Machine Learning

Module 1

MODULE-1	IntroductiontoMachineLearning	22AIM52.1,22AIM52.2	8Hours
Understanding Machine Learning: Definition and Types of Machine Learning-Application of Machine Learning- Machine Learning Algorithms: Supervised, Unsupervised, and Semi-Supervised Learning Algorithms. Machine Learning Models-**Model Evaluation Metrics:** Confusion Matrix, Precision, Recall, F1 Score-ROCCurveandAUC-ROC.**AdvancedTechniques:**FeatureScalingandNormalization-Encoding CategoricalVariables-Train-testSplitandCross-validation.			

---

1) Definition of Machine Learning

Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.

Machine Learning explores algorithms that can

- learn from data / build a model from data
- use the model for prediction, decision making or solving some tasks

Overview of AI/ML/DL vs Data Science:

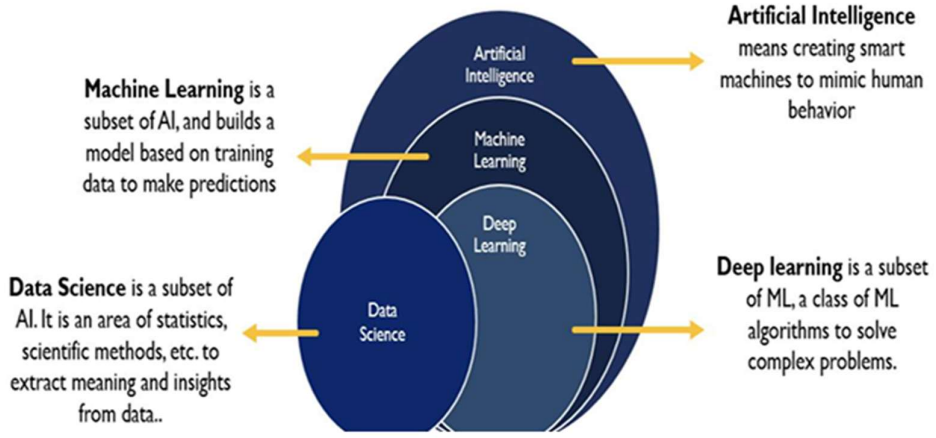


Figure1.2 AI/M/DL vs DS Overview

A subset of AI that focuses on teaching machines to learn patterns from data and make predictions without being explicitly programmed. The goal of ML is to build algorithms that improve automatically with experience.

Example: Spam email detection, product recommendations on Amazon, fraud detection.

Techniques used in ML are : Supervised learning, unsupervised learning, reinforcement learning.

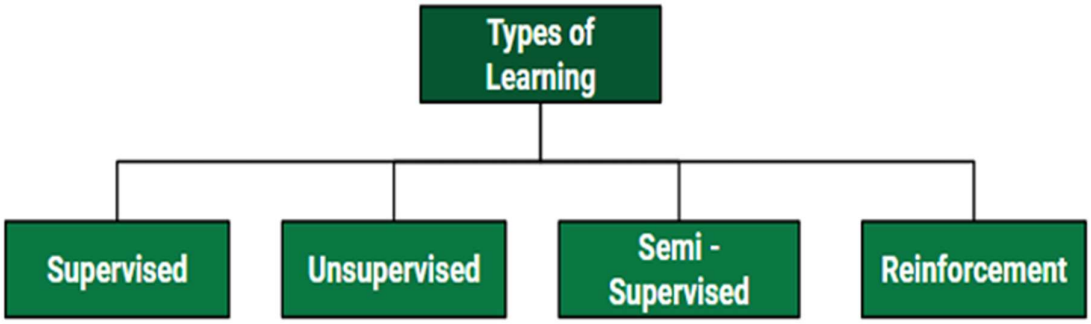
Terminologies in Machine Learning

1. **Dataset:** A collection of related data used for training or testing a machine learning model.
2. **Feature:** An individual measurable property or input variable of the data.
3. **Label:** The output or target value that the model is trained to predict.
4. **Training Data:** The data used to teach a machine learning model to learn patterns.
5. **Testing Data:** The data used to evaluate the performance and accuracy of a trained model.
6. **Algorithm:** A set of rules or steps a machine follows to learn patterns from data.
7. **Model:** The mathematical representation learned by an algorithm to map features to labels.
8. **Regression:** A type of model that predicts continuous numerical values.
9. **Classification:** A type of model that predicts discrete categories or classes.

