UNIT-II

SUPERVISED LEARNING

1 a Differentiate Supervised Learning and Unsupervised Learning [L4][CO5] [5M]

Difference Between Supervised and Unsupervised Learning

Introduction:

Machine learning helps computers learn from data and make decisions. It is divided into **Supervised Learning** and **Unsupervised Learning** based on the type of data used for training.

Supervised Learning	Unsupervised Learning
The model is trained using **labeled data** (input and correct output are given).	The model is trained using **unlabeled data** (only input is given, no correct output).
The model learns with the help of **feedback** (checks if the output is correct or not).	The model does **not** take feedback.
It is used to **predict results** based on past data.	It is used to **find hidden patterns** in the data.
Needs **human supervision** to train.	Does **not** need human supervision.
More **accurate and reliable**.	Less accurate compared to supervised learning.
Used in **Classification** (e.g., Spam detection) and **Regression** (e.g., Predicting house prices).	Used in **Clustering** (e.g., Customer segmentation) and **Association** (e.g., Market Basket Analysis).
Examples: **Linear Regression, Decision Tree, Support Vector Machine (SVM), Logistic Regression**.	Examples: **K-Means Clustering, K-Nearest Neighbors (KNN), Apriori Algorithm**.
Not close to true AI because it needs labeled data for training.	**Closer** to true AI because it learns from experience like humans.

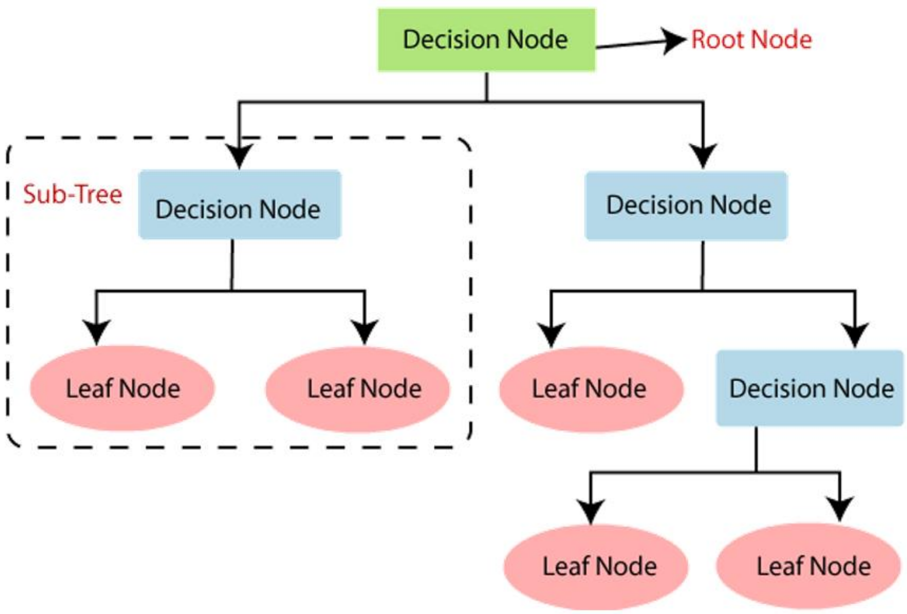
Conclusion:

- **Supervised learning** is useful when we have labeled data and want accurate predictions.
- **Unsupervised learning** is useful when we only have input data and want to find patterns.

Both types are important in machine learning and used in different real-world applications.

Simple Explanation of Decision Tree Classification

A **Decision Tree** is a method used in **Machine Learning** to classify data by asking a series of **Yes/No questions**. It looks like a **tree** with branches leading to different decisions.



