MACHINE LEARNING

UNIT I

Introduction:

Machine Learning is an AI technique that teaches computers to learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data . The algorithms adaptively improve their performance as the number of samples available for learning increases. Deep learning is a specialized form of machine learning.

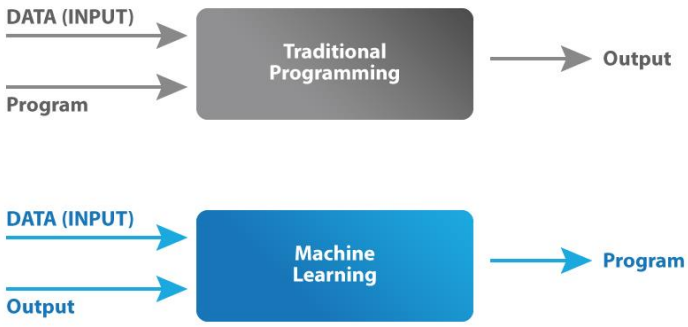
Machine learning is a powerful tool that can be used to solve a wide range of problems. It allows computers to learn from data, without being explicitly programmed. This makes it possible to build systems that can automatically improve their performance over time by learning from their experiences.

Machine learning is an application of artificial intelligence that uses statistical techniques to enable computers to learn and make decisions without being explicitly programmed. It is predicated on the notion that computers can learn from data, spot patterns, and make judgments with little assistance from humans.

It is a subset of Artificial Intelligence. It is the study of making machines more human-like in their behavior and decisions by giving them the ability to learn and develop their own programs. This is done with minimum human intervention, i.e., no explicit programming. The learning process is automated and improved based on the experiences of the machines throughout the process.

Good quality data is fed to the machines, and different algorithms are used to build ML models to train the machines on this data. The choice of algorithm depends on the type of data at hand and the type of activity that needs to be automated.

In traditional programming, we would feed the input data and a well-written and tested program into a machine to generate output. When it comes to machine learning, input data, along with the output, is fed into the machine during the learning phase, and it works out a program for itself. To understand this better, refer to the illustration below,



Traditional programming uses known algorithms to produce results from data:

Data + Algorithms = Results

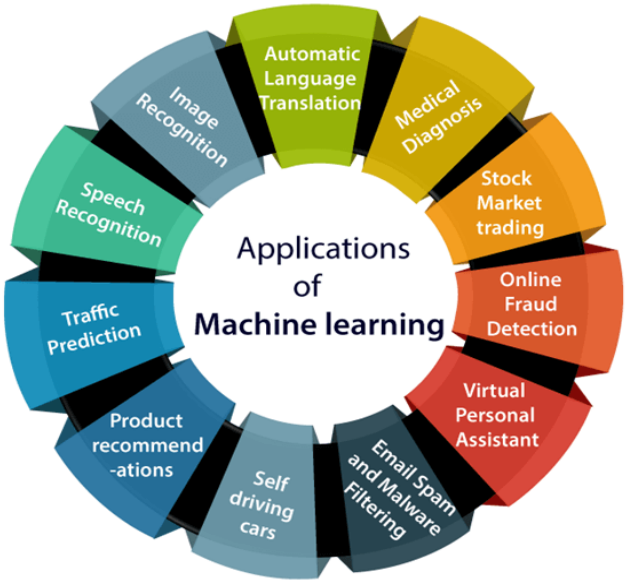
Machine learning creates new algorithms from data and results:

Data + Results = Algorithms

Uses/Applications of Machine Learning:

- Machine learning is widely used in many industries, including healthcare, finance, healthcare, manufacturing, marketing, finance, energy, insurance, human resources, government, digital media and entertainment, transportation, manufacturing, automobile, art & creativity and e-commerce. By learning machine learning, we can open up a wide range of career opportunities in these fields.
- Machine learning can be used to build intelligent systems that can make decisions and predictions based on data. This can help organizations make better decisions, improve their operations, and create new products and services.
- Machine learning is an important tool for data analysis and visualization. It allows us to extract insights and patterns from large datasets, which can be used to understand complex systems and make informed decisions.
- Machine learning is a rapidly growing field with many exciting developments and research opportunities. By learning machine learning, we can stay up-to-date with the latest research and developments in the field.

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:



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.

LANGUAGE FOR MACHINE LEARNING:

Programming languages for machine learning includes,

(1) Python

- Developed by Guido Van Rossum in 1991, Python is the most popular and widely-used programming language for machine learning.
- Python is a dynamic, high-level, interactive, multi-paradigm, scripting, object-oriented, high-level, procedural language.
- It is easy to write, the readable and understandable language has fewer and easy-to-read English syntax.
- It does not require saving a code and processing the code later.
- The codes are directly re-usable by importing the module or the package.
- Python is interpretable, scalable, and adaptable.
- It has various in-built libraries, modules, and packages.

