Decision Trees:-

They are the type of supervised machine learning model used for classification and regression tasks.

Structure :-

1) Nodes (feature / attribute)
2) Edges (outcome of a test / decision)
3) Leaves (final output)

Mathematics:-

1) Information gain
2) Impurity reduction.

Impurity Measures:-

To decide the best split,

i) Gini Impurity / Index:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$

→ p_i is the proportion of samples c is the classes.

ii) Entropy :-

$$Entropy(D) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Information gain:-

measures

It gains the reduction in impurity after a split.

$$IG(D, A) = Impurity(D) - \sum_{k=1}^{k} \frac{|D_k|}{|D|} Impurity(D_k)$$

D = original dataset
A = attribute used for split.
D_v = Subset of data for each split.
|D| = Total no. of samples.

Tree Pruning:-

Pruning reduces removes branching for overfitting.

Methods:-

Cost-Complexity Pruning:

$$\text{Minimize} \rightarrow \text{Cost}(T) = \sum_{m \in \text{leaves}(T)} \frac{N_m}{N} \text{Impurity}(D_m) + \alpha |T|$$

→ |T| is no. of leaves & α is the regularization parameter.

Prediction:-

Classification:-

→ Traverse the trees using the feature values of the input sample
→ The leaf nodes determine the predicted class.

Regression:-

→ Traverse the tree to the leaf node.
→ The output is the mean (or median) of the target values.

Optimization:-

→ greedy algos (CART) classification & Regression trees