# *Chapter 1: Introduction to Machine Learning*

## About Machine Learning

Machine learning is the field of study where the computers are given the capability to learn without being explicitly programmed. In short, we can say that the machine is given the ability to learn.

The first step in the process is to feed the data and then train the machines by using machine learning models and different algorithms. We select the algorithms based on the data and the tasks which we need to perform. For example, training the students for exams. Here, the students are given exam so that they can train themselves with what kind of questions which can come for the final exams. They go through the previous exam papers and lecture notes or videos each time when they appear. They are training their brain to answer all type of questions can come up in the exam for the future. In the similar way, the machines are trained on the previous data and identify the insights which we need to gain from the existing data and then test them on the new data which they receive.

## Types of ML

There are generically three types of learning;

1. ### *Supervised Learning*
   In supervised learning, the training data is feed to the algorithm to get desired output. Now this desired output is known to us. We try to check whether they are meeting with the same values or not. We either apply with regressions methods or classification methods to get the requirement. Some of the algorithms are linear regression, Logistic Regression, SVM, KNN and so on. We generally apply it on risk evaluation or forecasting sales.

2. ### *Unsupervised learning*
   In unsupervised learning, we can say that the machine is learning without the teacher or the training data is unlabeled. Here, we apply with association or clustering methods. Some of the algorithms are K – Means, C – Means, Apriori and so on. The machine will try to discover the patterns and understand the data. We generally apply where we need to identify the anomaly detection.

3. ### *Reinforcement learning*
   It is neither based on the supervised learning nor the unsupervised learning. In this learning, the algorithms learn to react to an environment on tier own. This is used in the field of robotics and gaming field. For a learning agent

## Methods or techniques of Machine Learning

A machine learning algorithm is also known as model. It is the mathematical expression which helps us to represent the data to identify the context or problem of the business. The aim is to find the insights from the data.

There are different methods or techniques, which is used in Machine Learning:

1. *Regression***:** This method falls under the supervised learning. The regression method helps us to predict the numerical value based on the previous data which we have collected. We use the mathematical equation of line ($y = m * x + b$) to model a dataset. We can use simple regression model for one factor, but multiple regression model if we have more than one factor.
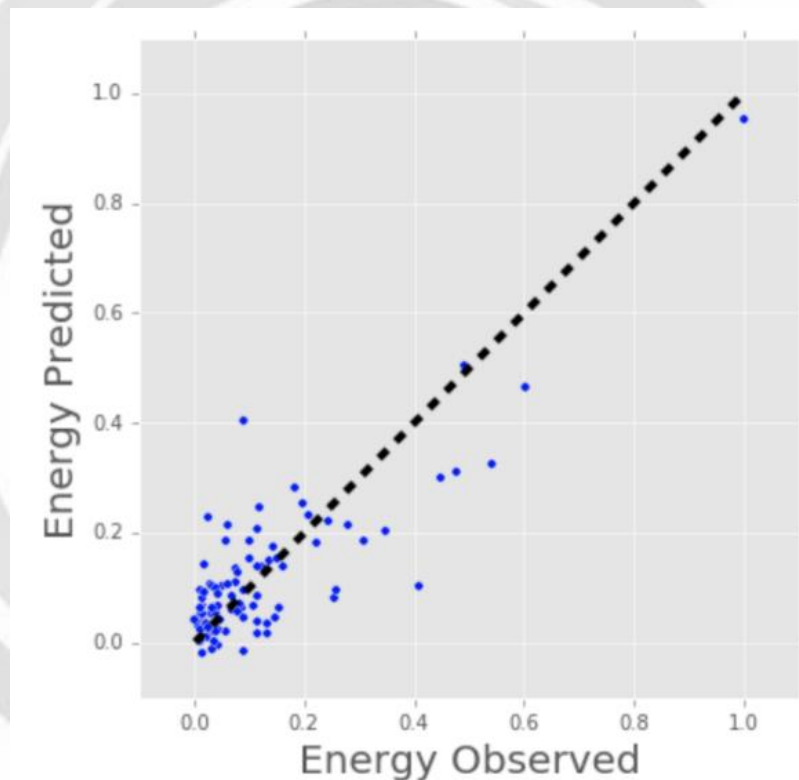


