Training Day-6 Report:

Decision Tree

- **Decision trees** are a popular and powerful tool used in various fields such as machine learning, data mining, and statistics. They provide a clear and intuitive way to make decisions based on data by modeling the relationships between different variables. This article is all about what decision trees are, how they work, their advantages and disadvantages, and their applications.

What is a Decision Tree?

A **decision tree** is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions. Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test, and each leaf node corresponds to a class label or a continuous value.

Structure of a Decision Tree

1. **Root Node**: Represents the entire dataset and the initial decision to be made.
2. **Internal Nodes**: Represent decisions or tests on attributes. Each internal node has one or more branches.
3. **Branches**: Represent the outcome of a decision or test, leading to another node.
4. **Leaf Nodes**: Represent the final decision or prediction. No further splits occur at these nodes.

How Decision Trees Work?

The process of creating a decision tree involves:

1. **Selecting the Best Attribute**: Using a metric like Gini impurity, entropy, or information gain, the best attribute to split the data is selected.
2. **Splitting the Dataset**: The dataset is split into subsets based on the selected attribute.
3. **Repeating the Process:** The process is repeated recursively for each subset, creating a new internal node or leaf node until a stopping criterion is met (e.g., all instances in a node belong to the same class or a predefined depth is reached).

Metrics for Splitting

- **Gini Impurity**: Measures the likelihood of an incorrect classification of a new instance if it was randomly classified according to the distribution of classes in the dataset.
    - Gini=1−∑i=1n(pi)2Gini=1−$\sum_{i=1n}(p_i)_2$, where *pi* is the probability of an instance being classified into a particular class.
- **Entropy**: Measures the amount of uncertainty or impurity in the dataset.

    - Entropy=−∑i=1$n$$p_i$log2($p_i$)Entropy=−∑$_{i=1n}$$p_i$log2($p_i$), where *pi* is the probability of an instance being classified into a particular class.
- **Information Gain**: Measures the reduction in entropy or Gini impurity after a dataset is split on an attribute.
    - InformationGain=Entropyparent−∑i=1n(|Di||D|∗Entropy(Di))InformationGain =Entropyparent−∑$_{i=1n}$(|D||D_i|∗Entropy(D_i)), where *Di* is the subset of *D* after splitting by an attribute.

Applications of Decision Trees

- **Business Decision Making**: Used in strategic planning and resource allocation.
- **Healthcare**: Assists in diagnosing diseases and suggesting treatment plans.
- **Finance**: Helps in credit scoring and risk assessment.
- **Marketing**: Used to segment customers and predict customer behavior.