Chap 01 Introduction to Machine Learning

Introduction: -

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information.

Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

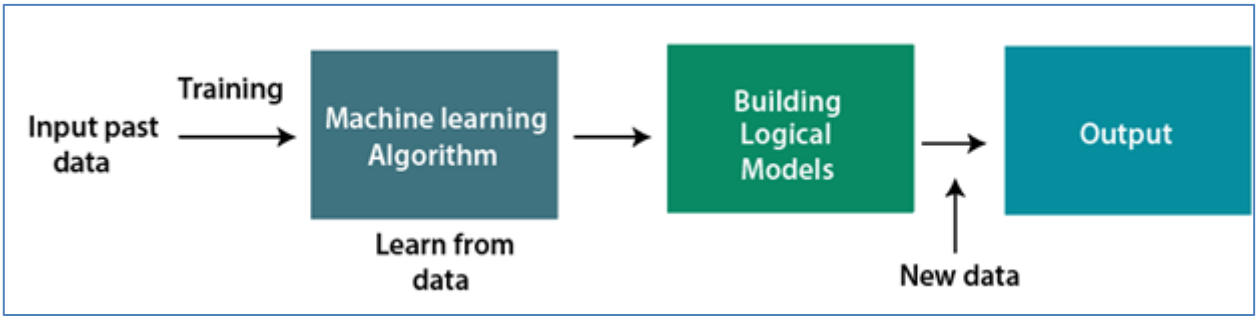
What is Machine Learning:

- 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.
- Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.
- The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as:
- "Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed."
- With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed.
- A machine has the ability to learn if it can improve its performance by gaining more data.
- "Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to "self-learn" from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. In short, machine learning algorithms and models learn through experience."

How does Machine Learning work:

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



💼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:

Machine Learning (ML) has become increasingly important and relevant due to several factors and needs across various industries and domains. Here are some key reasons why there is a significant need for machine learning**:**

1. ****Big Data Handling:**** With the explosion of digital data, traditional methods of data analysis and processing are often inadequate. ML algorithms can efficiently process and extract insights from massive datasets that would be impractical to handle manually.
2. ****Complexity and Scale:**** Many real-world problems have become too complex for human programmers to solve using traditional rule-based systems. ML can tackle intricate tasks by identifying patterns and relationships within data.
3. ****Predictive Analytics:**** Organizations need to predict future outcomes accurately to make informed decisions. ML models can analyze historical data to forecast trends and make predictions, whether in stock markets, weather forecasts, or customer behavior.
4. ****Automation and Efficiency:**** ML can automate repetitive tasks, improving efficiency and reducing human intervention. This is crucial in areas like manufacturing, supply chain optimization, and customer service.
5. ****Personalization:**** Customizing experiences for users or customers is vital in various industries. ML can analyze individual preferences and behaviors to provide personalized recommendations and content, as seen in streaming services, e-commerce, and marketing.
6. **Fraud Detection and Cybersecurity:** Identifying fraudulent activities or potential security breaches is a constant challenge. ML algorithms can learn from historical data to recognize patterns of fraud or suspicious behavior.
7. ****Healthcare Insights:**** ML can analyze vast amounts of medical data, aiding in disease diagnosis, treatment optimization, drug discovery, and even predicting potential outbreaks.
8. ****Natural Language Processing (NLP):**** With the rise of chatbots, virtual assistants, and language translation services, NLP powered by ML is crucial for enabling machines to understand and communicate in human languages.
9. ****Image and Video Analysis:**** ML-driven computer vision allows machines to interpret visual data, enabling applications like facial recognition, object detection, autonomous vehicles, and quality control in manufacturing.
10. ****Scientific Research:**** ML can accelerate scientific discoveries by analyzing complex data sets, simulating scenarios, and identifying potential patterns or solutions that might be missed by traditional methods.

Advantages of Machine Learning:

- **Automation of Tasks:** ML automates repetitive tasks and processes, saving time and effort, and reducing the need for human intervention.
- **Handling Complex Data:** ML can analyze vast and complex data sets, extracting valuable insights and patterns that might be challenging for humans to discern.
- **Continuous Improvement**: ML models can learn from new data, leading to continuous improvement and adaptation over time.
- **Predictive Analytics:** ML can predict future outcomes and trends based on historical data, helping businesses make informed decisions.

- **Personalization:** ML enables systems to tailor experiences to individual preferences, enhancing user satisfaction in various application.

Disadvantages of Machine Learning:

- **Data Dependency:** ML models require a substantial amount of quality data for training. Insufficient or biased data can lead to inaccurate or biased results.
- **Lack of Interpretability:** Some complex ML models are difficult to interpret, making it challenging to understand how they arrive at specific decisions.
- **Overfitting**: If an ML model is too complex, it may perform exceptionally well on training data but poorly on new, unseen data.
- **High Initial Costs:** Implementing ML can be expensive due to the need for specialized hardware, software, and skilled professionals.
- **Ethical Concerns:** ML models can perpetuate biases present in training data, leading to ethical and fairness issues.
- **Limited Creativity and Intuition:** ML lacks human creativity and intuition, making it unsuitable for tasks that require these qualities.

Use

Healthcare:

- Disease diagnosis and detection.
- Personalized treatment plans.
- Drug discovery and development.
- Predictive analytics for patient outcomes.

Finance:

- Credit scoring and risk assessment.
- Fraud detection and prevention.
- Algorithmic trading and investment strategies.
- Customer sentiment analysis for trading decisions.

Retail and E-commerce:

- Product recommendations and personalized shopping experiences.
- Price optimization.
- Demand forecasting and inventory management.
- Fraud detection in online transactions.

