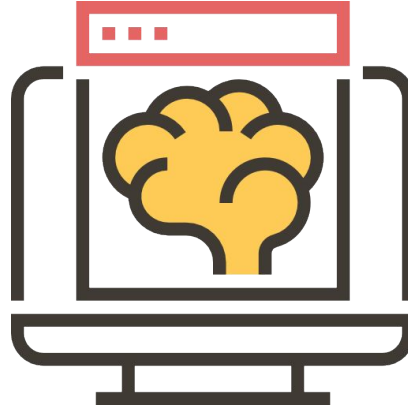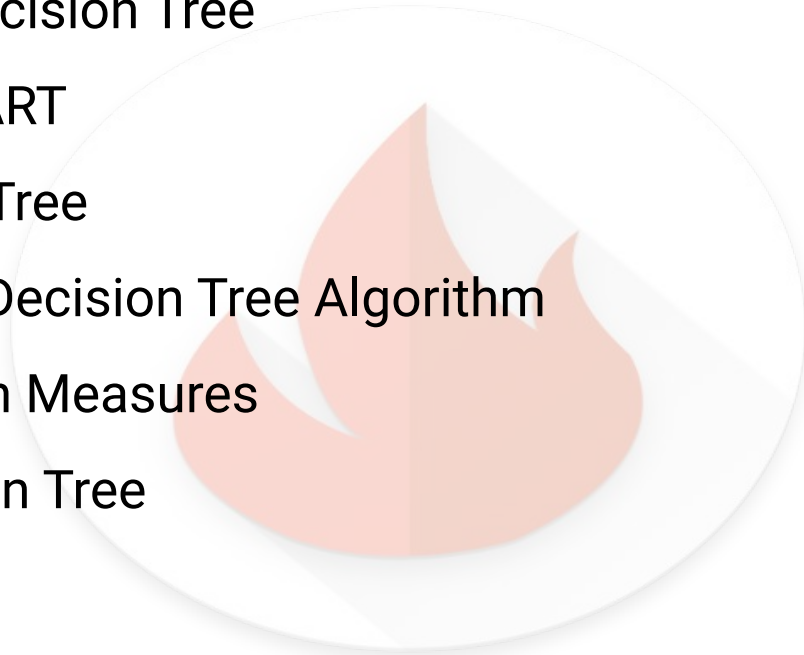# Supervised Learning Classification

Decision Tree

Agenda

- Introduction to Decision Tree

- Introduction to CART

- Uses of Decision Tree

- Terminologies in Decision Tree Algorithm

- Attribute Selection Measures
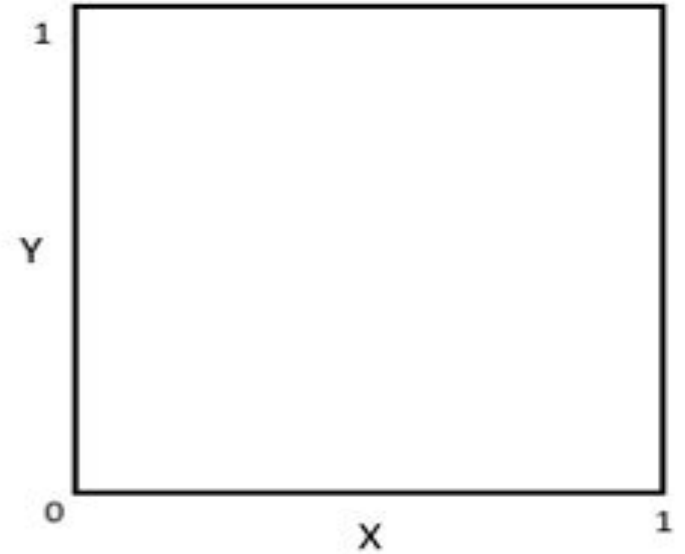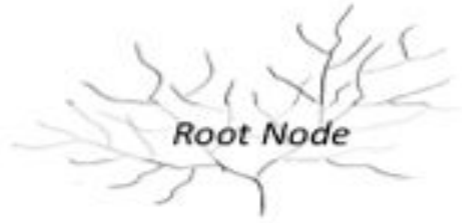
- Pruning in Decision Tree

Introduction to Decision Tree

1. It is Tree-structured classifier, where internal node represent feature of a dataset, branches represent the decision rules and each leaf node represent the outcome.

