UNIT-1

Introduction to Machine Learning

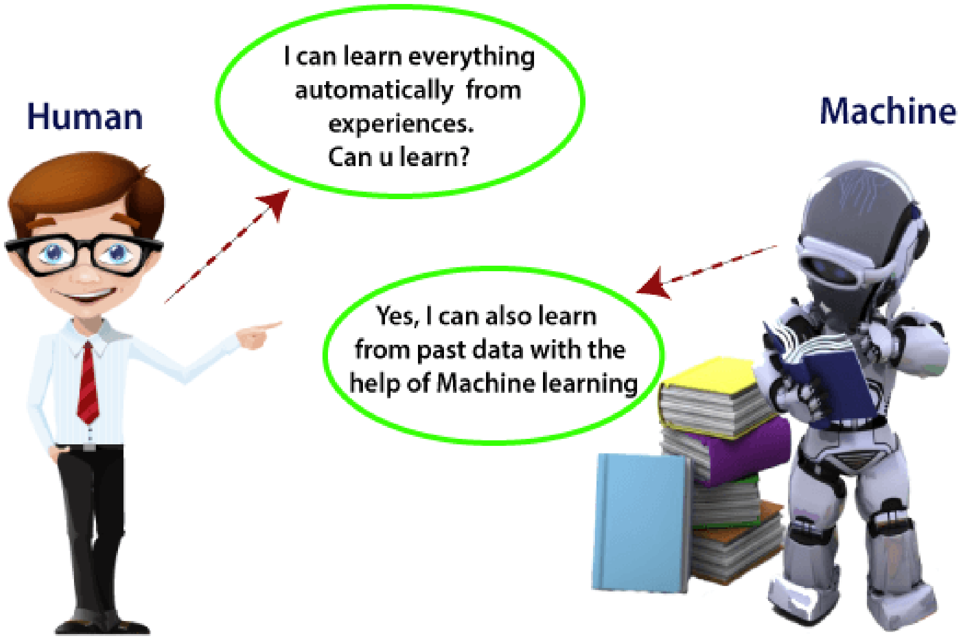
What is Machine Learning?

- In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions.
- But can a machine also learn from experiences or past data like a human does?
- So here comes the role of **Machine Learning.**
- Before answering the question 'What is machine learning?' more fundamental questions that peep into one's mind are
    - Do machines really learn?
    - If so, how do they learn?
    - Which problem can we consider as a well-posed learning problem? What are the important features that are required to well-define a learning problem?
- At the onset, it is important to formalize the definition of machine learning. This will itself address the first question, i.e. if machines really learn. There are multiple ways to define machine learning.
- But the one which is perhaps most relevant, concise and accepted universally is the one stated by Tom M. Mitchell, Professor of Machine Learning Department, School of Computer Science, Carnegie Mellon University.

- Tom M. Mitchell has defined machine learning as,
  'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.'

- What this essentially means is that a machine can be considered to learn if it is able to gather experience by doing a certain task and improve its performance in doing the similar tasks in the future.
- When we talk about past experience, it means past data related to the task.

- This data is an input to the machine from some source.
- In the context of the learning to play checkers, E represents the experience of playing the game, T represents the task of playing checkers and P is the performance measure indicated by the percentage of games won by the player.
- The same mapping can be applied for any other machine learning problem, for example, image classification problem.
- In context of image classification, E represents the past data with images having labels or assigned classes (for example whether the image is of a class cat or a class dog or a class elephant etc.), T is the task of assigning class to new, unlabelled images and P is the performance measure indicated by the percentage of images correctly classified.