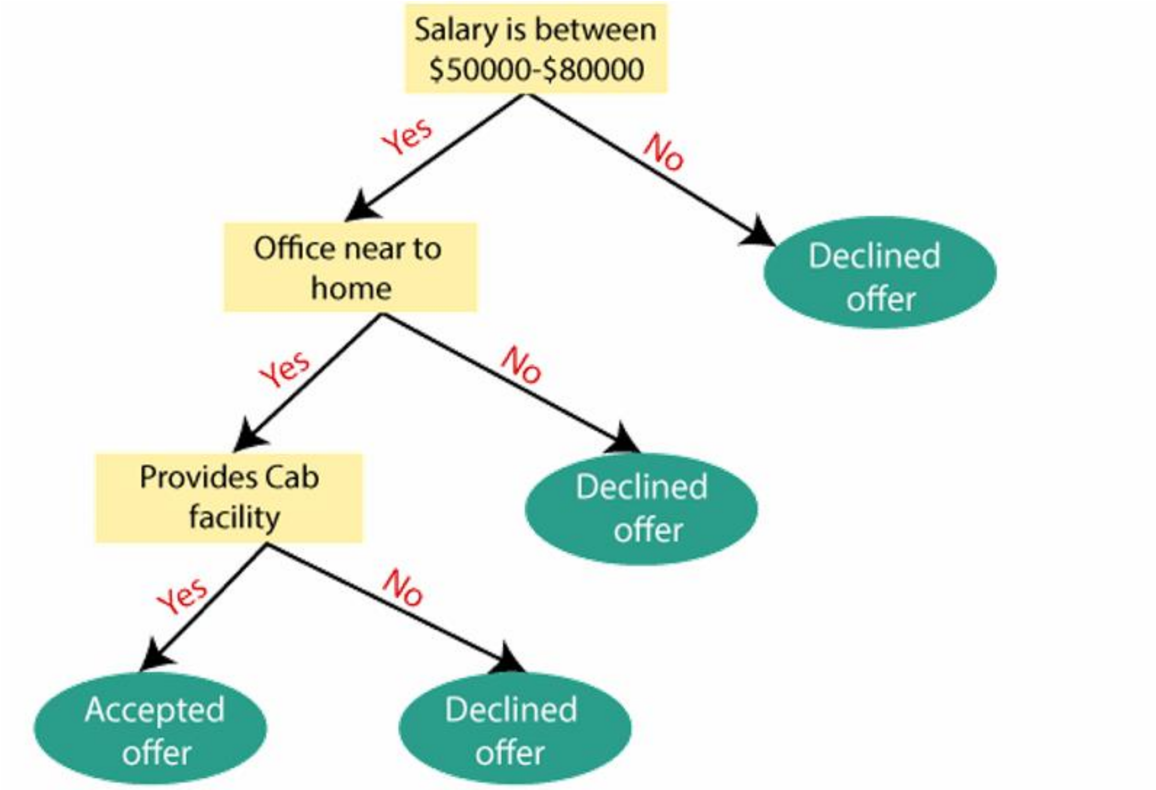
Important Parts of a Decision Tree

1. **Root Node** → The starting point of the tree.
2. **Decision Nodes** → Where decisions are made.
3. **Leaf Nodes** → The final result (Yes/No).
4. **Branches** → Paths showing decision rules.
5. **Splitting** → Dividing data into groups.
6. **Pruning** → Removing extra branches to keep the tree simple.

How a Decision Tree Works?

1. **Start with all the data** in the root node.
2. **Choose the best feature** to split data (like Age or Income).

3. **Divide the data** based on answers (Yes/No).
4. **Repeat the process** until all data is classified.



How to Choose the Best Feature?

📌 **Two Methods:**

1. **Information Gain** → More information means a better split.
2. Formula: IG=Entropy(S)−∑(Weighted Entropy of Features)

Entropy (measures impurity):

Entropy(S)=−P(yes)log2P(yes)−P(no)log2P(no)

3. **Gini Index** → Less impurity means a better split.

   Formula: GiniIndex=1−∑Pj2

Example: Will a Person Buy a Laptop?

Age	Income	Buys Laptop?
25	High	Yes
30	Medium	Yes
35	High	No
40	Low	No
45	Medium	Yes

1. **Find the best feature** (Age or Income) to start.
2. **Split the data** using that feature.
3. **Repeat until the decision tree is complete.**

Final Decision Tree Diagram

```
       Age?
      /   \
   <30     >=30
   / \     / \
Yes   No   Income?
      /     \
   High     Low/Medium
     |        |
    No       Yes
```

📌 **How to read this?**

- If **Age < 30** → Buys laptop ✅
- If **Age ≥ 30**, check **Income**:

  - **High Income** → Does not buy laptop ❌
  - **Low/Medium Income** → Buys laptop ✅

Advantages & Disadvantages

✅ **Good Things (Advantages)**
✓ Easy to understand.
✓ Works for numbers & categories.
✓ No complex math needed.

