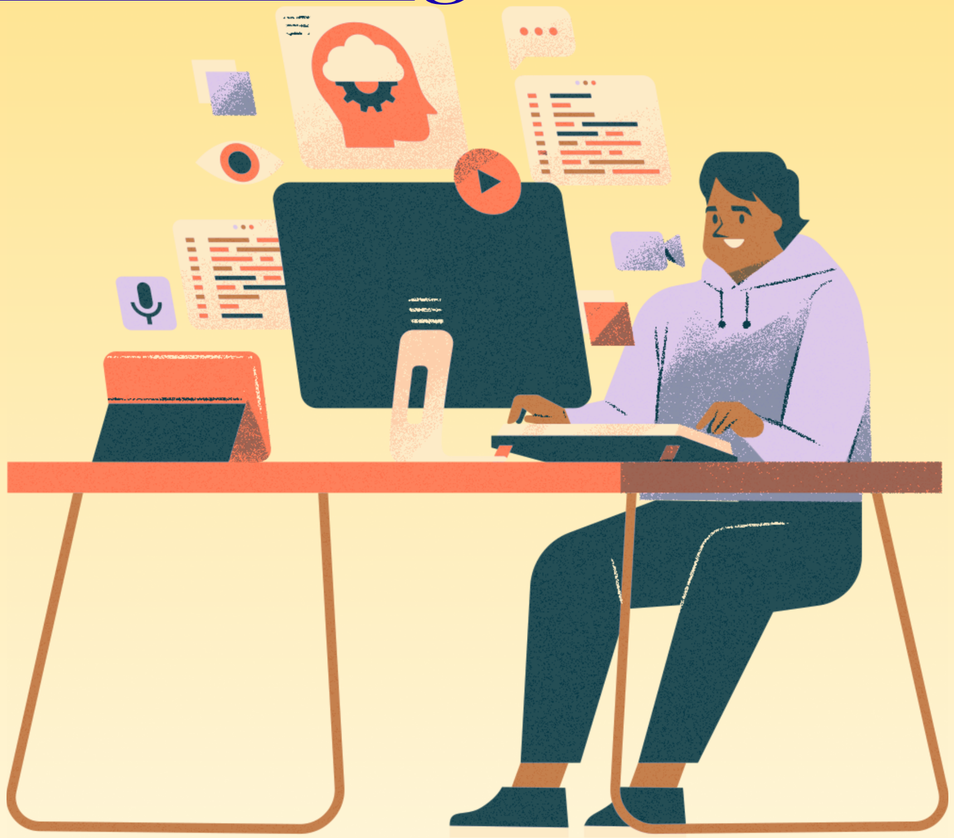
Machine Learning

Introduction To Machine Learning

unit 1



Contents

- Introduction to Machine Learning
- Comparison of Machine learning with traditional programming
- ML vs AI vs Data Science
- **Types of learning**: Supervised, Unsupervised, and semi-supervised, reinforcement learning techniques
- **Models of Machine learning:**
  Geometric model
  Probabilistic Models
  Logical Models
  Grouping and grading models
  Parametric and non-parametric models
- Important Elements of Machine Learning- Data formats, Learnability, Statistical learning approaches

Introduction to Machine Learning

Machine Learning is a branch of Artificial Intelligence that enables a system to automatically learn and improve from experience (data) without being explicitly programmed.

In traditional programming: Rules + Data ⇾ Output
In ML: Data + Output ⇾ Algorithm learns Rules (Model)

Example:
Spam Email Detection ⇾ The system learns patterns (like suspicious words, sender info) from past emails and classifies new emails as spam or not spam.

- Automates decision-making
- Improves with more data
- Widely used in healthcare, finance, e-commerce, robotics, self-driving cars, etc.

Comparison

Aspect	**Traditional Programming**	**Machine Learning**
Approach	Programmer defines rules explicitly.	System learns rules from data.
Input	Data + Rules	Data + Expected Output (labels)
Output	Result (Output)	Model (that can make predictions)
Flexibility	Limited – rules must be updated manually.	High – adapts automatically as data changes.
Handling Complexity	Struggles with large, unstructured data.	Handles complex, high-dimensional, unstructured data well.
Examples	Calculator, Payroll system, Ticket booking.	Spam detection, Face recognition, Recommendation systems.