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

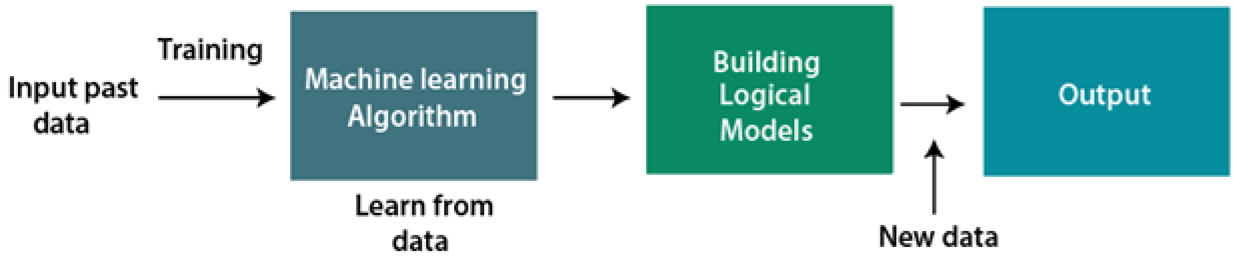
<u>Introduction to Machine Learning</u>

- A subset of artificial intelligence known as machine learning focuses primarily on the **<u>creation of algorithms that enable a computer to independently learn from data and previous experiences.</u>**
- Arthur Samuel first used the term **<u>"machine learning" in 1959</u>**. It could be summarized as follows:

- Without being explicitly programmed, **machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things.**
- Machine learning algorithms create a mathematical model that, without being explicitly programmed, aids in making predictions or decisions with the assistance of sample historical data, or training data.
- The performance will rise in proportion to the quantity of information we provide.
- For the purpose of developing predictive models, machine learning brings together statistics and computer science.
- Algorithms that learn from historical data are either constructed or utilized in machine learning.
- **A machine can learn if it can gain more data to improve its performance.**

How does Machine Learning work?

- A machine learning system builds prediction models, learns from previous data, and predicts the output of new data whenever it receives it.
- The amount of data helps to build a better model that accurately predicts the output, which in turn affects the accuracy of the predicted output.
- Let's say we have a complex problem in which we need to make predictions.
- Instead of writing code, we just need to feed the data to generic algorithms, which build the logic based on the data and predict the output.
- Our perspective on the issue has changed as a result of machine learning. The Machine Learning algorithm's operation is depicted in the following block diagram:

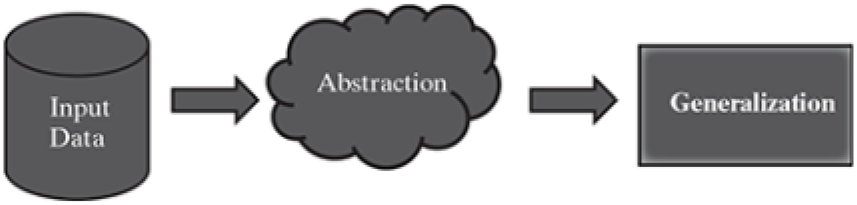


- The basic machine learning process can be divided into three parts.

  1. **Data Input:** Past data or information is utilized as a basis for future decision-making

  2. **Abstraction:** The input data is represented in a broader way through the underlying algorithm.

  3. **Generalization:** The abstracted representation is generalized to form a framework for making decisions.



- Moving to the machine learning paradigm, the vast pool of knowledge is available from the data input.
- However, rather than using it in entirety, a concept map, much in line with the animal group to characteristic mapping explained above, is drawn from the input data.
- This is nothing but knowledge abstraction as performed by the machine.
- In the end, the abstracted mapping from the input data can be applied to make critical conclusions.
- For example, if the group of an animal is given, understanding of the characteristics can be automatically made.
- Reversely, if the characteristic of an unknown animal is given, a definite conclusion can be made about the animal group it belongs to.
- This is generalization in context of machine learning.