2. *Classification***:** The other method of supervised ML, helps us to predict or explain the class value. For example, we can predict whether the value will be yes or no, that is we will have two outcomes. For example, the image contains car or the image does not contain car. The simplest classification algorithm is logistic regression. It helps us to estimate the probability of an occurrence of an event based on one or more inputs.
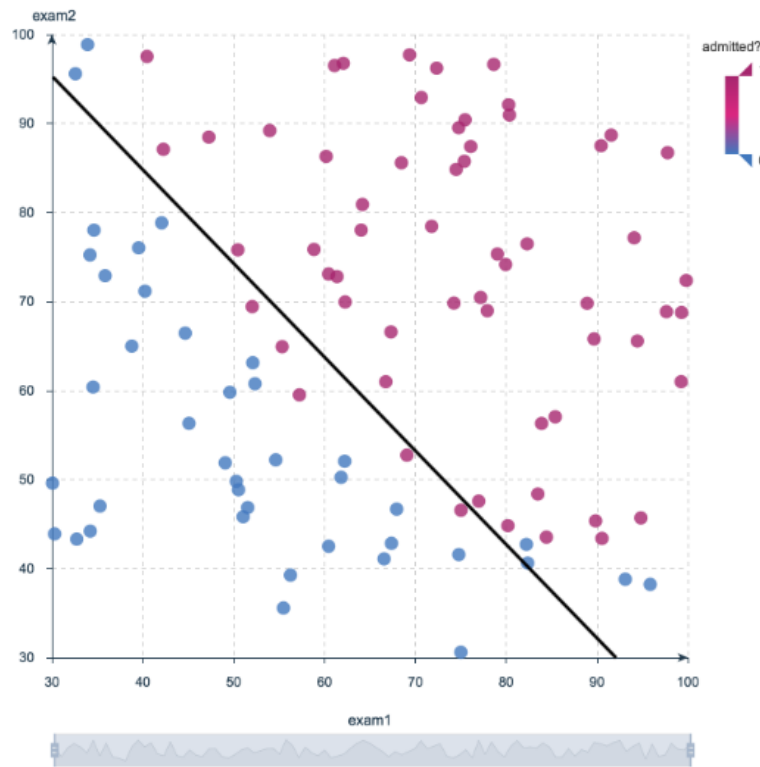
*3.* ***Clustering***: This is categorized into one of unsupervised ML. it helps us to identify the group or cluster of observations that have similar characteristics. The most popular clustering method is K-Means, where K represents the number of cluster that the user chooses to create. With the help of K-means it follows the below steps:

i) Randomly chooses K centers within the data.
ii) Assigns each data point to the closest of the randomly created centers.
iii) Re-computes the center of each cluster.
iv) If centers don't change (or change very little), the process is finished. Otherwise, we return to step ii.



