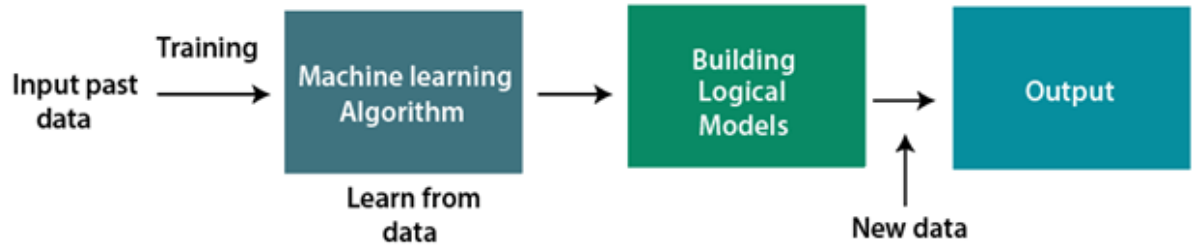
Introduction to Machine Learning

A subset of artificial intelligence known as machine learning focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences. Arthur Samuel first used the term "machine learning" in 1959

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.



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.

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.

❖ Challenges of Machine Learning

- Biased training data can lead to unfair or discriminatory outcomes.
- ML systems risk data breaches and raise ethical concerns about personal data usage.
- Complex models often lack transparency, making decision-making hard to explain.
- Automation can replace jobs, requiring workforce reskilling and adaptation.
- ML models demand high computational power and efficient optimization to scale effectively.

Artificial Intelligence	Machine Learning
Artificial intelligence (AI), where intelligence is defined as the acquisition of knowledge and the ability to apply knowledge.	Machine Learning (ML) means gaining skill or knowledge.
The goal is not accuracy but to increase the chance of business success.	The goal is to increase accuracy, but it does not care about business success
This leads to the development of a system that mimics a human being to behave in situations.	It involves designing self-learning algorithms.
The aim is to simulate natural intelligence to solve tough issues	The aim is to learn from the data on the specific task to maximize the performance of the machine.
Artificial Intelligence is a decision maker	ML enables the system to learn new things from the data.
It works as a smart working computer program	It is a simple concept machine that takes data and learns from data.
AI finds optimal solution	ML finds only solution, whether it is optimal or not.

History of Machine Learning

1. **1950s-1960s** – Alan Turing introduced the "Turing Test," and Arthur Samuel developed the first machine learning program (checkers-playing AI).
2. **1970s-1980s** – Early neural networks and decision trees emerged, but AI research faced funding cuts ("AI winter").
3. **1990s-2000s** – Support Vector Machines (SVM), Random Forests, and deep learning gained popularity, improving ML performance.
4. **2010s-Present** – Breakthroughs in deep learning, reinforcement learning, and transformer models led to AI-powered applications like ChatGPT, self-driving cars, and medical diagnostics.

Examples of Machine Learning Applications

1. **Healthcare** – Disease prediction, medical image analysis, and drug discovery (e.g., AI diagnosing cancer from scans).
2. **Finance** – Fraud detection, credit scoring, and algorithmic trading (e.g., detecting suspicious transactions).
3. **E-commerce** – Recommendation systems (e.g., Amazon suggesting products based on browsing history).
4. **Autonomous Vehicles** – Self-driving cars use ML for object detection and navigation (e.g., Tesla Autopilot).

5. **Virtual Assistants** – AI-powered chatbots like Siri, Alexa, and Google Assistant respond to voice commands.

Types of Machine Learning

1. **Supervised Learning** – The model learns from labeled data, where each input has a corresponding correct output. Example: Email spam detection (spam or not spam).
2. **Unsupervised Learning** – The model finds patterns in unlabeled data without specific outputs. Example: Customer segmentation in marketing.
3. **Semi-Supervised Learning** – A combination of labeled and unlabeled data, useful when labeling is expensive. Example: Fraud detection with few labeled fraud cases.
4. **Reinforcement Learning** – The model learns through rewards and penalties by interacting with an environment. Example: Self-learning AI in video games or robotics.

Machine Learning Life Cycle :

1. **Problem Definition** – Define the goal of the ML model (e.g., predicting house prices).
2. **Data Collection** – Gather relevant data from various sources like databases, APIs, or web scraping.
3. **Data Preprocessing** – Clean and prepare data by handling missing values, removing duplicates, and normalizing features.
4. **Model Selection** – Choose a suitable ML algorithm (e.g., Decision Trees, Neural Networks).
5. **Training the Model** – Feed data into the model to learn patterns.
6. **Model Evaluation** – Test the model's accuracy and performance using testing data.
7. **Hyperparameter Tuning** – Adjust model parameters for better accuracy.
8. **Deployment** – Integrate the trained model into a real-world application.
9. **Monitoring & Maintenance** – Continuously improve the model as new data becomes available.