Applications of Machine learning

1. **Image Recognition:**
   Used to identify objects, people, or places in images — e.g., Facebook's *DeepFace* for auto friend tagging.
2. **Speech Recognition:**
   Converts spoken language into text — used in Google Assistant, Siri, Alexa, and Cortana.
3. **Traffic Prediction:**
   Google Maps uses real-time GPS data and past traffic trends to predict road congestion.
4. **Product Recommendation:**
   E-commerce sites like Amazon and Netflix suggest products or shows based on user behavior using ML algorithms.
5. **Self-Driving Cars:**
   Cars like Tesla use machine learning (mainly unsupervised learning) to detect objects and drive autonomously.
6. **Email Spam and Malware Filtering:**
   ML algorithms like Naïve Bayes and Decision Trees filter spam emails and detect malware automatically.
7. **Virtual Personal Assistant:**
   Assistants like Alexa and Google Assistant use ML to process voice commands and perform tasks such as calling or scheduling.
8. **Online Fraud Detection:**
   ML models, especially neural networks, detect unusual transaction patterns to prevent online fraud.
9. **Stock Market Trading:**
   LSTM neural networks predict stock price trends and assist in algorithmic trading.
10. **Medical Diagnosis:**
    ML helps in diagnosing diseases and building 3D medical models to detect issues like brain lesions.
11. **Automatic Language Translation:**
    Google's GNMT uses neural machine translation to convert text between languages using sequence-to-sequence learning.

Types of Machine Learning



2) Supervised Machine Learning

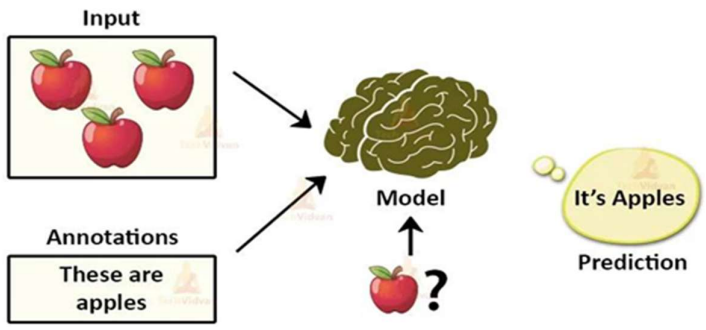
Definition:
Supervised learning uses a **labeled training dataset** to understand the relationships between input and output data. Data scientists manually create training datasets containing input data along with the corresponding labels.

Concept:
Supervised learning trains the model to apply the correct outputs to new input data in real-world use cases.
The trained model learns the underlying relationships between inputs and outputs, enabling it to **predict correct outputs** based on new, unlabeled real-world input data.

Example: Fruit Classification

Imagine we have a **basket full of different fruits** that we want the machine to identify.
The machine first looks at the image of a fruit and extracts **features** like its **shape, color, and texture**.
Then, it compares these features to the fruits it has already learned during training.

- If the fruit's features closely match those of an apple, the machine will predict that the fruit is an **Apple**.
- If the features match those of a banana, it will predict it as a **Banana**.



Training phase example:

- If the fruit is round, has a small depression at the top, and is red → labeled as **Apple**.
- If the fruit is long, curved, and greenish-yellow → labeled as **Banana**.