Abstraction:

- During the machine learning process, knowledge is fed in the form of input data.
- However, the data cannot be used in the original shape and form.
- As we saw in the example above, abstraction helps in deriving a conceptual map based on the input data.
- This map, or a model as it is known in the machine learning paradigm, is summarized knowledge representation of the raw data.
- The model may be in any one of the following forms:
    - Computational blocks like if/else rules
    - Mathematical equations
    - Specific data structures like trees or graphs
    - Logical groupings of similar observations.
- The choice of the model used to solve a specific learning problem is a human task.
- The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:
    - The type of problem to be solved: Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
    - Nature of the input data: How exhaustive the input data is, whether the data has no values for many fields, the data types, etc.
    - Domain of the problem: If the problem is in a business critical domain with a high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.
- Once the model is chosen, the next task is to fit the model based on the input data.
- Let’s understand this with an example.
- In a case where the model is represented by a mathematical equation, say ‘y = c + c x’ (the model is known as simple linear regression which we will study in a later chapter), based on the input data, we have to find out the values of c and c.
- Otherwise, the equation (or the model) is of no use.

- So, fitting the model, in this case, means finding the values of the unknown coefficients or constants of the equation or the model.
- This process of fitting the model based on the input data is known as training.
- Also, the input data based on which the model is being finalized is known as training data.

Generalization

- The first part of machine learning process is abstraction i.e. abstract the knowledge which comes as input data in the form of a model.
- However, this abstraction process, or more popularly training the model, is just one part of machine learning.
- The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions.
- This is achieved as a part of generalization.
- This part is quite difficult to achieve.
- This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics.
- But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems:
  1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
  2. The test data possess certain characteristics apparently unknown to the training data.
- Hence, a precise approach of decision-making will not work.
- An approximate or heuristic approach, much like gut feeling-based decision-making in human beings, has to be adopted.
- This approach has the risk of not making a correct decision – quite obviously because certain assumptions that are made may not be true in reality.
- But just like machines, same mistakes can be made by humans too when a decision is made based on intuition or gut-feeling – in a situation where exact reason-based decision-making is not possible.

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Need for Machine Learning

- The demand for machine learning is steadily rising. Because it is able to perform tasks that are too complex for a person to directly implement, machine learning is required.
- Humans are constrained by our inability to manually access vast amounts of data; as a result, we require computer systems, which is where machine learning comes in to simplify our lives.
- By providing them with a large amount of data and allowing them to automatically explore the data, build models, and predict the required output, we can train machine learning algorithms.
- The cost function can be used to determine the amount of data and the machine learning algorithm's performance.
- We can save both time and money by using machine learning.
- The significance of AI can be handily perceived by its utilization's cases, Presently, AI is utilized in self-driving vehicles, digital misrepresentation identification, face acknowledgment, and companion idea by Facebook, and so on.
- Different top organizations, for example, Netflix and Amazon have constructed AI models that are utilizing an immense measure of information to examine the client interest and suggest item likewise.

Following are some key points which show the importance of Machine Learning:

- Rapid increment in the production of data
- Solving complex problems, which are difficult for a human
- Decision making in various sector including finance
- Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**

1) Supervised Learning

- In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.
- The system uses labeled data to build a model that understands the datasets and learns about each one.
- After the training and processing are done, we test the model with sample data to see if it can accurately predict the output.
- The mapping of the input data to the output data is the objective of supervised learning.
- The managed learning depends on oversight, and it is equivalent to when an understudy learns things in the management of the educator.
- Spam filtering is an example of supervised learning.
- Supervised learning can be grouped further in two categories of algorithms:
    - **Classification**
    - **Regression**

2) Unsupervised Learning

- Unsupervised learning is a learning method in which a machine learns without any supervision.
- The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision.
- The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.
- In unsupervised learning, we don't have a predetermined result.
- The machine tries to find useful insights from the huge amount of data.
- It can be further classifieds into two categories of algorithms:
    - **Clustering**
    - **Association**