Dataset for Machine Learning :

A dataset is a collection of data used to train and test ML models. It consists of:

- **Features (Input Variables)** – Characteristics used for predictions (e.g., age, income for loan approval).
- **Labels (Target Variable)** – The output the model predicts (e.g., "approved" or "denied" for a loan).
- **Types of Datasets**:
    - **Structured Data** – Organized in tables (e.g., spreadsheets, databases).
    - **Unstructured Data** – Raw data like images, videos, text.

- **Public Datasets** – Available for ML training, such as ImageNet (images), Kaggle datasets, UCI ML Repository.

Data Pre-processing :

Before training, raw data needs cleaning and transformation:

1. **Handling Missing Data** – Replace missing values with averages (mean), most frequent values, or remove incomplete records.
2. **Removing Duplicates** – Eliminates redundant data to avoid bias.
3. **Normalization & Scaling** – Converts data to a standard format to improve accuracy (e.g., converting heights from cm to meters).
4. **Encoding Categorical Data** – Converts text labels into numbers (e.g., "Male" = 1, "Female" = 0).
5. **Feature Engineering** – Creating new relevant features to enhance model performance.

Training vs. Testing in ML

- **Training Data** – The dataset used to teach the ML model patterns. The model adjusts itself based on this data.
- **Testing Data** – A separate dataset used to check how well the trained model performs on unseen data.
- **Validation Data** – Sometimes used to fine-tune hyperparameters before final testing.

Example:

- Training data: 80% of the dataset (used for learning).
- Testing data: 20% of the dataset (used for evaluation).

Positive and Negative Class in ML

- **Positive Class** – The outcome of interest (e.g., detecting a disease: "disease present" = Positive).
- **Negative Class** – The absence of the condition (e.g., "no disease" = Negative).
- Used in **classification problems** such as spam detection:
    - **Spam email** = Positive Class
    - **Not spam** = Negative Class

Cross-Validation in ML :-

Cross-validation is a technique to ensure the ML model performs well on unseen data by splitting the dataset into multiple parts.

1. **K-Fold Cross-Validation** – The data is divided into "K" subsets. The model trains on (K-1) parts and tests on the remaining part, repeating the process K times.

2. **Leave-One-Out Cross-Validation (LOOCV)** – Each data point is used once as a test case while the rest are for training.
3. **Stratified Cross-Validation** – Ensures each fold has the same proportion of positive and negative classes.

Why is it useful?

- Reduces overfitting (memorizing training data instead of learning patterns).
- Provides a more reliable accuracy estimate of the ML model.

Supervised Learning :-

Supervised learning is a machine learning technique where a model learns from labeled data. It finds patterns in the data by mapping input features to output labels and makes predictions for unseen data.

Supervised learning is divided into two main types:

1. **Regression Algorithms** – Used for predicting continuous numerical values.
2. **Classification Algorithms** – Used for categorizing data into predefined labels.

Regression Algorithms (For Continuous Data Prediction)

Regression algorithms are used when the target variable is continuous (e.g., predicting stock prices, house prices, temperature, etc.).

1. Linear Regression

- **Definition:** Linear regression is the simplest form of regression that models the relationship between an independent variable (**X**) and a dependent variable (**Y**) using a straight-line equation: Y=mX+bY = mX + bY=mX+b where **m** is the slope (how much Y changes when X changes) and **b** is the intercept (the starting value of Y when X = 0).
- **Use Case:** Predicting sales based on advertising expenditure.

2. Polynomial Regression

- **Definition:** Extends linear regression by fitting a polynomial equation to the data to model nonlinear relationships: Y=aX2+bX+cY = aX^2 + bX + cY=aX2+bX+c This allows it to capture more complex trends in data.
- **Use Case:** Predicting the trajectory of an object in physics.

3. Ridge Regression

- **Definition:** A regularized form of linear regression that prevents overfitting by adding an L2 penalty term (sum of squared coefficients) to the cost function: ∑(Yi−Y^i)2+λ∑w2\sum (Y_i - \hat{Y}_i)^2 + \lambda \sum w^2∑(Yi−Y^i)2+λ∑w2 The penalty term **λ** helps keep the model simpler and prevents it from fitting noise in the data.
- **Use Case:** Used in financial forecasting when there is multicollinearity (high correlation between independent variables).