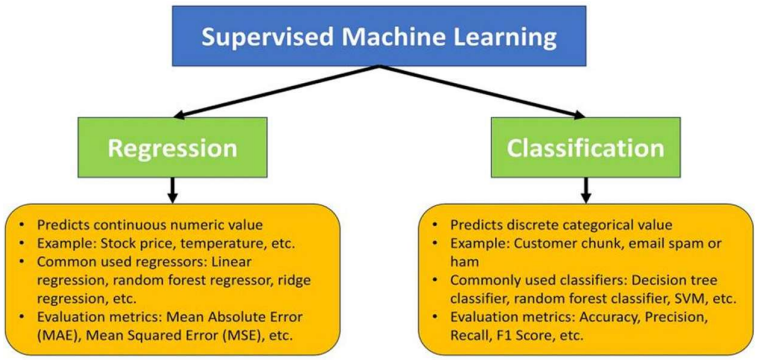
After training, when a new fruit (say a banana) is given, the machine will use what it has learned.
It analyzes the shape and color of the new fruit and classifies it as a **Banana**, placing it in the correct category.

Thus, the machine **learns from labeled training data** and **applies that knowledge** to recognize new, unseen fruits.

Working Logic

- Supervised learning algorithms are used when the **output is classified or labeled**.
- These algorithms **learn from past input data (training data)**, perform analysis, and use this to predict future outcomes within known classifications.
- **Accurate prediction** of test data requires **large datasets** to ensure sufficient understanding of patterns.
- The algorithm can be **further trained** by comparing the predicted outputs with actual results and **adjusting errors** to improve performance.

3) Types ofSupervised Machine Learning



Regression

Regression is a type of supervised learning where the model learns from labeled data to predict **continuous numerical outputs** based on input features.Two types of variables present in regression:

- Dependent Variable (Target): The variable we are trying to predict e.g house price.
- Independent Variables (Features): The input variables that influence the prediction e.g locality, number of rooms.

Regression analysis problem works with if output variable is a real or continuous value such as ―salary‖ or ―weight‖. Many different regression models can be used but the simplest model in them is linear regression.

Classification

Classification is a type of supervised learning where the model learns from labeled data to predict **discrete or categorical outputs** (i.e., class labels).

- Input (features): Measurable attributes of data (like age, symptoms, pixel values, text).
- Output (labels): A discrete category/class (like ―disease‖ or ―no disease‖).

The model‘s task is to map inputs → correct class labels, and then use that mapping to classify new, unseen data.

Example: Build a model that classifies emails into two categories: Spam or Not Spam

Real-world Examples of classification:

- Medical Diagnosis → Predict if a patient has a disease.
- Email Filtering → Spam vs. Not Spam.
- Sentiment Analysis → Positive / Negative review.

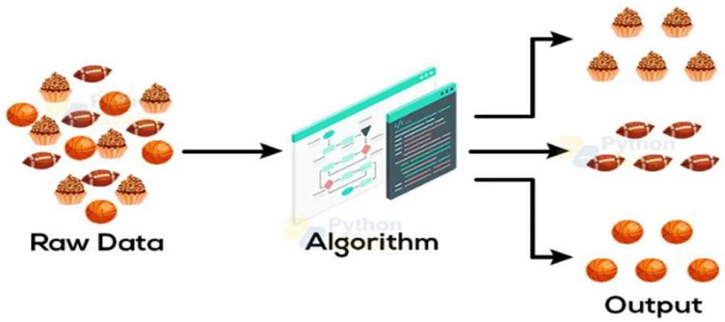
Advantages	Disadvantages
Works with labeled data, giving a clear idea about object classification and variables.	Cannot solve complex tasks if there is insufficient data.
Helps predict output based on prior experience and trained data.	May give wrong output if test data differs from training data or contains noise.
—	Requires high computational time and large amounts of training data.

4) Unsupervised Machine Learning

Definition:
Unsupervised learning is a type of machine learning that **analyzes and models data without labeled responses or predefined categories**.
Unlike supervised learning, where algorithms learn from input-output pairs, unsupervised learning algorithms work only with **input data** and aim to **discover hidden patterns, structures, or relationships** within the dataset without human supervision or prior knowledge.

Concept:

In unsupervised learning, the machine is trained using **unlabeled or untrained data**, and it tries to find useful insights, groupings, or associations on its own — **without any supervision**.