❌ **Bad Things (Disadvantages)**
✕ Can become too complex.
✕ Sensitive to bad data.

2 a) Describe classification techniques in supervised learning. [L2][CO1] [8M]

Classification in Supervised Learning

Classification is a method used in **Machine Learning** to **put things into groups**. The computer **learns from past data** and then **predicts the correct group** for new data.

For example, a **spam filter** in emails **learns from past emails** and **decides if a new email is spam or not**.

Types of Classification Methods

1. Logistic Regression

- Used when the answer is **Yes or No** (like pass/fail, spam/not spam).
- It draws a simple line to separate the two groups.

2. K-Nearest Neighbors (K-NN)

- Looks at the **closest neighbors** to decide the group.
- Example: If **most of your friends like a movie**, you might like it too!

3. Support Vector Machine (SVM)

- Finds the **best boundary** between two groups.
- It **creates a clear division** between categories.

4. Naïve Bayes

- **Uses probability** to decide the group.
- Example: It is used in **spam filters** to check if an email is spam.

5. Decision Tree Classification

- **Asks simple questions** step by step to make a decision.
- Example: If you want to **buy a phone**, you may ask:
    - **Is it affordable?**
    - **Does it have a good camera?**
    - **Is the battery life good?**
- Each question **helps in making a final decision**.

Ensemble Methods (Using Multiple Models Together)

1. Random Forest Classification

- **Uses many decision trees** to improve accuracy.
- It **takes the majority vote** to make the final decision.

2. Gradient Boosting Classification

- **Learns from mistakes** and improves step by step.
- Used in **stock market predictions** and **fraud detection**.

Conclusion

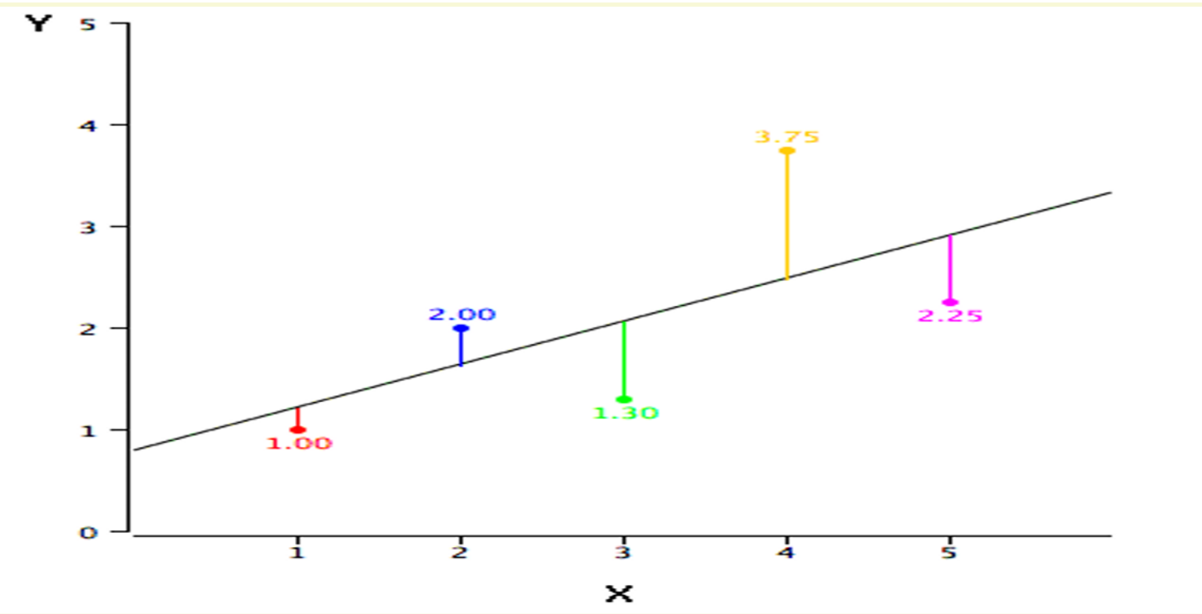
Classification helps computers **sort things into groups**. Different methods are used for different problems. It is very important in AI, business, and technology.

b List out various Regression techniques in Machine Learning. [L1][CO1] [6M]

Various Regression Techniques in Machine Learning:

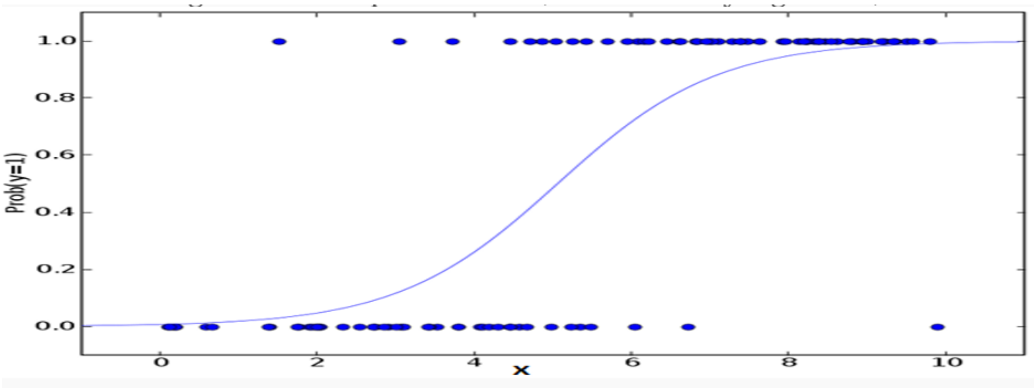
Regression is a technique used to predict continuous values, like predicting the temperature, stock prices, or sales. Here are some common regression techniques:

1. **Linear Regression**:
   - **What it does**: Predicts the value of a variable based on a straight line.
   - **How it works**: It finds a line (called the "best-fit line") that predicts the output using the input variable.
   - **Example**: Predicting house prices based on the size of the house.

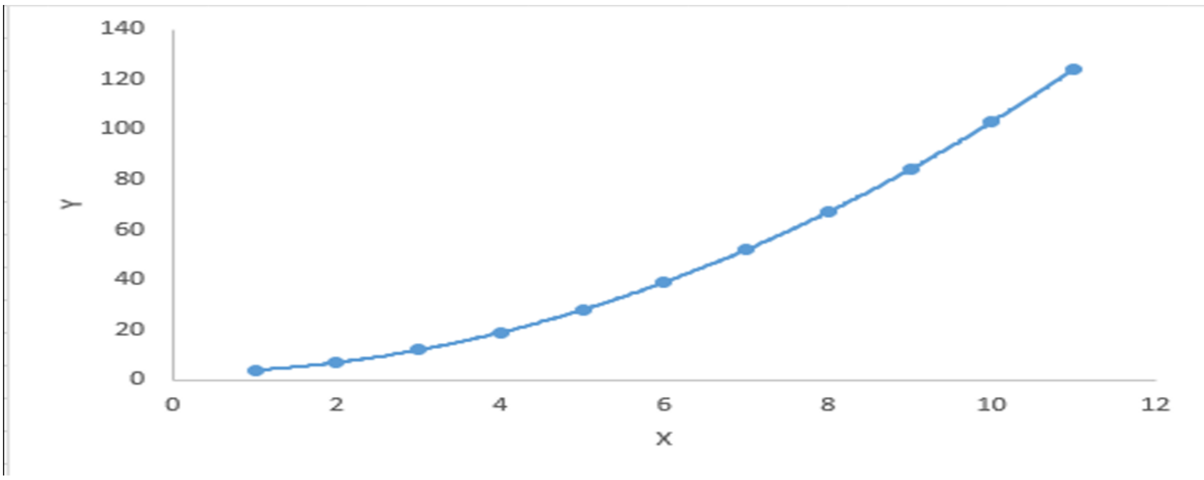


2. **Logistic Regression**:

- **What it does**: Used to predict a binary outcome (Yes/No, True/False).
- **How it works**: It estimates the probability of an event happening (e.g., success/failure).
- **Example**: Predicting if a student will pass or fail based on study hours.



3. **Polynomial Regression**:
   - **What it does**: Fits a curved line to the data instead of a straight one.
   - **How it works**: Uses an equation with powers of the input variable to predict the outcome.
   - **Example**: Predicting how temperature affects plant growth, where the relationship is not a straight line.