Clustering Buildings into Efficient (Green) and Inefficient (Red) Groups.
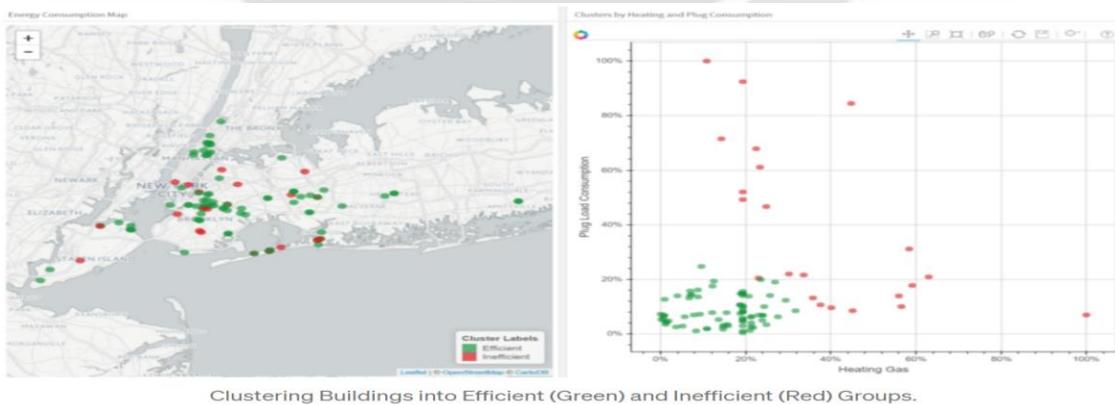
4. ***Dimensionality Reduction:*** From the data which we collect, there are plenty of variables which we need to process to get the insight. There are some data which we don't require, and hence we need to remove those. Here we use dimensionality reduction to remove those variables by identifying them. For example, from the image which we have collected there are thousands of pixels and not every pixel is used from analysis. We will use dimensionality reduction algorithms to make the data manageable. With the help of Principal Component Analysis (PCA), we will be able to reduce the dimension by finding new vectors which help in maximizing the linear variation of the data.
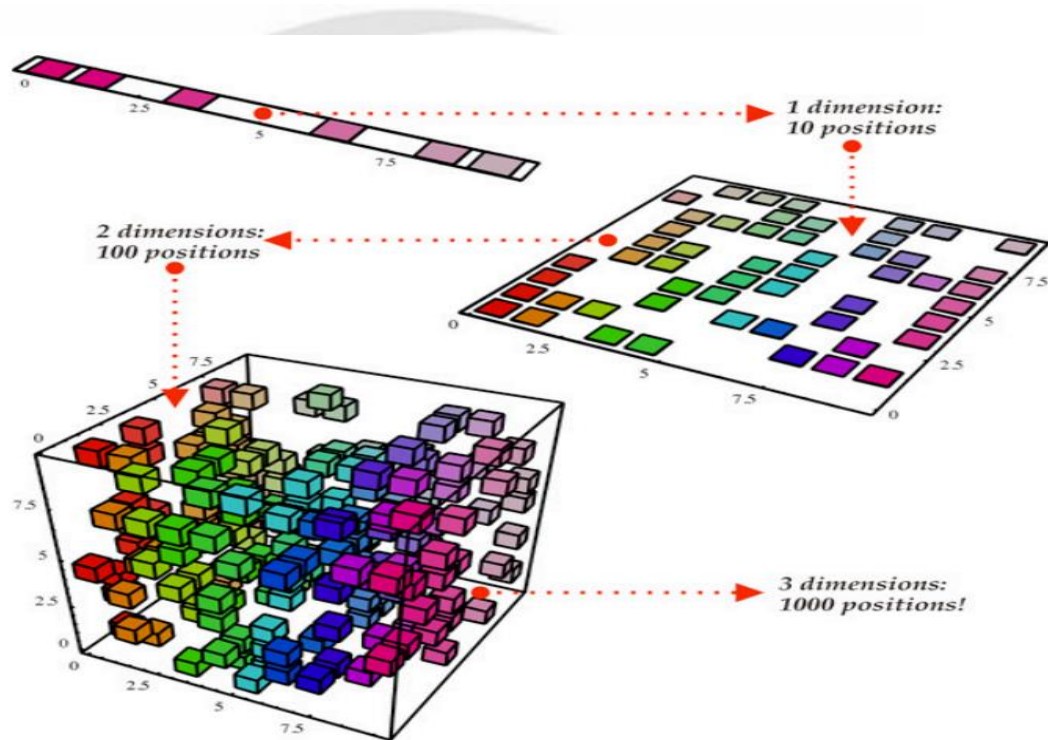


5. ***Ensemble Methods***: As the word ensemble speaks about, a group of something producing single effect, we use different algorithms together to make one good algorithm which helps us to complete the task. With the ensemble method, we combine several predictive models (supervised ML) to get a higher quality prediction instead of each model could provide in its own. For example, Random Forest method algorithm is an ensemble method that combines many decision trees trained with different samples of the data sets. The results which acquire will be better. The quality of the predictions from a Random Forest is higher than the quality of predictions which we can estimate it from single decision tree.

