

# Classification & Prediction

Aditya Raj

RA2311003030555

CSE(Core)-I

# What is Classification?

## **Content:**

- Classification is a supervised learning technique.
- It assigns data into predefined categories (labels).
- Uses training data to build a model.

## **Examples:**

- Email → Spam / Not Spam
- Transaction → Fraud / Legitimate

# What is Prediction?

## **Content:**

- Prediction estimates unknown or future values.
- Uses historical data to forecast outcomes.
- Often done using regression techniques.

## **Examples:**

- Predict house prices
- Forecast sales revenue

# Decision Tree Induction

## **Content:**

- A tree-like model used for classification.
- Nodes represent attributes.
- Branches represent decisions.
- Leaves represent class labels.

## **Advantages:**

- Easy to understand
- Visual representation
- Handles both numeric & categorical data

# Attribute Selection Measures

## **Content:**

Used to choose the best attribute for splitting.

## **Common Measures:**

- Information Gain
- Gain Ratio
- Gini Index

## **Goal:**

Select attribute that best separates data.



# ID3 & CART Algorithms

## **Content:**

### **ID3 Algorithm**

- Uses Information Gain.
- Works with categorical data.
- Builds tree top-down.

### **CART Algorithm**

- Uses Gini Index.
- Supports numeric data.
- Produces binary trees.

# Tree Pruning

## Content:

- Removes unnecessary branches from decision tree.
- Reduces overfitting.
- Improves model accuracy on new data.

## Types:

- Pre-pruning (stop early)
- Post-pruning (trim after building)

# Bayes' Theorem & Naive Bayes

## **Content:**

### **Bayes' Theorem:**

$$P(A|B) = (P(B|A) \times P(A)) / P(B)$$

### **Naive Bayes Classifier:**

- Assumes attributes are independent.
- Simple and fast.
- Works well for text classification.

**Example:** Spam detection



# Support Vector Machines (SVM)

## **Content:**

- Supervised learning algorithm for classification.
- Finds optimal boundary (hyperplane) between classes.
- Maximizes margin between data points.

## **Advantages:**

- High accuracy
- Effective in high-dimensional data

# K-Nearest Neighbors (KNN)

## **Content:**

- Lazy learning algorithm.
- Classifies based on nearest neighbors.
- Uses distance metrics (Euclidean distance).

## **Steps:**

1. Choose K value.
2. Find nearest neighbors.
3. Assign majority class.

**Thank You !**