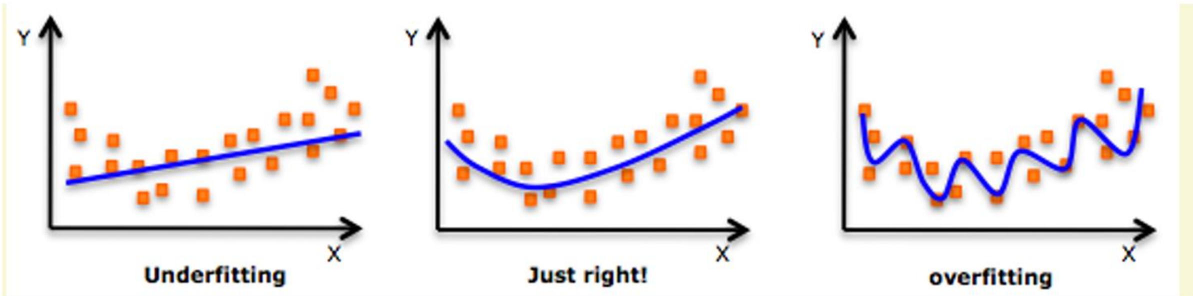


4. **Stepwise Regression**:

- **What it does**: Automatically selects the important features (variables) to make predictions.
- **How it works**: It adds or removes variables based on statistical tests to improve prediction.
- **Example**: Selecting which factors (e.g., age, income) affect sales the most.

Key Points:

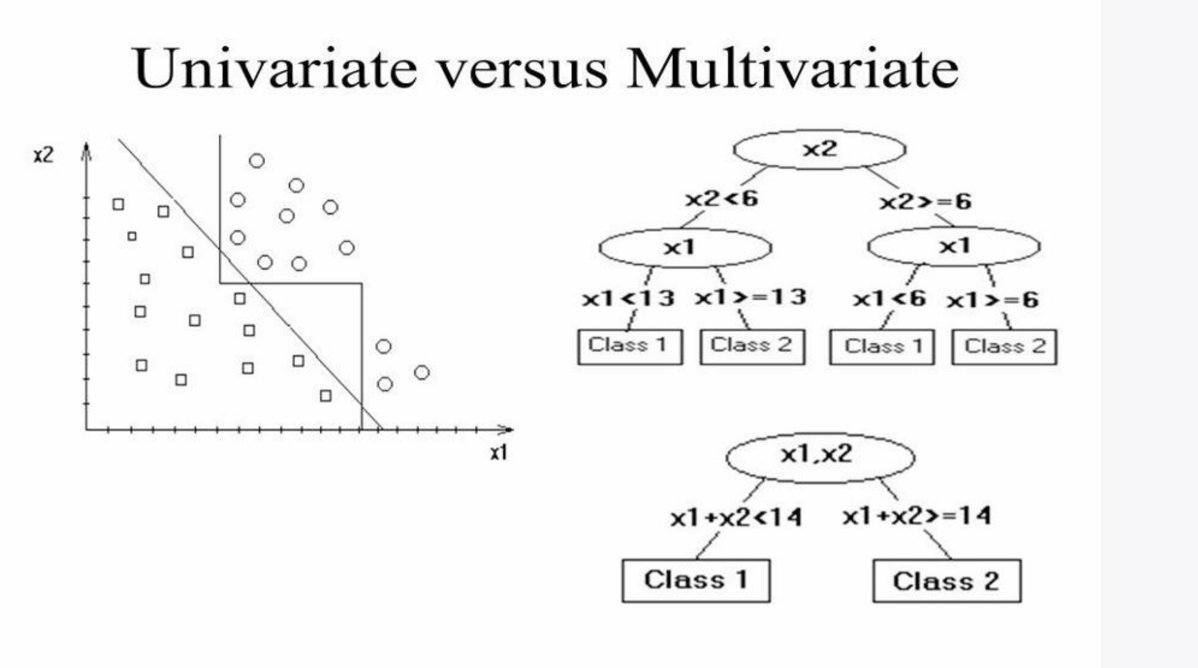
- Regression techniques help predict continuous values.
- **Linear regression** uses a straight line, while **polynomial regression** uses curves.
- **Logistic regression** is for predicting categories (Yes/No).
- **Stepwise regression** helps pick the best features for prediction.