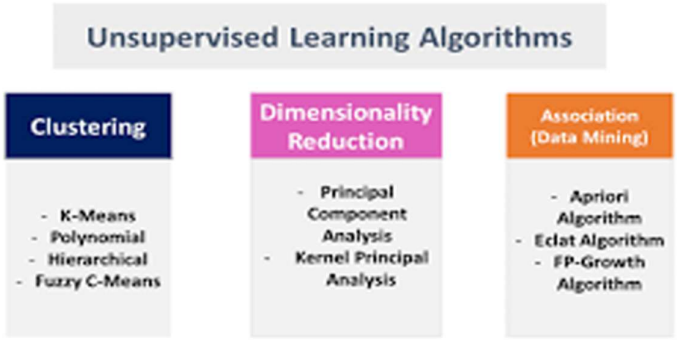


Working of Unsupervised Learning

The working of unsupervised machine learning can be explained in the following steps:

1. **Collect Unlabeled Data:**
   - Gather a dataset without predefined labels or categories.
   - *Example:* Images of various animals without any tags.
2. **Select an Algorithm:**
   - Choose a suitable unsupervised algorithm based on the goal, such as:
     - **Clustering:** K-Means
     - **Association Rule Learning:** Apriori
     - **Dimensionality Reduction:** PCA (Principal Component Analysis)
3. **Train the Model on Raw Data:**
   - Feed the entire unlabeled dataset to the algorithm.
   - The algorithm identifies similarities, relationships, or hidden structures within the data.
4. **Group or Transform Data:**
   - The algorithm organizes data into **groups (clusters)**, **rules**, or **lower-dimensional forms** automatically.
   - *Example:* It may group similar animals together or extract key patterns from large datasets.
5. **Interpret and Use Results:**
   - Analyze the discovered groups, rules, or features to gain insights.
   - The output can be used for **visualization, anomaly detection**, or as **input for other models**.

5) Types of Unsupervised Machine Learning



1. Clustering Algorithms

Definition:
Clustering is an **unsupervised learning technique** that groups unlabeled data into clusters based on **similarity** among data points.

Key Points:

- Groups data points with similar features or characteristics.
- Helps discover natural groupings or patterns in unclassified data.
- Works purely from input data without any output labels.
- Commonly used for **customer segmentation, anomaly detection, and data organization**.
- Enables better understanding of the data structure for further analysis or decision-making.

Common Clustering Algorithms:

- **K-Means Clustering:** Groups data into *K* clusters based on the distance between points.
- **Hierarchical Clustering:** Forms clusters by continuously merging or splitting groups in a tree structure.
- **DBSCAN (Density-Based Clustering):** Identifies clusters in dense regions and treats scattered points as noise.
- **Mean-Shift Clustering:** Finds clusters by shifting points toward denser regions.
- **Spectral Clustering:** Uses graph-based similarity to group data points.

---

2. Association Rule Learning

Definition:
Association Rule Learning is a **rule-based unsupervised learning technique** that finds interesting **relationships or patterns** between variables in large datasets.

Key Points:

- Discovers frequent item combinations and relationships between them.

- Expresses findings in **"if–then" rules** (e.g., *If a customer buys bread, they also buy butter*).
- Commonly used in **market basket analysis**, promotions, and cross-selling strategies.
- Helps businesses understand item purchase relationships.

Common Association Rule Learning Algorithms:

- **Apriori Algorithm:** Finds frequent itemsets step-by-step through iterative exploration.
- **FP-Growth Algorithm:** Faster alternative to Apriori; finds frequent patterns without candidate generation.
- **Eclat Algorithm:** Uses itemset intersections to efficiently find frequent patterns.
- **Tree-Based Algorithms:** Organize data hierarchically for efficient large-scale pattern discovery.

---

3. Dimensionality Reduction

Definition:
Dimensionality Reduction is the process of **reducing the number of features (dimensions)** in a dataset while retaining most of the important information.

Key Points:

- Simplifies complex data for easier analysis and visualization.
- Improves algorithm efficiency and reduces noise or overfitting.
- Focuses on the most relevant traits or patterns in the data.
- Commonly used to **enhance model performance and reduce computation time**.

