely that it is not a scam.

Based on the content, label, and the spam score of the new incoming mail, the algorithm decides whether it should land in the inbox or spam folder.

In supervised learning, learning data comes with description, labels, targets or desired outputs and the objective is to find a general rule that maps inputs to outputs. This kind of learning data is called labeled data. The learned rule is then used to label new data with unknown outputs.

Supervised learning involves building a machine learning model that is based on labeled samples. For example, if we build a system to estimate the price of a plot of land or a house based on various features, such as size, location, and so on, we first need to create a database and label it. We need to teach the algorithm what features correspond to what prices. Based on this data, the algorithm will learn how to calculate the price of real estate using the values of the input features

Supervised learning deals with learning a function from available training data. Here, a learning algorithm analyzes the training data and produces a derived function that can be used for mapping new examples. There are many supervised learning algorithms such as Logistic Regression, Neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers.

Common examples of supervised learning include classifying emails into spam and not-spam categories, labeling web pages based on their content, and voice recognition.

Real-Life Applications of Supervised Learning

- Risk Assessment

Supervised learning is used to assess the risk in financial services or insurance domains in order to minimize the risk portfolio of the companies.

- Image Classification

Image classification is one of the key use cases of demonstrating supervised machine learning. For example, Facebook can recognize your friend in a picture from an album of tagged photos.

- Fraud Detection

To identify whether the transactions made by the user are authentic or not.

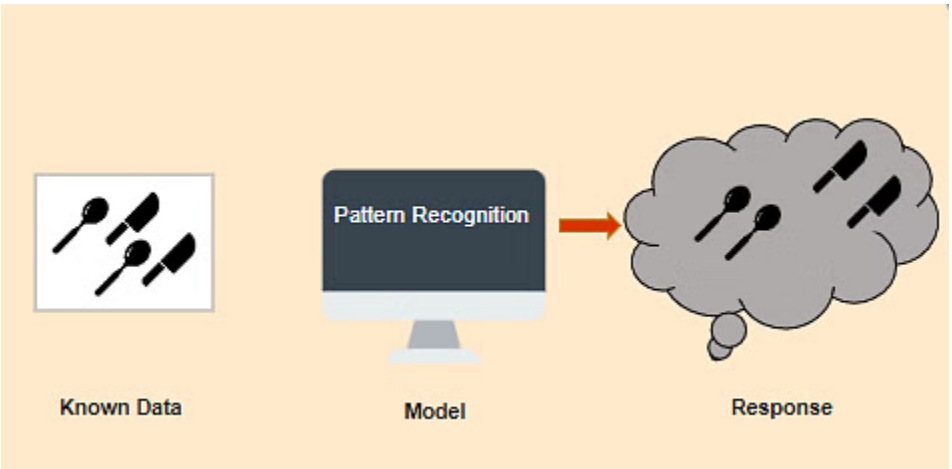
- Visual Recognition

The ability of a machine learning model to identify objects, places, people, actions, and images.

Unsupervised Learning

Unsupervised learning is used to detect anomalies, outliers, such as fraud or defective equipment, or to group customers with similar behaviors for a sales campaign. It is the opposite of supervised learning. There is no labeled data here.

In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response



Let's take a similar example is before, but this time we do not tell the machine whether it's a spoon or a knife. The machine identifies patterns from the given set and groups them based on their patterns, similarities, etc.

When learning data contains only some indications without any description or labels, it is up to the coder or to the algorithm to find the structure of the underlying data, to discover hidden patterns, or to determine how to describe the data. This kind of learning data is called unlabeled data.

Suppose that we have a number of data points, and we want to classify them into several groups. We may not exactly know what the criteria of classification would be. So, an unsupervised learning algorithm tries to classify the given dataset into a certain number of groups in an optimum way.