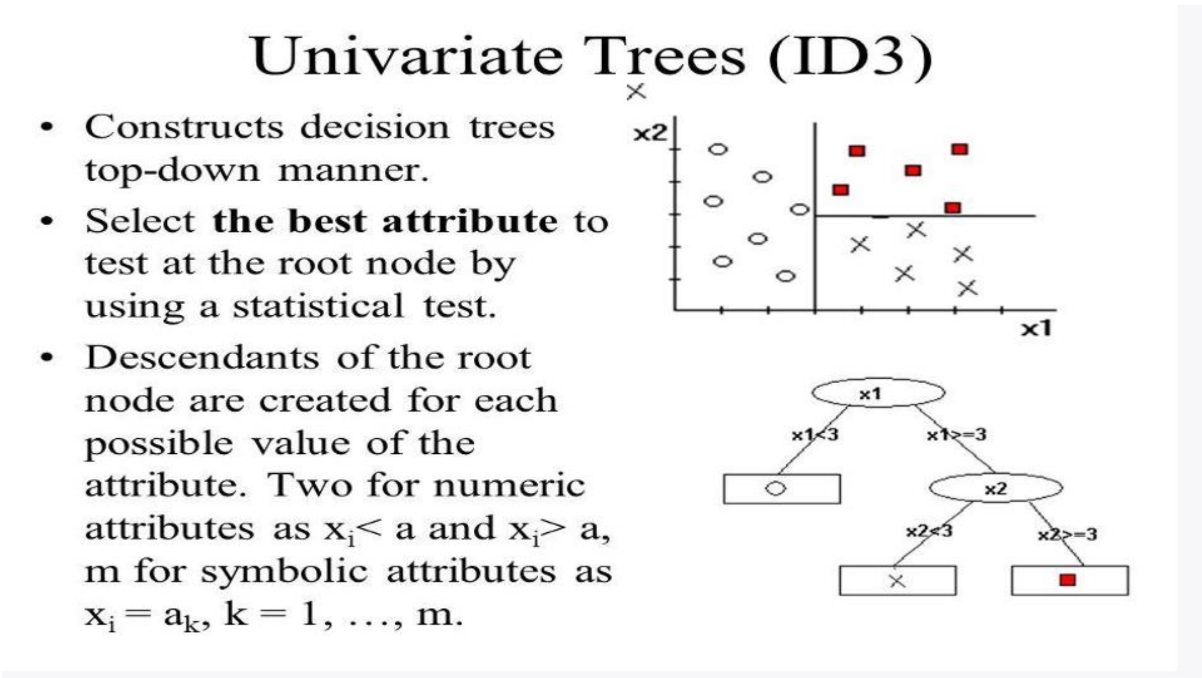


3.a) Compare Univariate and Multivariate Decision Trees. [L5][CO1][6M]

1. **Univariate Decision Tree**:
   - Uses **only one feature** (column of data) at each decision point to make a choice.
   - It's like asking one question at each step to divide the data.

2. **Multivariate Decision Tree**:
   - Uses **multiple features** at each decision point to make a choice.
   - It combines the answers from many trees to make a better decision.





Key Differences:

- **Univariate Tree**: Simple but can make mistakes if the data is complex.
- **Multivariate Tree**: More powerful, uses many trees and features, so it makes better and more accurate decisions.

In short:

- Univariate is simpler but can struggle with difficult problems.
- Multivariate is stronger and better for tough problems because it looks at more features and uses many trees to decide.

b) Explain about Pruning in supervised learning. [L2][CO1] [6M]

Pruning in Supervised Learning (Very Simple Explanation)

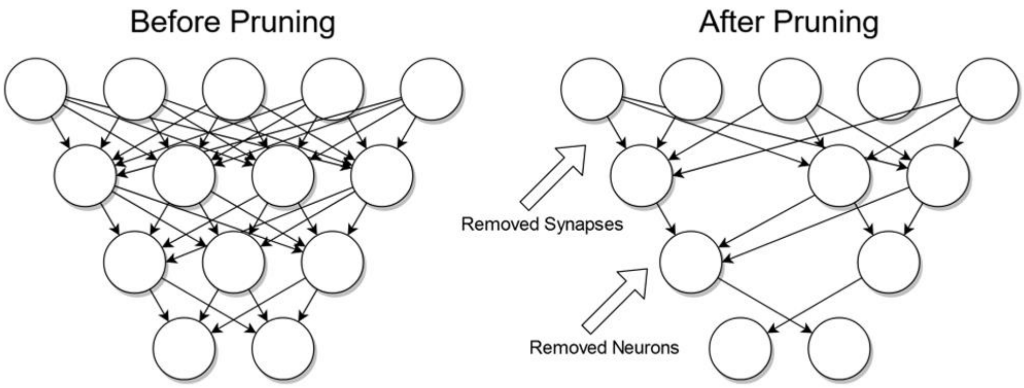
Pruning is when we make a decision tree simpler to help it make better predictions.

