## Algorithms

ID3 (Iterative Dichotomiser 3)

Uses entropy and information gain.

Limitations: No pruning, categorical-only features.

## C4.5

Improves ID3 by:

Handling continuous features (discretization).

Supporting missing values.

Using gain ratio to reduce bias toward high-branching features.

## **CART (Classification and Regression Trees)**

Uses Gini impurity for classification.

For regression, splits minimize Mean Squared Error (MSE).

Step-by-Step Learning Process

Start with the entire dataset at the root.

For each feature, calculate splitting criterion (e.g., Gini).

Select the feature with the best split.

Repeat recursively for child nodes.

Example: CART for Iris Dataset

Split 1: Petal Length  $\leq$  2.45 cm  $\rightarrow$  Class "Setosa."

Split 2: Petal Width > 1.75 cm  $\rightarrow$  Class "Virginica."

**Overfitting Solutions** 

Pre-pruning: Stop splitting if node depth > threshold.

Post-pruning: Remove unnecessary branches after full growth (e.g., cost-complexity pruning).

## **Practical Considerations**

Hyperparameters: max\_depth, min\_samples\_split.

Scalability: Use approximate algorithms (e.g., RainForest) for big data.