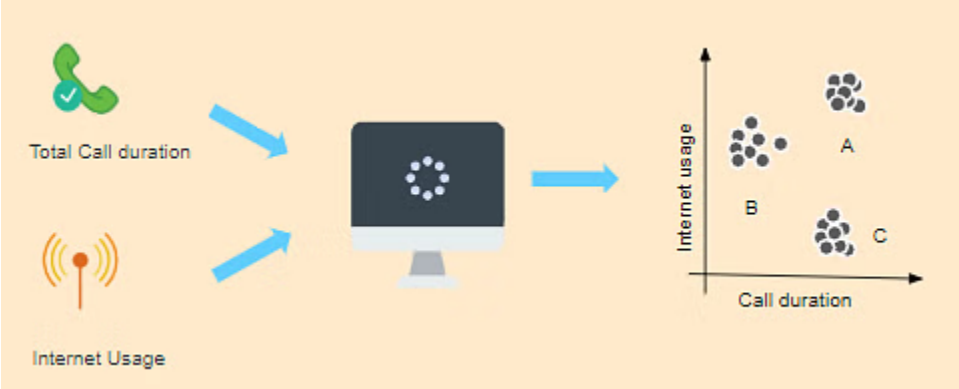
Unsupervised learning algorithms are extremely powerful tools for analyzing data and for identifying patterns and trends. They are most commonly used for clustering similar input into logical groups. Unsupervised learning algorithms include Kmeans, Random Forests, Hierarchical clustering and so on.

Unsupervised learning can be further grouped into types:

1. Clustering
2. Association

1. Clustering - Unsupervised Learning

Clustering is the method of dividing the objects into clusters that are similar between them and are dissimilar to the objects belonging to another cluster. For example, finding out which customers made similar product purchases.

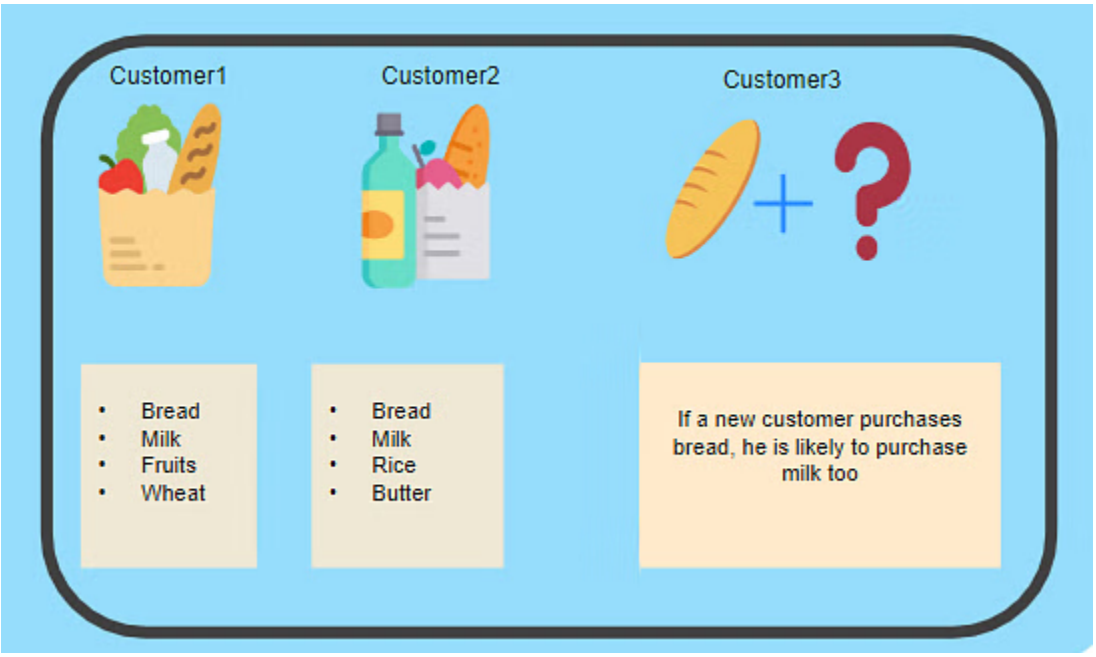


Suppose a telecom company wants to reduce its customer churn rate by providing personalized call and data plans. The behavior of the customers is studied and the model segments the customers with similar traits. Several strategies are adopted to minimize churn rate and maximize profit through suitable promotions and campaigns.

On the right side of the image, you can see a graph where customers are grouped. Group A customers use more data and also have high call durations. Group B customers are heavy Internet users, while Group C customers have high call duration. So, Group B will be given more data benefit plans, while Group C will be given cheaper call rate plans and group A will be given the benefit of both.

2. Association - Unsupervised Learning

Association is a rule-based machine learning to discover the probability of the co-occurrence of items in a collection. For example, finding out which products were purchased together.