- **Why pruning?** Sometimes, a decision tree becomes too complicated and learns things that don't matter. This is called **overfitting**, and it makes the tree bad at predicting new data.

There are two types of pruning:

1. **Pre-Pruning (Stopping Early):**
   - We stop the tree from growing too big by setting rules, like saying "don't grow too deep" or "don't split if there's not enough data."
2. **Post-Pruning (Cutting After Growing):**
   - We let the tree grow fully, then cut out the parts that don't help in making predictions.

Why it's useful:
Pruning helps the tree be simpler and more accurate when predicting new data.



4. a) Differentiate various Parametric and Non-Parametric Methods. [L4][CO1] [6M]

Aspect	Parametric Methods	Non-Parametric Methods
Number of Parameters	Fixed number of parameters	Flexible number of parameters
Test Focus	Tests means (average values)	Tests medians (middle values)
Data Type	Works with variables (numbers only)	Works with both variables and attributes (numbers & categories)
Assumptions	Strong assumptions about data	Fewer assumptions about data
Data Requirement	Requires less data	Requires more data
Distribution Assumption	Assumes data follows a normal distribution	No assumption about data distribution
Effect of Outliers	Results are affected by outliers	Results are not much affected by outliers
Computation Speed	Faster to compute	Slower to compute
Examples	Logistic Regression, Naïve Bayes	KNN, Decision Tree

Super Simple Explanation of Bayesian Decision Theory in Supervised Learning

Bayesian Decision Theory helps us make the **best choice** when we are **not sure** about something. It uses **probabilities** and **prior knowledge** to improve decision-making.

How It Works (Step by Step)

1. **Start with a Guess (Prior Knowledge)**

- Before looking at any data, we already have an idea (guess) about what might happen.
- Example: You know it rains **30% of the time** in your city.

2. **Get New Data (Likelihood Estimation)**
    - We observe new information and calculate how likely something is.
    - Example: You see **dark clouds** in the sky, so rain seems more likely.
3. **Update Our Guess (Bayesian Inference)**
    - We combine what we knew before (**prior**) with what we just saw (**new data**).
    - Example: After seeing the clouds, now you think there's a **70% chance** of rain.
4. **Make the Best Choice (Decision Rule)**
    - We pick the most **likely** option based on updated probabilities.
    - Example: Since the chance of rain is high, you **take an umbrella**.
5. **Set Boundaries for Decision (Decision Boundary)**
    - We define rules to separate different choices.
    - Example: If the chance of rain is above **50%**, take an umbrella. If below **50%**, don't.
6. **Make the Best Decision (Optimal Decision-Making)**
    - We make a choice that **reduces mistakes** and **gives the best result**.

Bayes' Theorem Formula (Easy Explanation)

$$P(\omega|x) = \frac{P(x|\omega) * P(\omega)}{P(x)}$$

Where:

- **P(ω|x)** → Probability of class **ω** (e.g., rain) after seeing data **x** (e.g., clouds).
- **P(x|ω)** → Probability of seeing **x** (clouds) if **ω** (rain) happens.
- **P(ω)** → Initial belief (prior) before looking at new data.
- **P(x)** → Normalization factor to keep probabilities valid.

Mathematically it can be written as:

$$\omega_i = argmax_i[P(\omega_i|x)]$$

Bayes' Decision Rule

Final Decision Rule

- Choose the class (rain or no rain) with the **highest probability**.
- Example: If **P(Rain|Clouds) > P(No Rain|Clouds)** → **Decide it will rain**.

Why It's Important?

- Helps in **Machine Learning** to classify data correctly.
- Reduces **errors** and improves **accuracy**.

Super Simple Summary

1. **Start with what we know (prior).**
2. **See new data (likelihood).**
3. **Update our knowledge (posterior).**
4. **Make the best decision.**

5.Summarize the following models.

