Introduction to Decision Trees

A decision tree is a flowchart-like structure used for decision-making and predictive modeling. It splits data into branches to reach a conclusion, much like a tree with branches.

Each internal node represents a 'test' on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

Decision trees are commonly used in classification and regression problems.

How Decision Trees Work

1. **Splitting**: The dataset is split into subsets based on an attribute value test.

2. **Decision Node**: When a sub-node splits into further sub-nodes, it's called a decision node.

3. **Leaf/Terminal Node**: Nodes that do not split are called Leaf or Terminal nodes.

4. **Pruning**: The process of removing sub-nodes of a decision node.

The most common algorithms to build a decision tree are ID3, C4.5, CART, and CHAID.

Advantages and Disadvantages

Advantages:

- Easy to understand and interpret.

- Requires little data preprocessing.

- Can handle both numerical and categorical data.

Disadvantages:

- Can create over-complex trees that do not generalize well (overfitting).

- Sensitive to noisy data.

- Greedy algorithms like ID3 do not guarantee the globally optimal tree.