Comparison

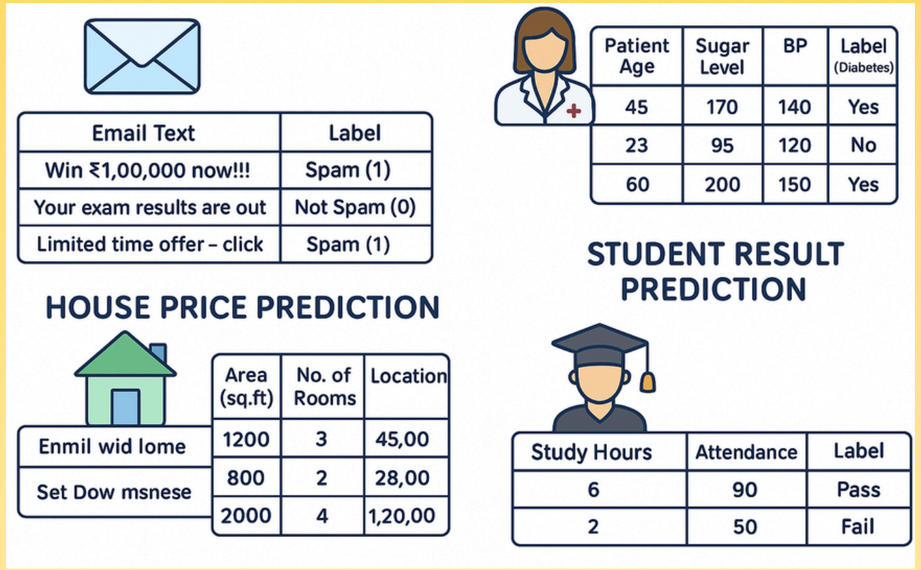
Aspect	**Artificial Intelligence (AI)**	**Machine Learning (ML)**	**Data Science (DS)**
Definition	Broad field of making machines think and act like humans.	Subset of AI that allows systems to learn from data.	Field that deals with collecting, processing, analyzing data to extract insights.
Scope	Very broad (reasoning, decision-making, problem-solving, learning).	Focused on algorithms that learn patterns from data.	Covers statistics, ML, data analysis, visualization.
Goal	Create intelligent systems.	Create models that improve with data.	Derive knowledge and insights from data.
Techniques Used	Robotics, Expert Systems, Natural Language Processing (NLP).	Regression, Classification, Clustering, Neural Networks.	Data cleaning, EDA, ML, Big Data tools (Hadoop, Spark).
Output	Intelligent behavior and decision-making.	Predictions and patterns.	Actionable insights, reports, predictive models.
Example	Siri, Self-driving cars, Chess-playing bots.	Spam filter, Product recommendation, Credit scoring.	Business analytics, Data dashboards, Predicting sales trends.

Types of learning

- Supervised Lerning
  - classification
  - Regression
- Unsupervised Lerning
  - clustering
  - dimensionality Reduction
- Semi-supervised Lerning
- Reinforcement Learning

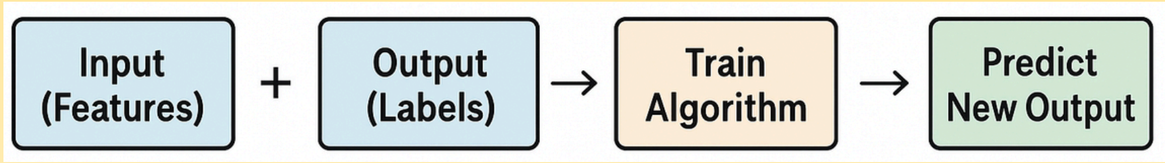
Supervised Learning

- Uses labeled data (Input + Output)
- Learns mapping from input ⇾ output
- Algorithm trained on past examples ⇾ predicts for new data



Examples:

- Spam Email Detection (Spam / Not Spam)
- Predicting House Prices



TASKS :

Classification

- Predicts discrete labels (Yes/No, Categories).
- Example: Spam Detection, Disease Diagnosis.
- Algorithms: Logistic Regression, Decision Trees, SVM, Naïve Bayes.