Ensemble methods consist of joining several weak learners in order to build a strong learner.
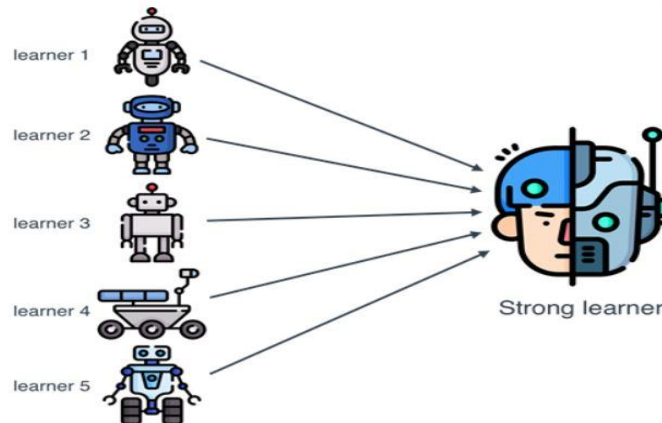


6. *Neural nets and Deep Learning*: When we encounter with a non-linear data, we use neural networks to add in the layers to data. By creating a structure of neural networks, the data becomes flexible enough to use the linear and logistic regression. The term Deep Learning (DL) comes from a neural network with many hidden layers. The Layers which we add in are known as the hidden layers which we can see in the picture.
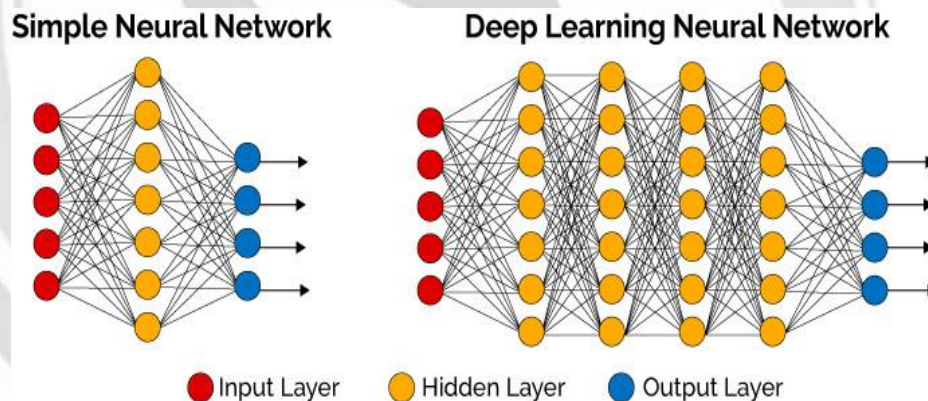


7. *Transfer Learning*: This specific helps us re-using the part trained neural networks and adopt to the similar task which we need. We can transfer the different layers within the neural network to add on the new neural network which we create to complete the task which we need. The advantage of this is that we need less data to train the neural network and both time and money are reduced.