2. CART As the name goes, it uses tree like model of a decision.

3. It is a supervised learning technique.

4. Also used for Both technique.

5. There are two nodes, first is **'Decision Node'** and another is **'Leaf Node'**.

Introduction to Decision Tree

1. Decision node used to make any decision, and have a multiple branches for example:- A student is passed based on mark criteria.

2. Leaf node are the output of those decision and do not contain any further branches.

3. **It is graphical representation for getting all the possible solutions to a decision on given conditions.**

4. It is called 'Decision Tree' because, similar to tree.

5. In order to Build a tree, we use the **CART** Algorithm, which stands for **Classification and Regression Tree Algorithm.**

Decision Tree Representation

Decision Tree Representation

Branches

Decision Node → Root Node

Decision Node

Decision Node

Leaf Node

Leaf Node
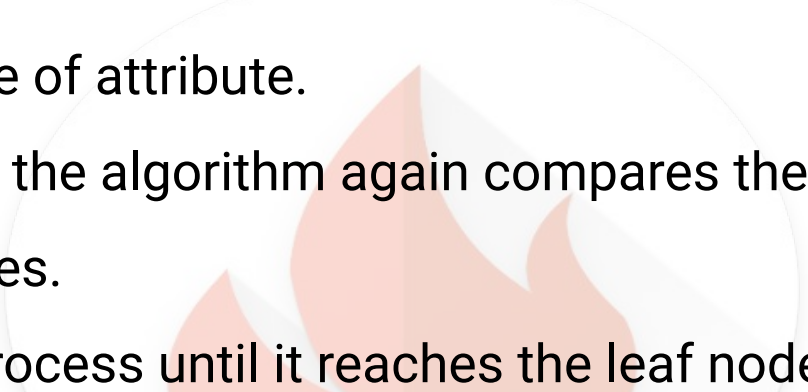
Leaf Node

Decision Node

Leaf Node

Leaf Node
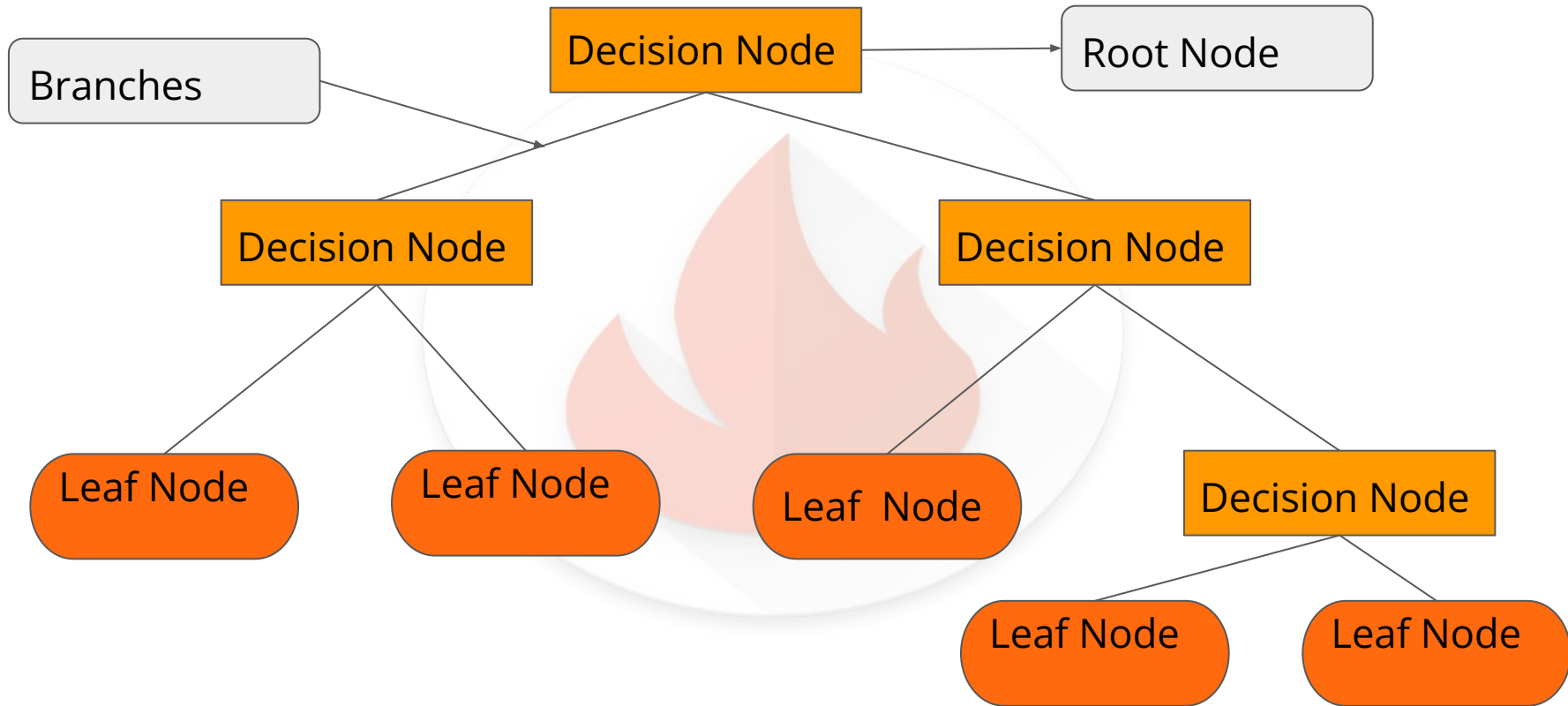
Uses of Decision Tree

1. Various algorithm in ML, so why Decision Tree?

2. Two reason for using the decision tree.

   a. Decision Trees usually mimic human thinking ability making a decision , so it is easy to understand.

   b. The logic behind the decision tree can be easily understand because it shows a tree-like structure.

Decision Tree Terminologies

1. **Root Node:-** Root node is from where the decision tree starts.

2. **Leaf Node:-** Leaf nodes are the final output.

3. **Sub Tree:-** A tree formed by splitting the tree.

4. **Pruning :-** Pruning is the process of removing the unwanted branches from the tree.