Regression

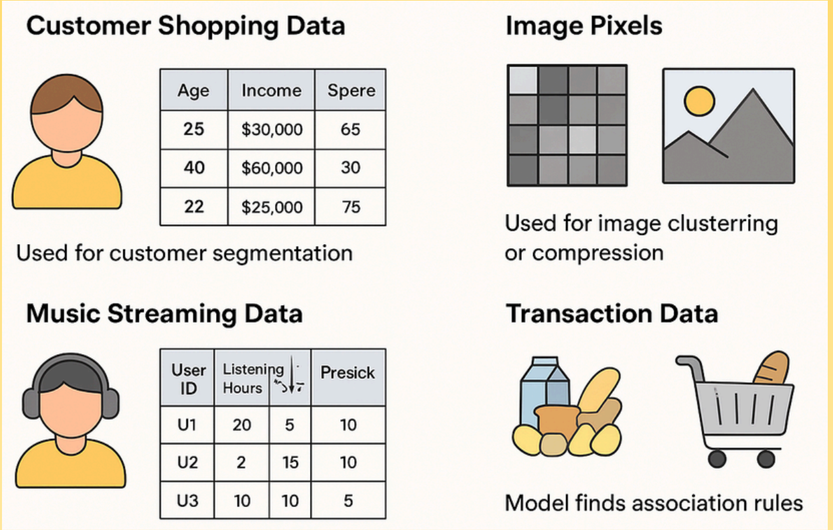
- Predicts continuous values (Numbers).
- Example: House Prices, Temperature Forecast.
- Algorithms: Linear Regression, Polynomial Regression, Support Vector Regression.

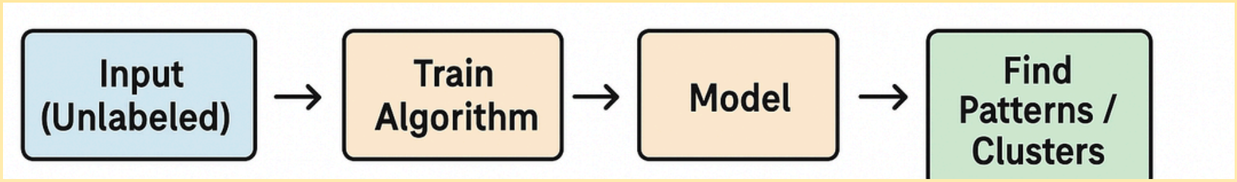
Unsupervised Learning

- works on unlabeled data
- Learns hidden patterns & structures without predefined output
- Output Type: Groups (clusters) or reduced feature representations.

Examples:

- Customer Segmentation (grouping buyers)
- Market Basket Analysis (items bought together)





TECHNIQUES:

Clustering

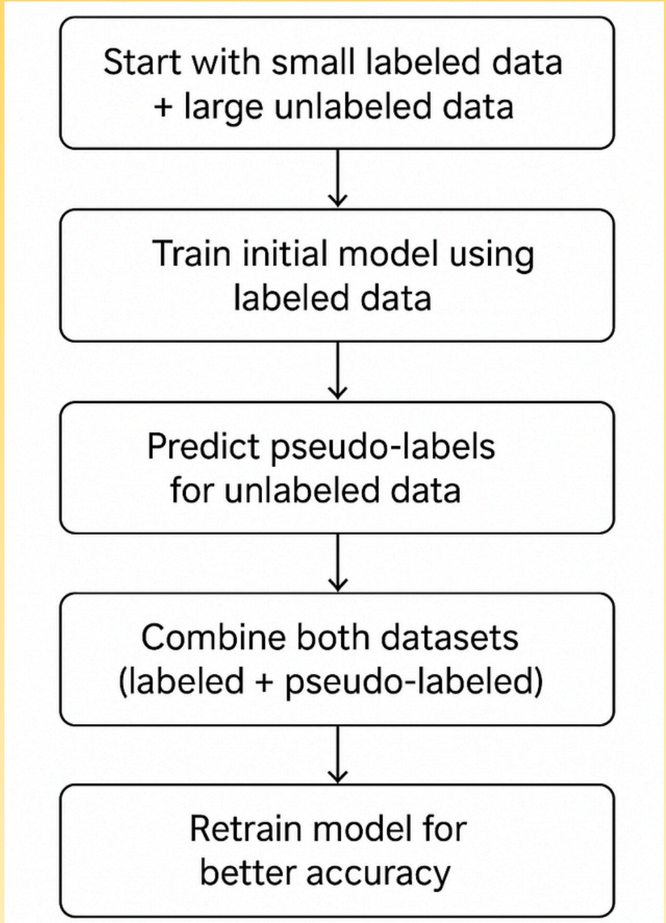
- Groups similar data points into clusters based on patterns or similarity.
- Output Type: Cluster labels (no predefined classes).
- Algorithms: K-Means, Hierarchical Clustering, DBSCAN