Common Dimensionality Reduction Algorithms:

- **PCA (Principal Component Analysis):** Converts correlated variables into uncorrelated principal components.
- **LDA (Linear Discriminant Analysis):** Reduces dimensions while maximizing class separability (for classification tasks).
- **NMF (Non-negative Matrix Factorization):** Decomposes data into additive, non-negative parts for simpler representation.
- **LLE (Locally Linear Embedding):** Preserves local neighborhood relationships while reducing dimensions.
- **Isomap:** Maintains global data structure by preserving distances along the data manifold.

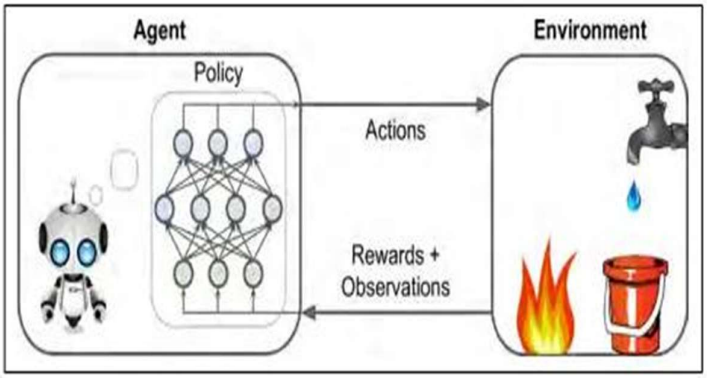
	Advantages of Unsupervised Learning	Disadvantages of Unsupervised Learning
	Can be used for complicated tasks as it works on **unlabeled datasets** and doesn't require large labeled data.	May produce **less accurate results** since data is unlabeled and there are no predefined outputs.
	Easier to obtain **unlabeled data** for training compared to labeled datasets.	**More difficult to work with**, as there's no mapping between inputs and outputs.

	Helps **discover hidden patterns or structures** in raw data.	Finding the **true underlying structure** can be challenging and sometimes subjective.

6) Reinforcement Learning (RL)

Definition:
Reinforcement Learning is a type of machine learning where an **agent learns through trial and error** by interacting with an **environment**. The agent performs actions and receives **rewards or penalties** as feedback, helping it to learn the best strategy (policy) to achieve a goal.

Key Idea:
The agent learns to make a sequence of decisions to **maximize cumulative rewards** over time without human supervision.



Elements of Reinforcement Learning

Element	Description
Agent	Learner or decision-maker.
Environment	External system the agent interacts with.
Actions	Possible moves or operations the agent can perform.
Rewards	Feedback received after an action (positive or negative).
Policy	Strategy that defines how the agent selects actions.

Core Components

Component	Description	Example
1. Policy	Defines the agent's behavior; maps states to actions.	Self-driving car deciding when to stop.
2. Reward Signal	Represents the goal and provides feedback.	Fewer collisions or shorter travel time.
3. Value Function	Estimates long-term benefit of a state considering future rewards.	Avoiding risky driving for long-term safety.
4. Model	Simulates the environment to predict outcomes.	Predicting other vehicles' movements.

Working of Reinforcement Learning

1. The agent observes the **current state** of the environment.
2. It selects an **action** based on its policy.
3. The environment transitions to a **new state** and gives a **reward/penalty**.
4. The agent updates its **policy/value function** using this feedback.
5. The process repeats — the agent balances **exploration** (trying new actions) and **exploitation** (using known actions) to **maximize total reward**.

Advantages

No.	Advantages
1	Solves **complex real-world problems** where traditional methods fail.
2	Mimics **human learning** behavior, leading to accurate decision-making.
3	Focuses on **long-term performance** rather than short-term results.