Let's say that a customer goes to a supermarket and buys bread, milk, fruits, and wheat. Another customer comes and buys bread, milk, rice, and butter. Now, when another customer comes, it is highly likely that if he buys bread, he will buy milk too. Hence, a relationship is established based on customer behavior and recommendations are made.

Real-Life Applications of Unsupervised Learning

- Market Basket Analysis

It is a machine learning model based on the algorithm that if you buy a certain group of items, you are less or more likely to buy another group of items.

- Semantic Clustering

Semantically similar words share a similar context. People post their queries on websites in their own ways. Semantic clustering groups all these responses with the same meaning in a cluster to ensure that the customer finds the information they want quickly and easily. It plays an important role in information retrieval, good browsing experience, and comprehension.

- Delivery Store Optimization

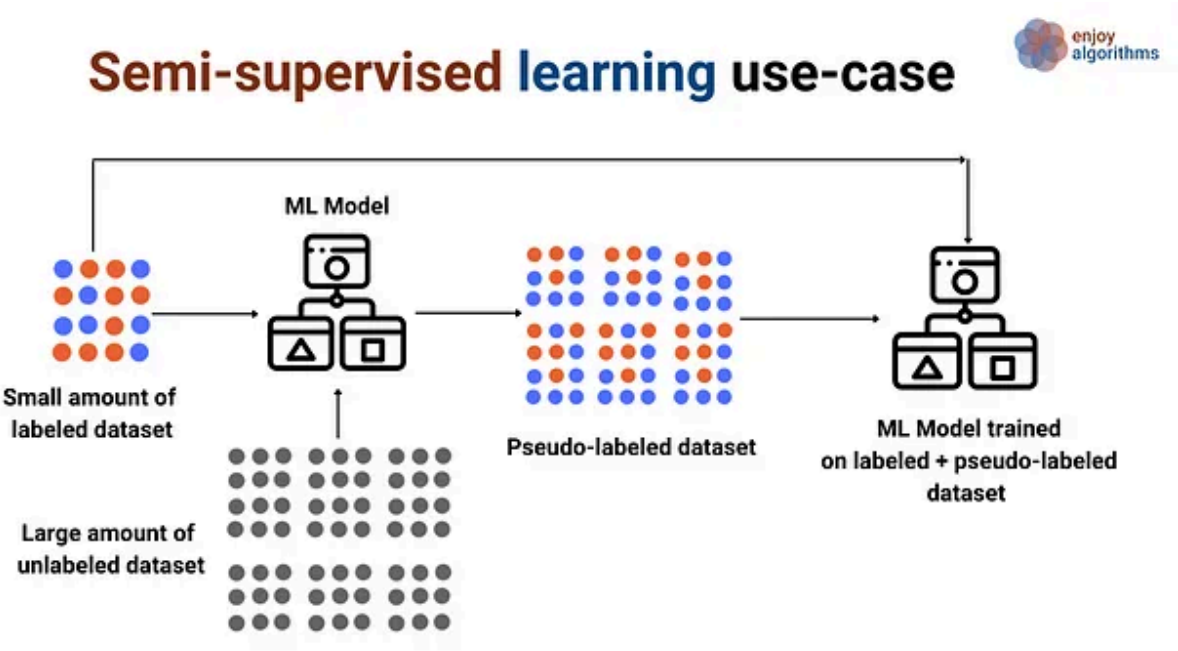
Machine learning models are used to predict the demand and keep up with supply. They are also used to open stores where the demand is higher and optimizing roots for more efficient deliveries according to past data and behavior.

- Identifying Accident Prone Areas

Unsupervised machine learning models can be used to identify accident-prone areas and introduce safety measures based on the intensity of those accidents.

Semi-supervised Learning

If some learning samples are labeled, but some others are not labeled, then it is semi-supervised learning. It makes use of a large amount of unlabeled data for training and a small amount of labeled data for testing. Semi-supervised learning is applied in cases where it is expensive to acquire a fully labeled dataset while more practical to label a small subset. For example, it often requires skilled experts to label certain remote sensing images, and lots of field experiments to locate oil at a particular location, while acquiring unlabeled data is relatively easy.



For example, suppose there is a large chunk of data in the image above, and a small amount of labeled dataset is present. We can train the model using that small amount of labeled data and then predict on the unlabelled dataset. Prediction on an unlabelled dataset will attach the label with every data sample with little accuracy, termed as a Pseudo-labeled dataset. Now a new model can be trained with the mixture of the true-labeled dataset and pseudo-labeled dataset.