Dimensionality Reduction

- Reduces the number of features while keeping important information.
- Purpose: Simplifies data, removes noise, speeds up training, helps visualization.
- Examples: Compressing images while preserving structure

Semi-supervised Learning

- Uses few labeled + many unlabeled data
- Mix of Supervised & Unsupervised
- Reduces labeling cost
- Examples: Medical diagnosis, Speech recognition, Web classification



TECHNIQUES :

- Self-training ↠ Model trains on labeled data ↠ predicts labels for unlabeled ↠ retrains with them.
- Co-training ↠ Two models learn from different views/features of same data ↠ teach each other.
- Graph-based methods ↠ Data points are nodes in a graph ↠ labels spread through connections.

Reinforcement Learning

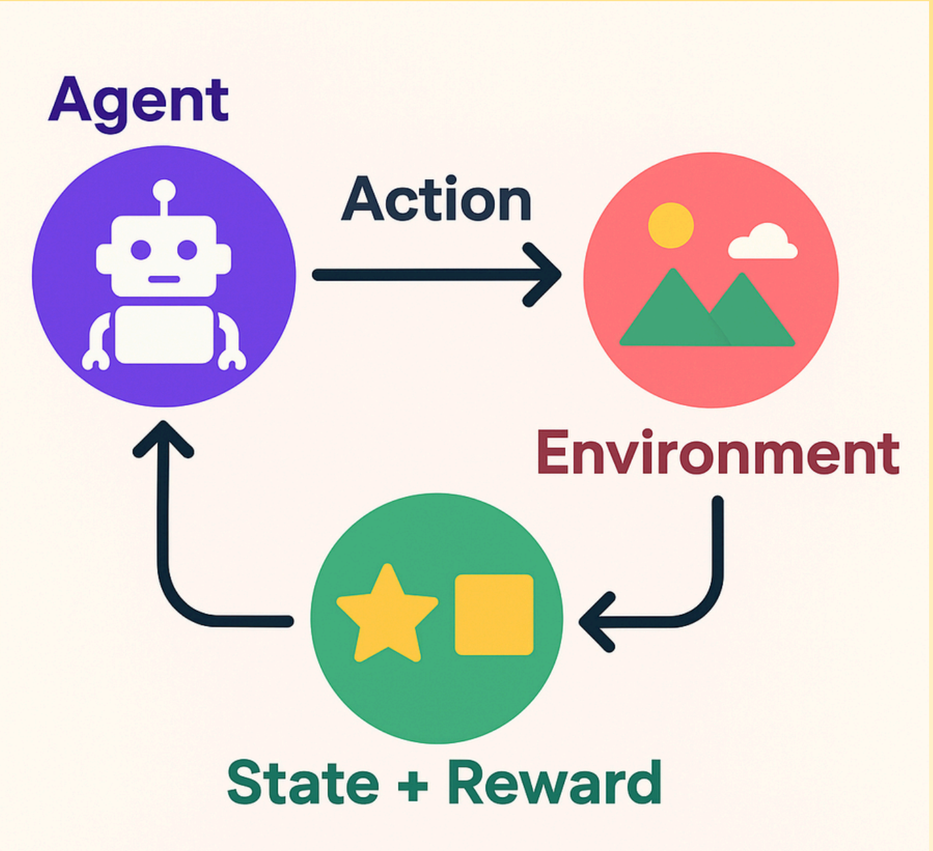
- A type of machine learning where an agent learns to make decisions by interacting with an environment to maximize rewards.