Disadvantages

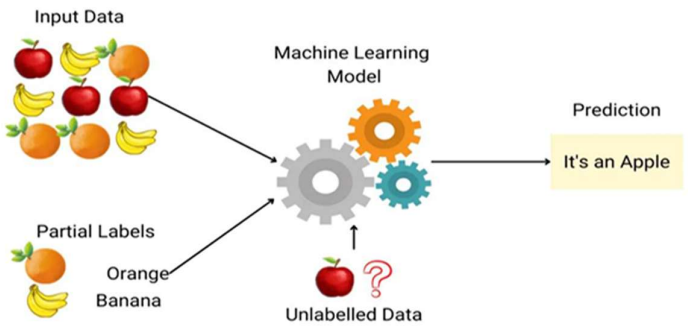
No.	Disadvantages
1	Not suitable for **simple problems**.
2	Requires **large data** and **high computational power**.
3	**Overtraining** can cause too many state updates, reducing effectiveness.

7) Semi-Supervised Learning (SSL)

Definition:
Semi-Supervised Learning is a **machine learning approach that combines both labeled and unlabeled data** during training. It lies **between supervised and unsupervised learning**, using a small amount of labeled data along with a large amount of unlabeled data to improve learning accuracy.

Concept Explanation

- Supervised learning requires **large labeled datasets**, which are **expensive and time-consuming** to prepare.
- Unsupervised learning works on **unlabeled data** but provides **limited accuracy** and control.
- **Semi-supervised learning bridges this gap** by using a small labeled dataset to guide learning from a much larger unlabeled dataset.



Example:
In image recognition, only a few images are labeled (e.g., “cat”, “dog”), and the rest are unlabeled. The model learns from both to improve classification accuracy.

Working of Semi-Supervised Learning

1. **Input Data:** Mix of small labeled data + large unlabeled data.
2. **Clustering:** Uses unsupervised methods to find patterns or similarities in the data.
3. **Label Propagation:** The labeled data helps assign pseudo-labels to similar unlabeled samples.
4. **Model Training:** The algorithm uses both true and pseudo-labeled data to train efficiently.
5. **Prediction:** Produces better generalization than purely supervised or unsupervised models.

Advantages of Semi-Supervised Learning

No.	Advantages
1	Simple and easy to understand; does not encounter major anomalies.
2	Efficient in predicting output from input data.
3	Reduces cost and effort of manual labeling.
4	Overcomes drawbacks of both supervised and unsupervised learning.

Disadvantages of Semi-Supervised Learning

No.	Disadvantages
1	Iteration results may not be stable; outputs can vary.
2	Cannot be applied effectively to complex network-level data.
3	May have **lower accuracy** compared to fully supervised models.

Applications

- Text classification (spam filtering, sentiment analysis)
- Speech recognition
- Image and video labeling
- Fraud detection

Criteria	Supervised ML	Unsupervised ML	Reinforcement ML
Definition	Learns by using labelled data	Trained using unlabelled data without any guidance.	Works on interacting with the environment
Type of data	Labelled data	Unlabelled data	No – predefined data
Type of problems	Regression and classification	Association and Clustering	Exploitation or Exploration
Supervision	Extra supervision	No supervision	No supervision
Algorithms	Linear Regression, Logistic Regression, SVM, KNN etc.	K – Means, C – Means, Apriori	Q – Learning, SARSA
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Application	Risk Evaluation, Forecast Sales	Recommendation System, Anomaly Detection	Self Driving Cars, Gaming, Healthcare

8) Confusion Matrix

Definition

A **Confusion Matrix** is a performance measurement tool used for **classification problems** where the output has two or more classes.
It compares the **actual values** with the **predicted values** to evaluate the accuracy of a model.

It is a **table** that helps visualize how many predictions were **correct** and **incorrect** across different classes.

Terminologies in Confusion Matrix

Term	Meaning	Description / Example
True Positive (TP)	Correctly predicted positive	Patient has COVID and is correctly diagnosed as COVID-positive
True Negative (TN)	Correctly predicted negative	Patient is healthy and correctly diagnosed as negative
False Positive (FP)	Incorrectly predicted positive (Type I Error)	Patient is healthy but diagnosed as COVID-positive

False Negative (FN)	Incorrectly predicted negative (Type II Error)	Patient has COVID but diagnosed as healthy