## Transfer Learning
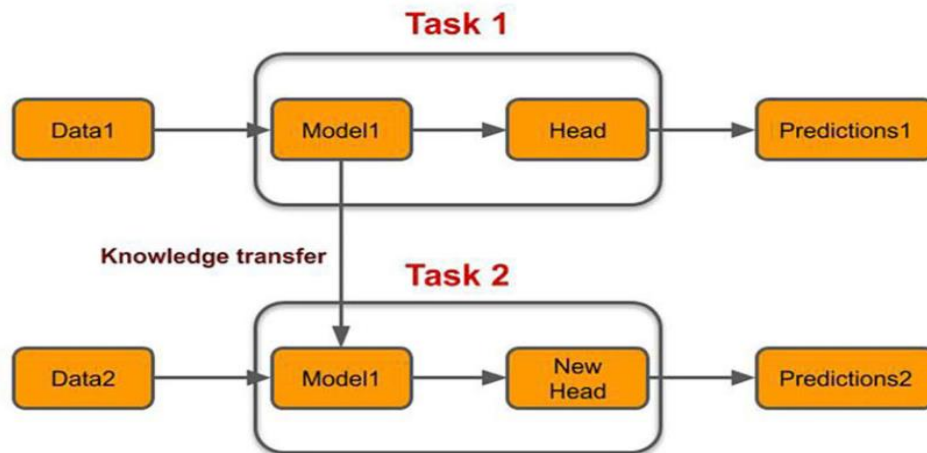


8. ***Reinforcement Learning:*** The term reinforcement also means by gaining strength with adding on something to it. In the same manner, the when the agent gains the experience it is able to maximize the rewards. Hence, we can say that reinforcement learning is one of ML methods which help an agent to learn through experience.
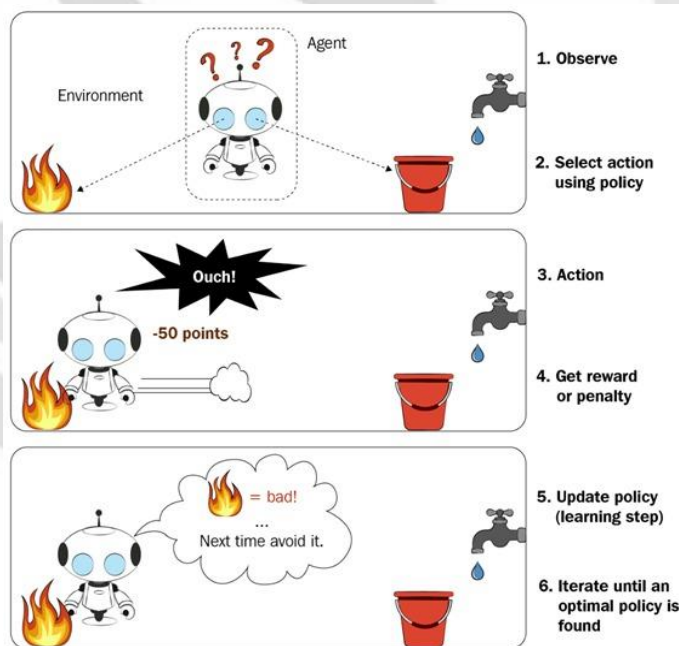


9. ***Natural Language Processing:*** The best example is when we start typing the text in Google chrome or your phone; we might have noticed that the words are auto completed and most of the times, the words which we need comes the next. This is because of the Natural

Language Processing (NLP). NLP is one the technique which is used widely to make the machine understand the human language. And based each user behavior pattern of using the language it will suggest the words. Here, we can say that we are teaching machine the human language. This process is done with the help Natural Language toolkit (NLTK).
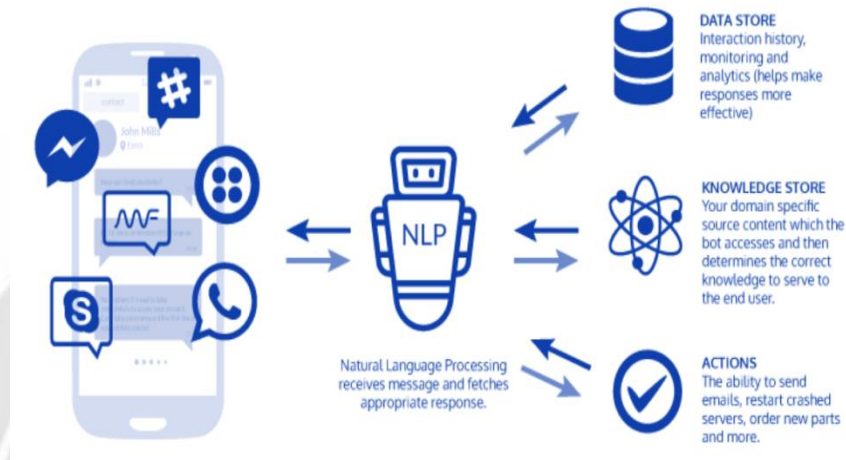