Process:

- Agent ↠ Takes an action.
- Environment ↠ Responds with a new state and reward.
- Agent ↠ Learns from feedback and improves strategy (policy).

Examples:

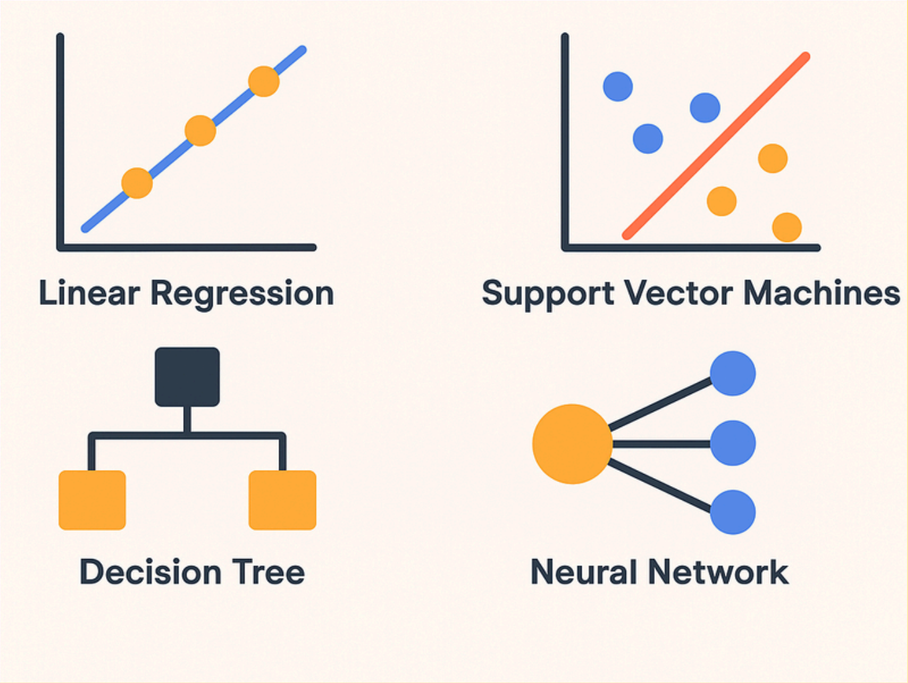
- Game playing (e.g., Chess, Go, Atari).
- Robotics (self-learning robots).
- Self-driving cars (navigation and control).



Models of Machine learning

📊 Geometric Models

- **Definition**: Represent data in geometric space (lines, planes, hyperplanes).
- **Idea**: Separate or classify data based on distance, margin, or similarity.
- **Examples:**
- k-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Use Case: Image recognition, pattern detection.



- **Data as Geometry** – Treats data points as vectors in high-dimensional space.
- **Decision Boundaries** – Learning is about finding lines, planes, or hyperplanes that separate data classes.
- **Distance & Similarity** – Uses geometric measures (Euclidean distance, cosine similarity) for classification and clustering.