4. Lasso Regression

- **Definition:** Similar to Ridge Regression but uses an L1 penalty (sum of absolute values of coefficients) instead of L2. It can shrink some coefficients to exactly zero, effectively performing **feature selection**: ∑(Yi−Y^i)2+λ∑|w|\sum (Y_i - \hat{Y}_i)^2 + \lambda \sum |w|∑(Yi−Y^i)2+λ∑|w|
    - **Difference from Ridge Regression:** Lasso can eliminate unimportant features by setting their weights to zero, whereas Ridge just reduces their impact.
- **Use Case:** Selecting important features in medical research prediction models.

5. Support Vector Regression (SVR)

- **Definition:** Based on **Support Vector Machines (SVM)**, SVR finds the best-fit hyperplane that allows most data points to be within a certain margin while minimizing error. It ignores small deviations and focuses on general trends.
- **Use Case:** Forecasting stock prices or predicting real estate values.

Classification Algorithms (For Categorical Data Prediction)

Classification algorithms categorize input data into discrete classes (e.g., spam detection, disease diagnosis, etc.).

1. Logistic Regression

- **Definition:** Unlike linear regression, logistic regression predicts **probabilities** rather than continuous values. It uses the **sigmoid function** to convert output into a probability between 0 and 1: P(Y=1)=11+e−(b0+b1X)P(Y=1) = \frac{1}{1 + e^{-(b_0 + b_1X)}}P(Y=1)=1+e−(b0+b1X)1 If **P(Y=1) > 0.5**, the outcome is classified as 1 (positive class); otherwise, it's classified as 0 (negative class).
- **Use Case:** Used for email spam classification (spam or not spam).

2. Decision Trees

- **Definition:** A tree-based algorithm that makes decisions by splitting the dataset into smaller subsets based on conditions. Each internal node represents a **decision rule**, and each leaf node represents an **outcome**.
- **Use Case:** Customer segmentation for targeted marketing campaigns.

3. Random Forest

- **Definition:** An **ensemble learning method** that combines multiple **decision trees** to improve accuracy and reduce overfitting. It selects the most common prediction (for classification) or the average prediction (for regression).
- **Use Case:** Credit risk assessment in banking.

4. Support Vector Machines (SVM)

- **Definition:** SVM finds the **optimal hyperplane** that best separates different classes by maximizing the margin between them. It is effective for high-dimensional data and robust against outliers.
- **Use Case:** Facial recognition and handwriting detection.

5. Naïve Bayes

- **Definition:** A probabilistic classifier based on **Bayes' theorem**, assuming that features are **independent** of each other: P(A|B)=P(B|A)P(A)P(B)P(A|B) = \frac{P(B|A) P(A)}{P(B)}P(A|B)=P(B)P(B|A)P(A) It is **fast** and works well with **large datasets**.
- **Use Case:** Sentiment analysis in social media (positive or negative sentiment).

6. K-Nearest Neighbors (KNN)

- **Definition:** A non-parametric algorithm that classifies a new data point based on the **majority class** of its **K nearest neighbors**. The value of **K** determines how many neighbors to consider.
- **Use Case:** Used in **handwritten digit recognition** (e.g., MNIST dataset).

Unsupervised Learning

What is Unsupervised Learning?

Unsupervised learning is a machine learning technique where models learn patterns and structures from **unlabeled data** without explicit supervision. Unlike **supervised learning**, where models learn from labeled data, unsupervised learning **identifies hidden patterns, relationships, or groupings** in data.

🔹 **Example:** A company wants to segment customers into different groups based on purchasing behavior. Since there are no predefined labels, an unsupervised learning algorithm can cluster customers into groups with similar traits.

Types of Unsupervised Learning

1.Clustering

- Clustering is the process of grouping similar data points together.
- It is useful for customer segmentation, anomaly detection, and organizing datasets.

2.Dimensionality Reduction

- Reduces the number of features while retaining essential information.
- Helps in data visualization and improving model performance.

Clustering Algorithms (For Grouping Similar Data)

1. K-Means Clustering

- **Definition:**
  K-Means is a partition-based clustering algorithm that divides data into **K** clusters. It

assigns each data point to the **nearest cluster centroid** and updates the centroids iteratively.