Marketing and Advertising:

- Customer segmentation and targeting.
- Ad campaign optimization.
- Sentiment analysis for brand perception.
- Click-through rate prediction.

Manufacturing and Industry:

- Predictive maintenance for machinery and equipment.
- Quality control and defect detection.
- Supply chain optimization.

Classification of Machine Learning:

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

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

1 Supervised Learning

- Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.
- The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.
- The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering.

(A) How Supervised Learning Works?

Suppose we have a dataset of different types of shapes which rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

includes square,

- If the given shape has four sides, and all the sides are equal, then it will be labelled as a Square.
- If the given shape has three sides, then it will be labelled as a triangle.
- If the given shape has six equal sides, then it will be labelled as hexagon.

Now, after training, we test our model using the test set, and the task of the model is to identify the shape. The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

Following are the steps involved in Supervised Learning:

- First Determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training dataset, test dataset, and validation dataset. o Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- o Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- o Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets. o Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

💎 Machine learning Life cycle:

- Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed.
- 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.

Phase of ml

- Gathering Data
- Data preparation
- Data Wrangling
- Analyse Data
- Train the model
- Test the model
- Deployment

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.

- **Identify various data sources**
- **Collect data**
- **Integrate the data obtained from different sources**

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.

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.

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.

- **Missing Values**
- **Duplicate data**
- **Invalid data**
- **Noise**

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.

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.

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.

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 Preprocessing: -

Data preprocessing is a crucial step in the machine learning pipeline, responsible for transforming raw data into a format that is suitable for training machine learning models. It involves a series of techniques that clean, transform, and prepare the data to enhance its quality and ensure that it is appropriate for the chosen machine learning algorithm.

Why Data Preprocessing is Essential:

1. Improves Data Quality: Data preprocessing enhances the quality of data by identifying and addressing issues such as missing values, outliers, inconsistencies, and redundancies.
2. Enhances Model Performance: Clean and well-prepared data leads to better model performance, improving accuracy, generalizability, and reducing the risk of overfitting.
3. Reduces Model Training Time: Preprocessed data reduces the computational burden on machine learning algorithms, leading to faster training and evaluation.
4. Enables Efficient Feature Engineering: Preprocessing facilitates feature engineering, the process of transforming raw data into meaningful features that are more informative for the model.

Common steps in data preprocessing:

Data Cleaning

Data Cleaning uses methods to handle incorrect, incomplete, inconsistent, or missing values. Some of the techniques for Data Cleaning include -

- **Handling Missing Values**
    - Input data can contain missing or **NULL** values, which must be handled before applying any Machine Learning or Data Mining techniques.
    - Missing values can be handled by many techniques, such as removing rows/columns containing NULL values and imputing NULL values using mean, mode, regression, etc.
- **De-noising**
    - De-noising is a process of removing noise from the data. Noisy data is meaningless data that is not interpretable or understandable by machines or humans. It can occur due to data entry errors, faulty data collection, etc.
    - De-noising can be performed by applying many techniques, such as binning the features, using regression to smoothen the features to reduce noise, clustering to detect the outliers, etc

Data Integration:

- **Combining Data**: If the data comes from multiple sources, integrate it into a single dataset.
- **Resolving Data Inconsistencies**: Address inconsistencies in naming conventions, units, or formats between different datasets.

Data Transformation:

- **Encode categorical variables**: Convert categorical variables into numerical representations. This can be done using techniques like one-hot encoding or label encoding.
- **Feature scaling**: Standardize or normalize numerical features to ensure that they contribute equally to the model. Common methods include Min-Max scaling or Z-score normalization.
- **Log transformation**: Apply logarithmic transformations to skewed data to make it more normally distributed.

Data Reduction

Data Reduction is used to reduce the volume or size of the input data. Its main objective is to reduce storage and analysis costs and improve storage efficiency. A few of the popular techniques to perform Data Reduction include -

- **Dimensionality Reduction** - It is the process of reducing the number of features in the input dataset. It can be performed in various ways, such as selecting features with the highest importance, Principal Component Analysis (PCA), etc.
- **Numerosity Reduction** - In this method, various techniques can be applied to reduce the volume of data by choosing alternative smaller representations of the data. For example, a variable can be approximated by a regression model, and instead of storing the entire variable, we can store the regression model to approximate it.
- **Data Compression** - In this method, data is compressed. Data Compression can be lossless or lossy depending on whether the information is lost or not during compression.

Data Splitting:

Training and Testing Sets: Split the dataset into training and testing sets to assess the model's performance on unseen data.

Cross-Validation: For more robust evaluation, use techniques like k-fold cross-validation.

Parameters	Supervised machine learning	Unsupervised machine learning
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data that is not labeled
Computational Complexity	Simpler method	Computationally complex
Accuracy	Highly accurate	Less accurate
No. of classes	No. of classes is known	No. of classes is not known
Data Analysis	Uses offline analysis	Uses real-time analysis of data
Algorithms used	Linear and Logistics regression, Random forest, Support Vector Machine, Neural Network, etc.	K-Means clustering, Hierarchical clustering, Apriori algorithm, etc.
Output	Desired output is given.	Desired output is not given.
Training data	Use training data to infer model.	No training data is used.
Complex model	It is not possible to learn larger and more complex models than with supervised learning.	It is possible to learn larger and more complex models with unsupervised learning.
Model	We can test our model.	We can not test our model.
Called as	Supervised learning is also called classification.	Unsupervised learning is also called clustering.
Example	Example: Optical character recognition.	Example: Find a face in an image.