Models of Machine learning

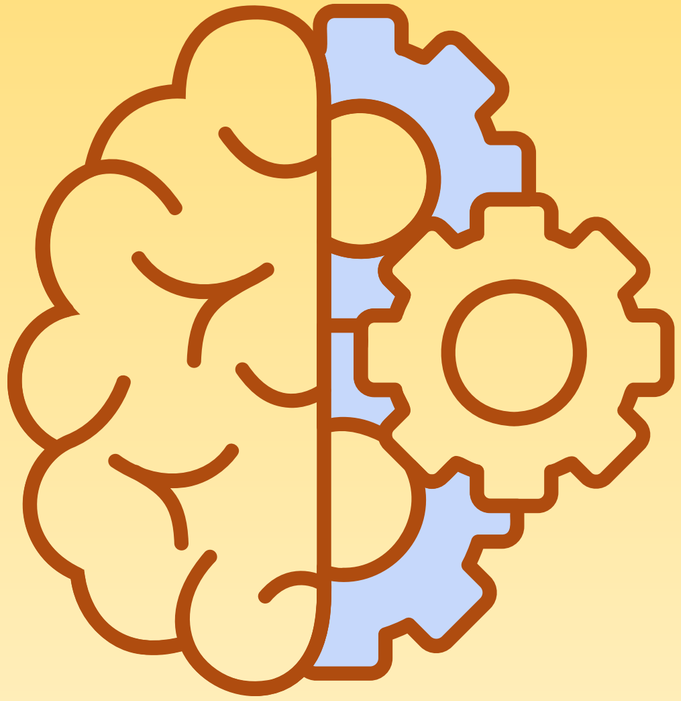
🎲 Probabilistic Models

- **Definition**: Use probability theory to model uncertainty in data.
- **Idea**: Predict outcomes by calculating likelihoods.
- Examples:
  - Naïve Bayes
  - Hidden Markov Models
- **Use Case**: Spam filtering, speech recognition.

- **Represents Uncertainty** – Assigns probabilities to outcomes instead of rigid predictions.
- **Models Relationships** – Links observed data (XXX), hidden variables (ZZZ), and outcomes (YYY).
- **Inference** – Computes probabilities of hidden variables/predictions (e.g., Bayes' theorem).
- **Prediction** – Provides outcome probabilities (e.g., 70% chance of 5, 20% chance of 4, 10% chance of 6).
- **Learning from Data** – Learns probability distributions from training data.

Models of Machine learning

🧩 Logical Models



- **Definition**: Learn rules and conditions from data.
- **Idea**: If–Then logic for decision making.
- **Examples**:
  - Decision Trees
  - Rule-based Systems
- **Use Case**: Decision Trees Rules like:

  IF age < 18 AND income = low ↣ class = “Student”

  Medical diagnosis:

  IF fever = yes AND cough = yes ↣ possible flu.

- **Purpose** – Used for binary classification (Yes/No, 0/1).
- **Logistic Function** – Applies the sigmoid function to map values into probability range [0,1].
- **Output** – Predicts the probability of an instance belonging to a class.

Models of Machine learning

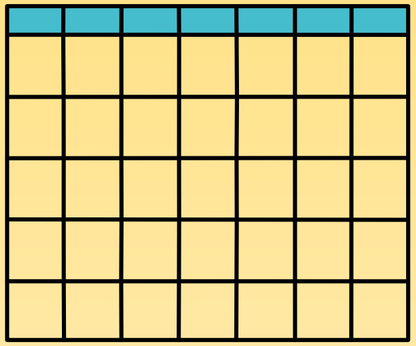
Aspect	Grouping Models	Grading Models
Aspect	Divide data into clusters based on similarity.	Assign scores/ranks to data points.
Output	Cluster labels (e.g., Cluster 1, Cluster 2).	Ordered ranking/grades (e.g., High, Medium, Low).
Goal	Discover hidden structure in data.	Prioritize or evaluate importance.
Techniques	K-Means, Hierarchical Clustering.	Recommendation Systems, Ranking Algorithms.
Use Cases	Market segmentation, Customer groups.	Search engine ranking, Student grading.

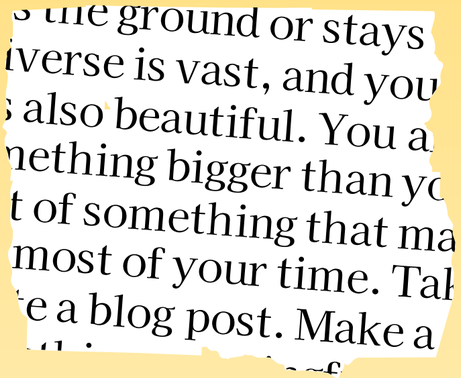
Models of Machine learning

Aspect	Parametric Models	Non-Parametric Models
Definition	Fixed set of parameters regardless of data size.	Parameters grow with data size, more flexible.
Examples	Linear Regression, Logistic Regression.	KNN, Decision Trees.
Data Requirement	Works well with small datasets.	Requires large amounts of data.
Speed	Fast and efficient.	Slower due to complexity.
Flexibility	Less flexible, assumes fixed form.	Highly flexible, adapts to data patterns.