- **How it Works:**
    1. Choose the number of clusters (**K**).
    2. Randomly initialize **K** cluster centroids.
    3. Assign each data point to the **nearest centroid**.
    4. Recalculate the centroids based on the assigned points.
    5. Repeat steps 3 and 4 until convergence.
- **Use Case:** Customer segmentation in marketing.

2. Hierarchical Clustering

- **Definition:**
  Unlike K-Means, hierarchical clustering creates a **tree-like structure (dendrogram)** to represent nested clusters.
- **Types:**
    - **Agglomerative Clustering** (Bottom-Up) – Each data point starts as its own cluster, and clusters merge step by step.
    - **Divisive Clustering** (Top-Down) – Starts with a single cluster and splits it recursively.
- **Use Case:** Organizing genes in bioinformatics.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Definition:**
  DBSCAN groups data points **based on density** rather than predefined cluster numbers. It can **detect outliers** as noise.
- **Advantages:**
    - Can detect clusters of **arbitrary shapes**.
    - Works well when clusters vary in size and density.
- **Use Case:** Fraud detection in banking.

Dimensionality Reduction Algorithms (For Handling High-Dimensional Data)

1. Principal Component Analysis (PCA)

- **Definition:**
  PCA is a statistical technique that **reduces the number of features** while preserving the most **important information**.
- **How it Works:**

1. Computes the **covariance matrix** of the data.
2. Identifies the **principal components** (directions of maximum variance).
3. Projects data onto these principal components.

- **Use Case:** Used in image compression and visualization of high-dimensional data.

2. t-SNE (t-Distributed Stochastic Neighbor Embedding)

- **Definition:**
  t-SNE is a **non-linear** dimensionality reduction technique used for **visualizing high-dimensional data** in 2D or 3D space.
- **Use Case:** Visualizing clusters in image recognition.

3. Autoencoders (Neural Network-Based)

- **Definition:**
  Autoencoders are **artificial neural networks** that learn an **efficient encoding** of data and can reconstruct the original data from the encoded representation.
- **Use Case:** Anomaly detection in cybersecurity.

Real-World Applications of Unsupervised Learning

📌 **Market Segmentation:** Grouping customers based on shopping behavior.
📌 **Anomaly Detection:** Identifying fraudulent transactions or network intrusions.
📌 **Recommender Systems:** Suggesting products (e.g., Netflix movie recommendations).
📌 **Image Compression:** Reducing image size while retaining quality.
📌 **Medical Diagnosis:** Detecting unknown diseases by grouping similar symptoms.

Supervised Learning

- Learns from labeled data, meaning each input has a known correct output.
- The model is trained to map inputs to outputs accurately.
- Used for prediction and classification tasks.
- Example: Email spam detection (emails are labeled as "spam" or "not spam").

Unsupervised Learning

- Works with unlabeled data, meaning there is no predefined output.
- The model finds patterns, structures, or relationships in data.
- Used for clustering and association tasks.
- Example: Customer segmentation (grouping customers based on behavior).

Reinforcement Learning

- Learns by interacting with an environment and receiving rewards or penalties.
- The model makes decisions to maximize long-term rewards.
- Used in self-learning systems like robotics, gaming AI, and recommendation systems.
- Example: Chess-playing AI learning moves by trial and error.

Types of Supervised Learning Algorithms

1. **Regression Algorithms** (Predict continuous values)
    - Linear Regression → Predicts house prices based on size.
    - Polynomial Regression → Models non-linear relationships, like population growth.
    - Support Vector Regression (SVR) → Uses Support Vector Machines for regression tasks.
2. **Classification Algorithms** (Categorize data into classes)
    - Logistic Regression → Classifies whether an email is spam or not.
    - Decision Tree → Splits data into decision-based categories.
    - Random Forest → Uses multiple decision trees for better accuracy.
    - Support Vector Machine (SVM) → Finds the best boundary between categories.
    - Naïve Bayes → Based on probability, used for text classification.
    - k-Nearest Neighbors (k-NN) → Classifies data based on nearest neighbors.

Types of Unsupervised Learning Algorithms

1. **Clustering Algorithms** (Group similar data points)
    - k-Means Clustering → Groups customers based on purchasing behavior.
    - Hierarchical Clustering → Creates a hierarchy of clusters.
    - DBSCAN → Detects clusters based on density.
2. **Association Algorithms** (Find relationships between variables)
    - Apriori Algorithm → Identifies frequent item sets (e.g., customers buying milk also buy bread).
    - FP-Growth → Faster association rule mining for large datasets.