- It is portable i.e can run on any operating system including such as Windows, Linux, UNIX, and Macintosh.
- Python is vastly used in all the spheres of machine learning problems. It can also be integrated with SQL. It is used for data mining and wrangling, data visualizations.
- It provides a variety of deep and machine learning frameworks.
- Python is the preferred programming language of choice for machine learning for some of the giants in the IT world including Google, Instagram, Facebook, Dropbox, Netflix, Walt Disney, YouTube, Uber, Amazon, and Reddit.

Various libraries exist for machine learning tasks such as:

- **Numpy**: It is designed for scientific computing.
- **Pandas**: It is applied for data analysis.
- **Scipy**: It is helpful for scientific & technical computing.
- **Sklearn**: It is for the implementation of machine learning algorithms.
- **NLTK, Spacy**: These are applicable for working with textual data.
- **Sci-Kit image and OpenCV**: These are used to work with the image data.
- **Librosa**: It is used for working with audio data such as audio pre-processing, and audio and music analysis.

(2) R

R language can also be used by non-programmer including data miners, data analysts, and statisticians. R programming is a preferred choice for machine learning applications that use a lot of statistical data. With user-friendly IDE's like RStudio and various tools to draw graphs and manage libraries. R language provides a variety of tools to train and evaluate machine learning algorithms for predicting future events making machine learning easy and approachable.

- R is an open-source programming language making it a highly cost-effective choice for machine learning projects of any size.
- R is considered a powerful choice for machine learning because of the machine learning techniques it provides. It provides data visualisation, data sampling, data analysis, model evaluation, supervised/unsupervised machine learning techniques.
- R is highly flexible and also offers cross-platform compatibility.

Various libraries exist for machine learning tasks such as:

- **MICE**: It is specially designed to work with missing values.
- **Tidyr and Dplyr**: Both of these are applied for data manipulation that is for clearing and organizing data. It is fast and reduces clutter in coding.
- **Ggplot2**: It is the data visualization package for illustrating the data.
- **Tidyquant**: It is applied for business and financial analysis.
- **CARET, PARTY, and randomFOREST**: It is especially used for working with classification and regression problems, PARTY and rpart for creating recursive data partitions, and randomFOREST for creating decision trees.

(3) Java

- Java is gaining popularity among machine learning engineers who hail from a Java development background as they don't need to learn a new programming language like Python or R to implement machine learning.
- Java has plenty of third party libraries for machine learning.
- Java makes application scaling easier for the development of large and complex machine learning applications from scratch.
- Java Virtual Machine is one of the best platforms for machine learning as engineers can write the same code on multiple platforms. JVM also helps machine learning engineers create custom tools at a rapid pace and has various IDE's that help improve overall productivity. Java works best for speed-critical machine learning projects as it is fast executing.
- Java Virtual Machine (JVM) allows the developers to write codes that are identical across multiple platforms. Java applications are scalable and are also built fastly.
- Java is useful for various processes in data science such as cleaning data, data importation and exportation, statistical analysis, deep learning, NLP, and data visualization.

The popular tools and libraries for machine learning are:

- **Weka**: It is a free, portable library having many general-purpose utilities for machine learning including algorithms, data mining, data analysis, and predictive modeling.
- **JavaML**: is an in-built machine learning library that provides a collection of machine learning algorithms implemented in Java. It is a Java API with a user-friendly interface to implement the compilation of machine learning and data mining algorithms.

- **Deeplearning4j**: It is a distributed deep learning library and is open source. It offers a computing framework supporting machine learning algorithms and high processing capabilities.
- **Massive Online Analysis**: An open-source software, it is useful for real-time analytics and applied for data mining on data streams on a real-time basis.

(4) Julia

- Julia is an open-source, high-level, general-purpose dynamic programming language.
- It is both functional and object-oriented and is favored by developers as it has easy syntax. It is accessible and easily understandable.
- This scripting language is useful for high-performance numerical analysis and computational statistics.
- Julia is powering machine learning applications at big corporations like Apple, Disney, Oracle, and NASA

Julia has the following powerful tools:

- **Flux**: It comes with the same functionality as Tensorflow. Lightweight ML library, useful tools to help us to use the full power of Julia
- **MLBase.jl**: It is used for data processing and manipulation, evaluation of models, cross-validation, and model tuning.
- **TensorFlow.jl**: It offers the option to express computations as data flow graphs.
- **ScikitLearn.jl**: It allows preprocessing, clustering, model selection.

(5) LISP

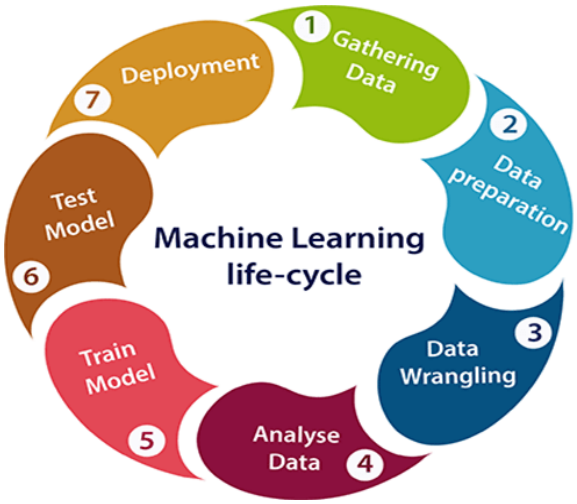
- List Processing (LISP) was developed by the American computer scientist and cognitive scientist, John McCarthy in 1958. This artificial intelligence focussed language is the second oldest programming language and is very much in use.
- LISP is a dynamically written programming language that has led to the creation of many popular machine learning programming languages including Python, Julia, and Java.
- LISP has a fast prototyping capacity, is very flexible, and aids in easy and dynamic object creation. It is specifically useful for logic problems and machine learning.
- LISP has features of a domain-specific language that is embedded with code, compiled, and can run code in 30+ programming languages.
- The first AI chatbot ELIZA was developed using LISP and even today machine learning practitioners can use it to create chatbots for eCommerce.