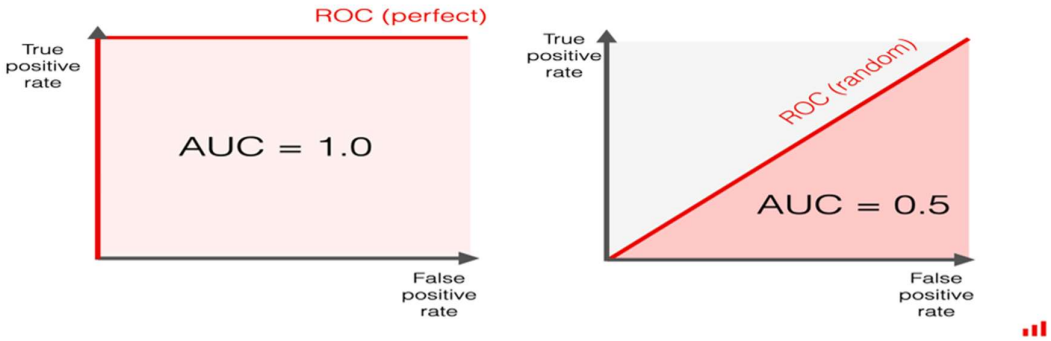
Metrics Based on Confusion Matrix

Metric	Formula	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Measures overall correctness of the model.
Precision	TP / (TP + FP)	Measures how many predicted positives are actually positive.
Recall (Sensitivity / TPR)	TP / (TP + FN)	Measures how many actual positives are correctly predicted.
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean of precision and recall; balances both.
Specificity (TNR)	TN / (TN + FP)	Measures how well the model identifies negative cases.

9) **Short note on AUC and ROC**

ROC (Receiver Operating Characteristic) Curve

- ROC curve is a **graphical representation** that shows the performance of a classification model at different threshold values.
- It plots **True Positive Rate (TPR)** on the y-axis and **False Positive Rate (FPR)** on the x-axis.
- The curve helps to understand how well the model distinguishes between positive and negative classes.
- A model with a curve closer to the top-left corner indicates **better performance**.



AUC (Area Under the Curve)

- AUC represents the **area under the ROC curve**.
- It measures the **overall ability of the model to classify** positive and negative instances correctly.
- The value of AUC ranges from **0 to 1**:
  - **AUC = 1:** Perfect model
  - **AUC = 0.5:** No discrimination (same as random guessing)
  - **AUC < 0.5:** Poor model performance
- Higher AUC means a **better performing model**.

Relationship between Threshold, TPR & FPR

Threshold	Effect	Recall (TPR)	FPR
High Threshold (0.95)	Model is conservative (less True Positives)	Decreases	Decreases
Medium Threshold (0.8)	Balanced prediction	Moderate	Moderate
Low Threshold (0.5)	Model predicts positive easily	Increases	Increases

1. Feature Scaling and Normalization (10 Marks)

Definition:

Feature Scaling is the process of adjusting the range of independent variables or features of data so that they contribute equally to the model.
Normalization is a specific type of scaling where data values are rescaled to a fixed range, usually between 0 and 1.

Need:

- Different features may have different units and ranges.
- Models like KNN, SVM, Logistic Regression, and Neural Networks are sensitive to scale.
- Prevents domination of high-magnitude features and improves convergence speed in optimization algorithms.

Types of Feature Scaling:

1. **Min–Max Normalization:**
   Rescales features to a range of [0,1].
   Formula:
   $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$
2. **Standardization (Z-score Normalization):**
   Centers data around mean 0 and standard deviation 1.
   Formula:
   $$X' = \frac{X - \mu}{\sigma}$$
3. **Robust Scaling:**
   Uses median and interquartile range (IQR) to reduce effect of outliers.
   Formula:
   $$X' = \frac{X - \text{Median}(X)}{\text{IQR}(X)}$$

Normalization:

- A technique to bring all data values within the same range, often [0,1].
- Maintains relative relationships between values but changes magnitude.

- Used when features have different scales and data does not follow a normal distribution.

Applications:

- Gradient descent-based algorithms.
- Distance-based algorithms (KNN, K-Means).
- Neural Networks for stable and faster training.