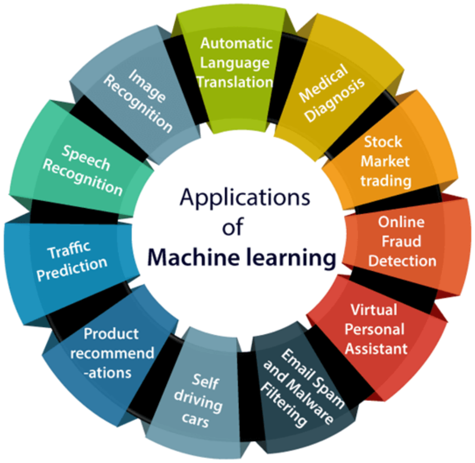
3) Reinforcement Learning

- Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action.
- The agent learns automatically with these feedbacks and improves its performance.
- In reinforcement learning, the agent interacts with the environment and explores it.
- The goal of an agent is to get the most reward points, and hence, it improves its performance.
- The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Applications of Machine learning

- Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day.
- We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc.

- Below are some most trending real-world applications of Machine Learning:


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

1. Image Recognition:

- Image recognition is one of the most common applications of machine learning.
- It is used to identify objects, persons, places, digital images, etc.
- The popular use case of image recognition and face detection is, Automatic friend tagging suggestion:
- Facebook provides us a feature of auto friend tagging suggestion.
- Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.
- It is based on the Facebook project named "Deep Face," which is responsible for face recognition and person identification in the picture.

2. Speech Recognition

- While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.
- Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition."
- At present, machine learning algorithms are widely used by various applications of speech recognition.
- Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

- If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.
- It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:
- Real Time location of the vehicle form Google Map app and sensors
- Average time has taken on past days at the same time.
- Everyone who is using Google Map is helping this app to make it better.
- It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

- Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user.
- Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.
- Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.
- As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars:

- One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars.
- Tesla, the most popular car manufacturing company is working on self-driving car.
- It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. Email Spam and Malware Filtering:

- Whenever we receive a new email, it is filtered automatically as important, normal, and spam.
- We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning.
- Below are some spam filters used by Gmail:
    - Content Filter
    - Header filter
    - General blacklists filter
    - Rules-based filters
    - Permission filters
- Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree, and Naïve Bayes classifier are used for email spam filtering and malware detection.

7. Virtual Personal Assistant:

- We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri.
- As the name suggests, they help us in finding the information using our voice instruction.
- These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.
- These virtual assistants use machine learning algorithms as an important part.
- These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

8. Online Fraud Detection:

- Machine learning is making our online transaction safe and secure by detecting fraud transaction.
- Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal money in the middle of a transaction.
- So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction.
- For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round.
- For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

9. Stock Market trading:

- Machine learning is widely used in stock market trading.
- In the stock market, there is always a risk of up and downs in shares, so for this machine learning's long short term memory neural network is used for the prediction of stock market trends.

10. Medical Diagnosis:

- In medical science, machine learning is used for diseases diagnoses.
- With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.
- It helps in finding brain tumors and other brain-related diseases easily.

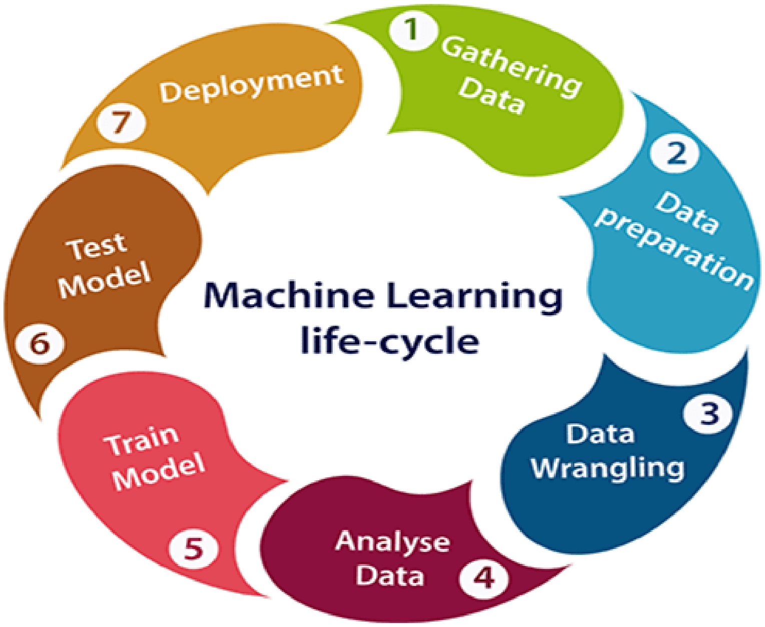
11. Automatic Language Translation:

- Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages.

- Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.
- The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

Machine learning Life cycle

- Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed.
- But how does a machine learning system work? So, it can be described using the life cycle of machine learning.
- Machine learning life cycle is a cyclic process to build an efficient machine learning project.
- The main purpose of the life cycle is to find a solution to the problem or project.
- Machine learning life cycle involves seven major steps, which are given below:
    - **Gathering Data**
    - **Data preparation**
    - **Data Wrangling**
    - **Analyze Data**
    - **Train the model**
    - **Test the model**
    - **Deployment**



- The most important thing in the complete process is to understand the problem and to know the purpose of the problem.
- Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.
- In the complete life cycle process, to solve a problem, we create a machine learning system called "model", and this model is created by providing "training".
- But to train a model, we need data, hence, life cycle starts by collecting data.

1. Gathering Data

- Data Gathering is the first step of the machine learning life cycle.
- The goal of this step is to identify and obtain all data-related problems.
- In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices.
- It is one of the most important steps of the life cycle.
- The quantity and quality of the collected data will determine the efficiency of the output.
- The more will be the data, the more accurate will be the prediction.

- **This step includes the below tasks:**
- Identify various data sources
- Collect data
- Integrate the data obtained from different sources
- By performing the above task, we get a coherent set of data, also called as a dataset. It will be used in further steps.

2. Data preparation

- After collecting the data, we need to prepare it for further steps.
- Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
- In this step, first, we put all data together, and then randomize the ordering of data.
- This step can be further divided into two processes:

    - **Data exploration:**
    - It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.
    - A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.

    - **Data pre-processing:**
    - Now the next step is preprocessing of data for its analysis.

3. Data Wrangling

- Data wrangling is the process of cleaning and converting raw data into a useable format.
- It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step.
- It is one of the most important steps of the complete process.

- Cleaning of data is required to address the quality issues.
- It is not necessary that data we have collected is always of our use as some of the data may not be useful.
- In real-world applications, collected data may have various issues, including:
    - Missing Values
    - Duplicate data
    - Invalid data
    - Noise

So, we use various filtering techniques to clean the data.

- It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

4. Data Analysis

- Now the cleaned and prepared data is passed on to the analysis step. This step involves:
    - **Selection of analytical techniques**
    - **Building models**
    - **Review the result**
- The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.
- It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.
- Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

- Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.

- We use datasets to train the model using various machine learning algorithms.
- Training a model is required so that it can understand the various patterns, rules, and, features.

6. Test Model

- Once our machine learning model has been trained on a given dataset, then we test the model.
- In this step, we check for the accuracy of our model by providing a test dataset to it.
- Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment

- The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.
- If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system.
- But before deploying the project, we will check whether it is improving its performance using available data or not.
- The deployment phase is similar to making the final report for a project.

Machine learning models: Classification, Regression, Clustering, Reinforcement

- Machine Learning models are very powerful resources that automate multiple tasks and make them more accurate and efficient.
- ML handles new data and scales the growing demand for technology with valuable insight.
- It improves the performance over time.
- This cutting-edge technology has various benefits such as faster processing or response, enhancement of decision-making, and specialized services.
- A model of machine learning is a set of programs that can be used to find the pattern and make a decision from an unseen dataset.
- These days NLP (Natural language Processing) uses the machine learning model to recognize the unstructured text into usable data and insights.
- You may have heard about image recognition which is used to identify objects such as boy, girl, mirror, car, dog, etc.
- A model always requires a dataset to perform various tasks during training.
- In training duration, we use a machine learning algorithm for the optimization process to find certain patterns or outputs from the dataset based upon tasks.