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**
- **Analyse 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:

- **DataExploration:**
  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.
- **Datapre-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.

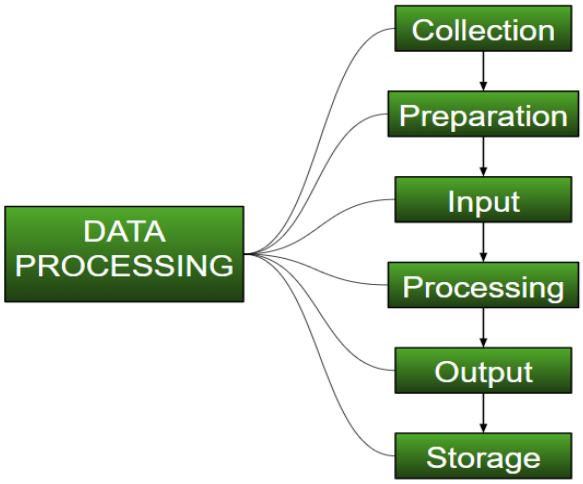
7. Deployment

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.

DATA PROCESSING:

Data Processing is the task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modeling, and statistical knowledge, this entire process can be automated. The output of this complete process can be in any desired form like graphs, videos, charts, tables, images, and many more, depending on the task we are performing and the requirements of the machine. This might seem to be simple but when it comes to massive organizations like Twitter, Facebook, Administrative bodies like Parliament, UNESCO, and health sector organizations, this entire process needs to be performed in a very structured manner. So, the steps to perform are as follows:



- **Collection:**
  The most crucial step when starting with ML is to have data of good quality and accuracy. Data can be collected from any authenticated source like data.gov.in, Kaggle or UCI dataset repository. For example, while preparing for a competitive exam, students study from the best study material that they can access so that they learn the best to obtain the best results. In the same way, high-quality and accurate data will make the learning process of the model easier and better and at the time of testing, the model would yield state-of-the-art results.
  A huge amount of capital, time and resources are consumed in collecting data. Organizations or researchers have to decide what kind of data they need to execute their tasks or research.
  Example:
  Working on the Facial Expression Recognizer, needs numerous images having a variety of human expressions. Good data ensures that the results of the model are valid and can be trusted upon.

- **Preparation:**
  The collected data can be in a raw form which can't be directly fed to the machine. So, this is a process of collecting datasets from different sources, analyzing these datasets and then constructing a new dataset for further processing and exploration. This preparation can be performed either manually or from the automatic approach. Data can also be prepared in numeric forms also which would fasten the model's learning.
  Example:
  An image can be converted to a matrix of N X N dimensions, the value of each cell will indicate the image pixel.
- **Input:**
  Now the prepared data can be in the form that may not be machine-readable, so to convert this data to the readable form, some conversion algorithms are needed. For this task to be executed, high computation and accuracy is needed.
  Example:
  Data can be collected through the sources like MNIST Digit data(images), Twitter comments, audio files, video clips.
- **Processing:**
  This is the stage where algorithms and ML techniques are required to perform the

instructions provided over a large volume of data with accuracy and optimal computation.

- **Output:**
  In this stage, results are procured by the machine in a meaningful manner which can be inferred easily by the user. Output can be in the form of reports, graphs, videos, etc
- **Storage:**
  This is the final step in which the obtained output and the data model data and all the useful information are saved for future use.

Advantages of data processing in Machine Learning:

1. **Improved model performance**:
   Data processing helps improve the performance of the ML model by cleaning and transforming the data into a format that is suitable for modeling.
2. **Better representation of the data**:
   Data processing allows the data to be transformed into a format that better represents the underlying relationships and patterns in the data, making it easier for the ML model to learn from the data.
3. **Increased accuracy**:
   Data processing helps ensure that the data is accurate, consistent, and free of errors, which can help improve the accuracy of the ML model.

Disadvantages of data processing in Machine Learning:

1. **Time-consuming**:
   Data processing can be a time-consuming task, especially for large and complex datasets.
2. **Error-prone:**
   Data processing can be error-prone, as it involves transforming and cleaning the data, which can result in the loss of important information or the introduction of new errors.
3. **Limited understanding of the data:**
   Data processing can lead to a limited understanding of the data, as the transformed data may not be representative of the underlying relationships and patterns in the data.