(i) Linear regression

(ii) Logistic regression

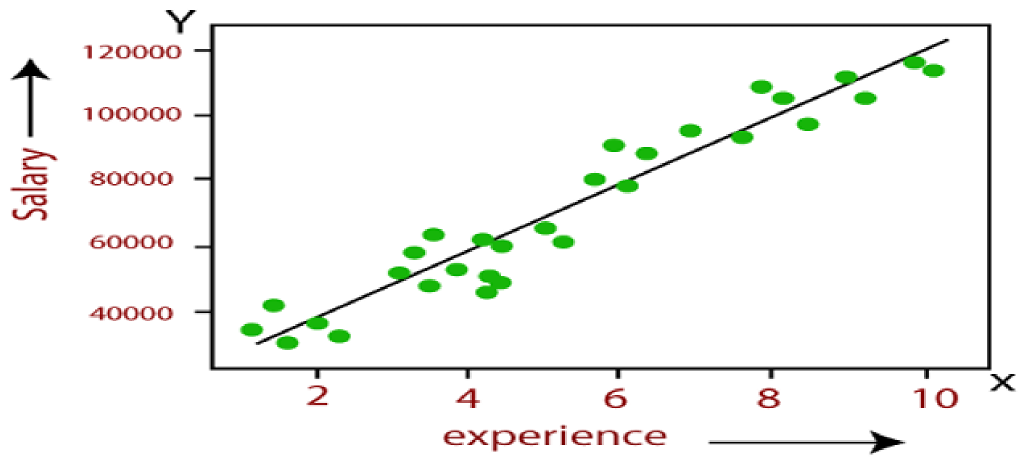
[L2][CO1] [12M]

(i) Linear Regression:

- Used to predict **continuous values** (e.g., salary, price, age).
- Finds a **straight-line** relationship between **independent** and **dependent** variables.
- If one independent variable → **Simple Linear Regression**.
- If multiple independent variables → **Multiple Linear Regression**.
- Equation:

$$y = a_0 + a_1 x + \varepsilon$$

Where a_0 and a_1 are constants, and ε is error.

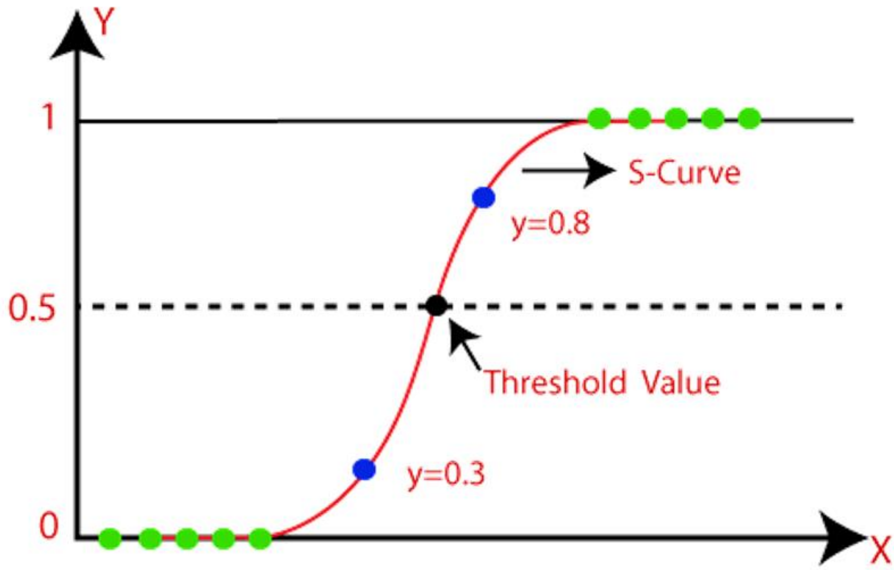


(ii) Logistic Regression:

- Used for **classification** (e.g., yes/no, 0/1, spam/not spam).
- Predicts a **categorical output** (not continuous).
- Uses a **sigmoid function** to give values between **0 and 1** (probabilities).
- Example: Will it rain today? (Yes = 1, No = 0).
- Equation:

$$P(Y=1) = \frac{1}{1+e^{-(a_0+a_1 x)}}$$

Where a_0 and a_1 are constants.



- The equation for logistic regression is:

$$\mathbf{\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n}$$