---

2. Encoding Categorical Variables (10 Marks)

Definition:

Encoding Categorical Variables is the process of converting categorical or qualitative data into numerical format so that it can be used by machine learning algorithms.

Need:

- Most algorithms can only process numerical data.
- Encoding makes categorical data interpretable for mathematical operations.
- Prevents misinterpretation of categories as having numerical order unless intended.

Techniques of Encoding:

1. **Label Encoding:**
    - Assigns an integer value to each category.
    - Example: {Red: 0, Blue: 1, Green: 2}.
    - Suitable for tree-based models.
    - May introduce false order for nominal data.
2. **One-Hot Encoding:**
    - Creates a binary column for each category.
    - Example: “Color” with {Red, Blue} becomes Red = [1,0], Blue = [0,1].
    - Eliminates order relationship but increases dimensionality.
3. **Ordinal Encoding:**
    - Assigns ordered integers to categories that have a defined ranking.
    - Example: Low = 1, Medium = 2, High = 3.
4. **Target Encoding:**
    - Replaces category with mean of target variable for that category.
    - Useful for high-cardinality features.
5. **Binary Encoding:**
    - Converts categories into binary digits and represents them as separate columns.
    - More memory-efficient than one-hot encoding.
6. **Frequency Encoding:**
    - Replaces each category with its frequency or count in the dataset.

Applications:

Used during data preprocessing for regression, classification, and clustering tasks to prepare categorical attributes for numeric model inputs.

---

3. Train–Test Split and Cross-Validation (10 Marks)

Definition:

Train–Test Split and Cross-Validation are methods used to evaluate the performance and generalization capability of machine learning models.

A. Train–Test Split

Concept:

The dataset is divided into two parts:

- **Training Set:** Used to train the model.
- **Testing Set:** Used to evaluate how well the model performs on unseen data.

Typical Split Ratios:

- 70% Training and 30% Testing
- 80% Training and 20% Testing

Advantages:

- Simple and fast to implement.
- Provides an initial estimate of model performance.

Limitations:

- Performance may depend heavily on how data is split.
- May not generalize well if dataset is small.

B. Cross-Validation

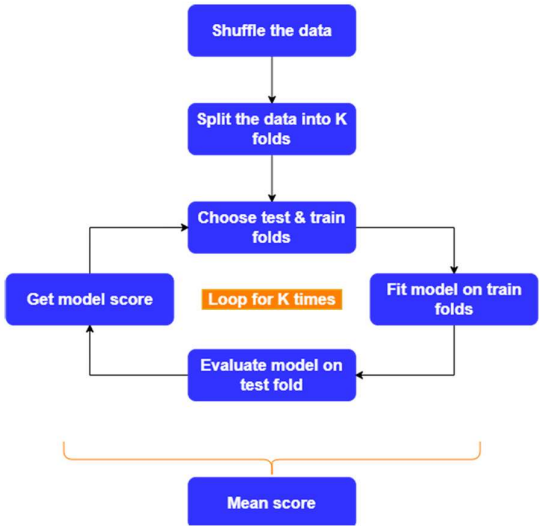
Definition:

Cross-Validation is a resampling technique that divides the dataset into multiple subsets (folds) and tests the model on each fold to get an average performance score.

Procedure (K-Fold Cross-Validation):

1. Divide data into k equal parts (folds).
2. Train the model on $k–1$ folds and test on the remaining fold.
3. Repeat the process k times, each time using a different fold as the test set.

4. Average the results to obtain the final performance metric.



Types of Cross-Validation:

1. **K-Fold Cross-Validation** – Most commonly used.
2. **Stratified K-Fold** – Maintains the same class proportion across folds.
3. **Leave-One-Out Cross-Validation (LOOCV)** – Each data point is used once as a test case.
4. **Repeated K-Fold** – Repeats K-Fold multiple times for stability.
5. **Time-Series Cross-Validation** – Used for time-dependent data.

Advantages:

- Provides a more reliable estimate of model performance.
- Reduces the risk of overfitting or underfitting.
- Makes better use of available data for both training and validation.