10. **Word Embedding:** Word Embedding is the method in which the words are represented in the numerical form. It might be the frequency of the word or weighted frequency of the word which represents it. The wrod2vec is one of the methods which we use to convert the word to numeric vectors and use it to find the synonyms or do arithmetic calculations. For example, there are different words posted on twitter by thousands of users. And among these users of the data which we have collected, we known who all have bought the houses. We will combine word2vec to logistic regression and predict the probability of a new twitter use buying the house.
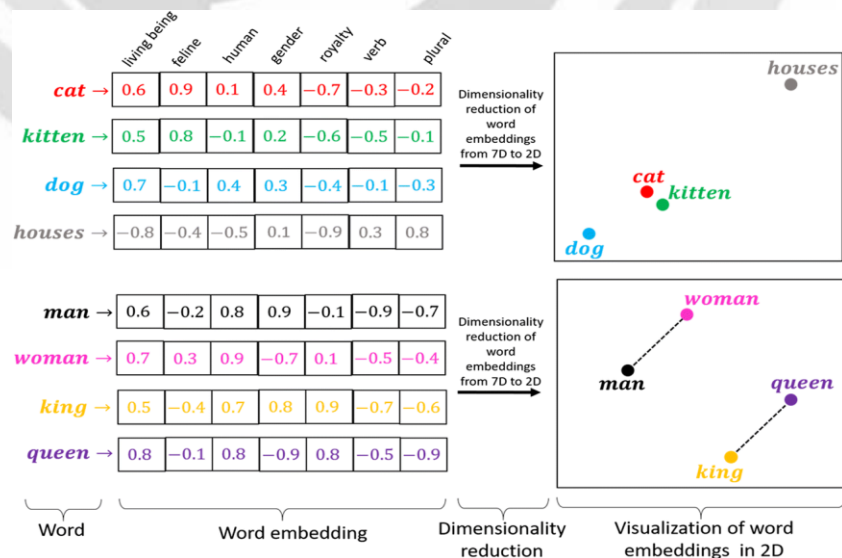
## Life Cycle of Machine Learning

1. *Gathering the data***:** This is the first step, with the specific goal to identify and obtain all the data related to problems. We need to identify the sources of the data also like files, data base, internet and so on. Once collected from the different sources, we will integrate those data to the required tool which we use.
2. *Data preparation***:** Once, we explore the data to identify what the data consist of. We look into the format of the data, the characteristics and quality. In this step, we also try to understand the correlation between the existing data and outliers.
3. *Data Wrangling***:** This is the processing of cleaning the data based on the requirement. We will be only selecting the variables which will used for the analysis, we might be removing the missing values or duplicate values which we don't require. Sometimes the missing values are not removed and instead will be managed accordingly may be by replacing it means. We will validate the invalid data.
4. *Data Analysis***:** Once the data is cleaned and prepared, we will select the analytical techniques which we need to use. Here, we will be using one or more machine learning algorithms to do the data analysis which we require. We build the specific model which gives us the best possible solution. We might use techniques like classification, clustering, regression, etc. based on how it is suited.
5. *Train the model***:** In this step, we will train the model to improve the better outcomes of the problem. We will be data sets to train the model using various machine algorithms. Training is required to understand the patterns or the behavior of the data.
6. *Testing the model***:** Once our machine model is been trained in a given data set, we test on the different other dataset to check the accuracy of the model. Testing the model helps us to identify the accuracy of the model based on the requirement of the project.
7. *Deployment***:** And last not the least, we need to deploy the model in the real-world system. Before deployment, we check whether the model is improving the performance based on the available or not.

By following these steps, we are able train different machine learning models to do analysis of the data and also create different automated bots and machine helping the user to make the different task easy.