Data Formats

Definition: The way input data is structured and represented.



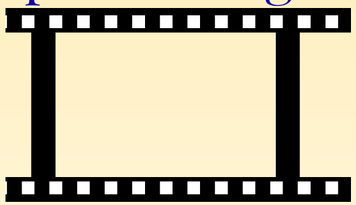


Types:

- **Structured Data**: Tables, rows & columns (e.g., databases).
- **Unstructured Data**: Text, images, audio, video.
- **Semi-Structured Data**: JSON, XML, log files.

```
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this
   weekend!
  <body>
</note>
```

Importance: Good format = easier preprocessing & better learning.



Learnability

Definition: The ability of an algorithm to learn patterns from data.

Quality & Size of Dataset – More and cleaner data improves learning.

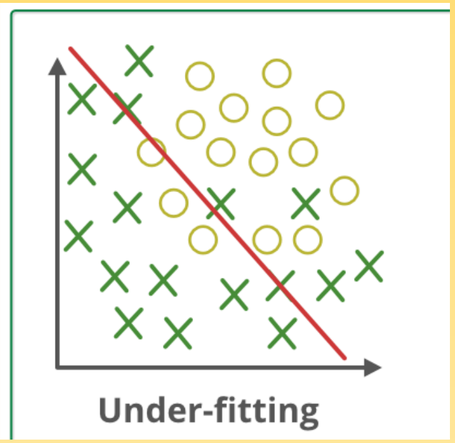
Model Complexity –

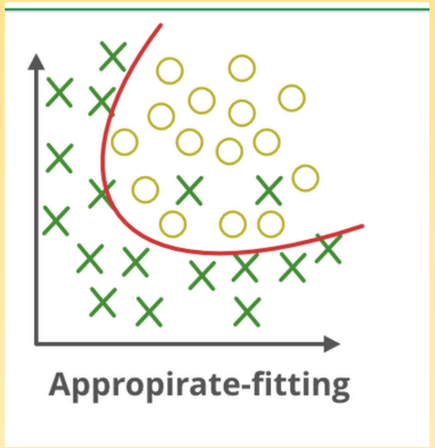
- If the model is too simple, it cannot capture important patterns ⇾ Underfitting.
- If the model is too complex, it memorizes training data instead of learning general patterns ⇾ Overfitting.

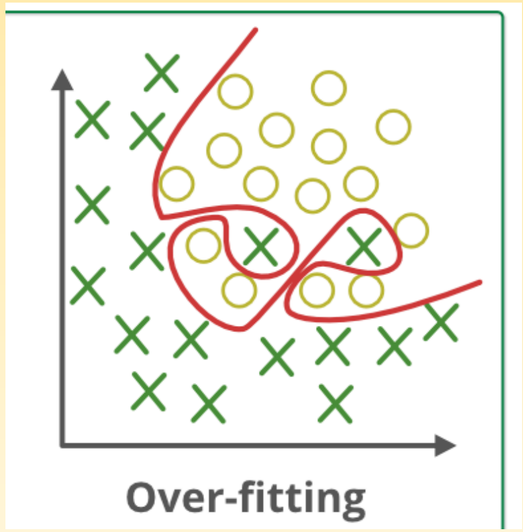
Availability of Labels –

- Supervised Learning ⇾ requires labeled data.
- Unsupervised Learning ⇾ works without labels.

Goal: Ensure the model generalizes well on unseen data.







Statistical Learning Approaches

- **Definition**: Mathematical foundation of ML based on statistics & probability.

Main Ideas:

- Estimate relationships between variables.
- Minimize error/loss using statistical methods.
- Use probability to handle uncertainty in predictions.

Examples:

- Linear Regression (predicting continuous values).
- Logistic Regression (classification).
- Bayesian Learning (probabilistic approach).

THANK YOU