- Machine learning models can be broadly categorized into four main paradigms based on the type of data and learning goals:

1. Supervised Models

- Supervised learning is the study of algorithms that use labeled data in which each data instance has a known category or value to which it belongs.
- This results in the model to discover the relationship between the input features and the target outcome.

1.1 Classification

- The classifier algorithms are designed to indicate whether a new data point belongs to one or another among several predefined classes.

- Imagine when you are organizing emails into spam or inbox, categorizing images as cat or dog, or predicting whether a loan applicant is a credible borrower.
- In the classification models, there is a learning process by the use of labeled examples from each category.
- In this process, they discover the correlations and relations within the data that help to distinguish class one from the other classes.
- After learning these patterns, the model is then capable of assigning these class labels to unseen data points.

Common Classification Algorithms:

- **Logistic Regression**: A very efficient technique for the classification problems of binary nature (two types, for example, spam/not spam).

- **Support Vector Machine (SVM):** Good for tasks like classification, especially when the data has a large number of features.

- **Decision Tree:** Constructs a decision tree having branches and proceeds to the class predictions through features.

- **Random Forest:** The model generates an "ensemble" of decision trees that ultimately raise the accuracy and avoid overfitting (meaning that the model performs great on the training data but lousily on unseen data).

- **K-Nearest Neighbors (KNN):** Assigns a label of the nearest neighbors for a given data point.

1.2 Regression

- Regression algorithms are about forecasting of a continuous output variable using the input features as their basis.

- This value could be anything such as predicting real estate prices or stock market trends to anticipating customer churn (how likely customers stay) and sales forecasting.
- Regression models make the use of features to understand the relationship among the continuous features and the output variable.
- That is, they use the pattern that is learned to determine the value of the new data points.

- **Common Regression Algorithms**

- **Linear Regression:** Fits depth of a line to the data to model for the relationship between features and the continuous output.
- **Polynomial Regression:** Similiar to linear regression but uses more complex polynomial functions such as quadratic, cubic, etc, for accommodating non-linear relationships of the data.
- **Decision Tree Regression:** Implements a decision tree-based algorithm that predicts a continuous output variable from a number of branching decisions.
- **Random Forest Regression:** Creates one from several decision trees to guarantee error-free and robust regression prediction results.
- **Support Vector Regression (SVR):** Adjusts the Support Vector Machine ideas for regression tasks, where we are trying to find one hyperplane that most closely reflects continuous output data.

2. Unsupervised Models

- Unsupervised learning involves a difficult task of working with data which is not provided with pre-defined categories or label.

2.1 Clustering

- Visualize being given a basket of fruits with no labels on them. The fruits clustering algorithms are to group them according to the inbuilt similarities.
- Techniques like K-means clustering are defined by exact number of clusters ("red fruits" and "green fruits") and then each data point (fruit) is assigned

to the cluster with the highest similarity within based on features (color, size, texture).

- Contrary to this, hierarchical clustering features construction of hierarchy of clusters which makes it easier to study the system of groups.
- Spatial clustering algorithm Density-Based Spatial Clustering of Applications with Noise (DBSCAN) detects groups of high-density data points, even in those areas where there is a lack of data or outliers.

4. Reinforcement learning Models

- Reinforcement learning takes a dissimilar approach from supervised learning and unsupervised learning.
- Different from supervised learning or just plain discovery of hidden patterns, reinforcement learning adopt an agent as it interacts with the surrounding and learns.
- This agent is a learning one which develops via experiment and error, getting rewarded for the desired actions and punished for the undesired ones.
- The main purpose is to help players play the game that can result in the highest rewards.