5. **Child Node :-** The root node of the tree is called the parent node, and other node are called the child nodes.
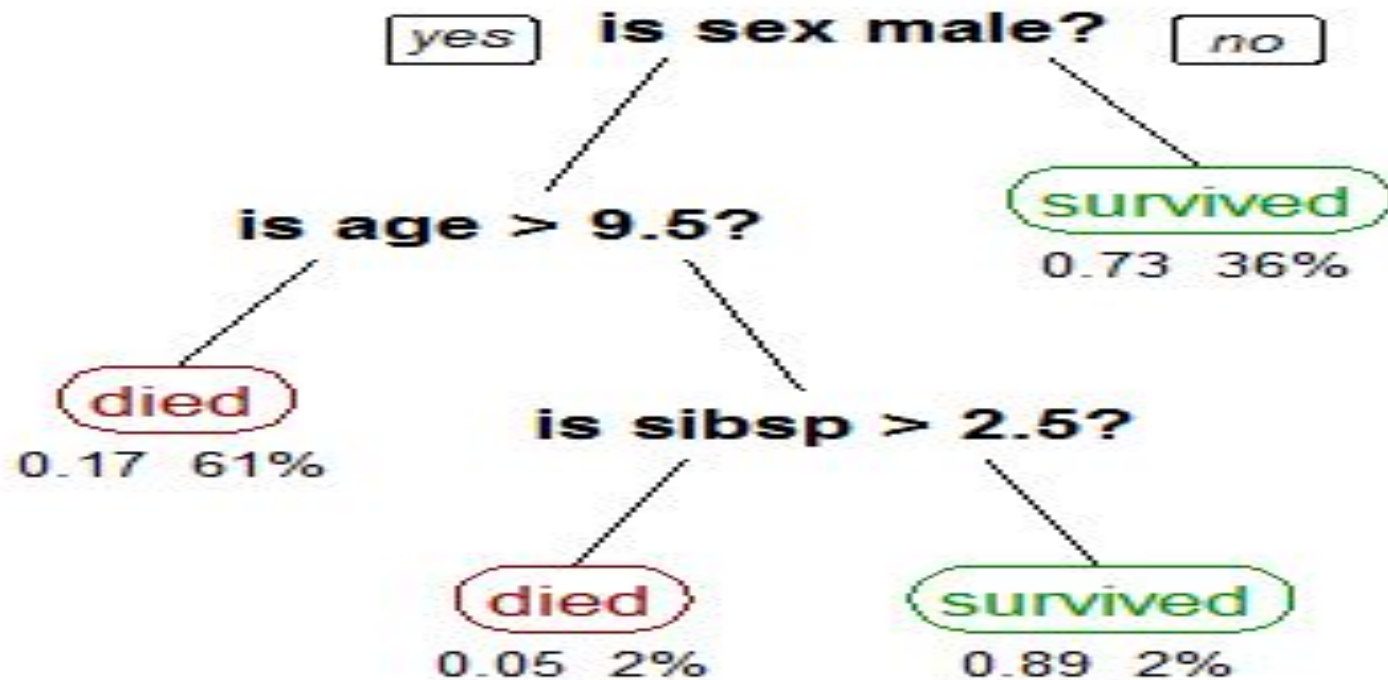
How does work?

1. Algorithm start from root node of tree.

2. Compare the value of attribute.

3. For the next node, the algorithm again compares the attribute value with the other sub-nodes.

4. It continues the process until it reaches the leaf node of the tree.

Basic Steps

1. **Step 1:-** Begin the tree with the root node, says S which contains the complete dataset.

2. **Step 2:-** Find the best attribute in the dataset using ASM.(Attribute Selection Measure)

3. **Step 3:-** Divide the S into subsets that contain possible values for the best attributes.

4. **Step 4:-** Generate the decision tree node, which contains best attribute.

5. **Step 5:-** Continue the process until a stage is reached.

Basic Example

is sex male? yes / no

- yes → **is age > 9.5?**
  - **died** 0.17 61%
  - **is sibsp > 2.5?**
    - **died** 0.05 2%
    - **survived** 0.89 2%
- no → **survived** 0.73 36%

Attribute Selection Measures

1. **Information Gain:-**

    a. It is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.

    b. It Calculates how much information a feature provides us about a class.

    c. Based on information gain, we split the node.

    d. Decision Tree always tries to maximize the value of information gain.

Attribute Selection Measures

Information Gain = Entropy(S) - [(Weighted Avg) * Entropy (each feature)]

Entropy

Entropy is a metric to measure the impurity in a given attribute.

Entropy calculated as,

S = -P(yes) log2 P(yes) - P(no) log2 P(no)

S = Total Sample

Attribute Selection Measures

Gini Index:-

1. It is measure of impurity or purity.

2. It measure an incorrect classification of new instance.

3. An attribute with the low gini index should be preferred as compared to the high gini index.

4. It only creates binary splits.

5. Gini index calculated as,

   Gini Index = 1 - sum(P)^2

Attribute Selection Measures

Gini Index:-

1. It is measure of impurity or purity.

    a. **Pure:-** A selected sample of dataset all data belongs to same class.

    b. **Impure:-** Data is mixture of different classes.
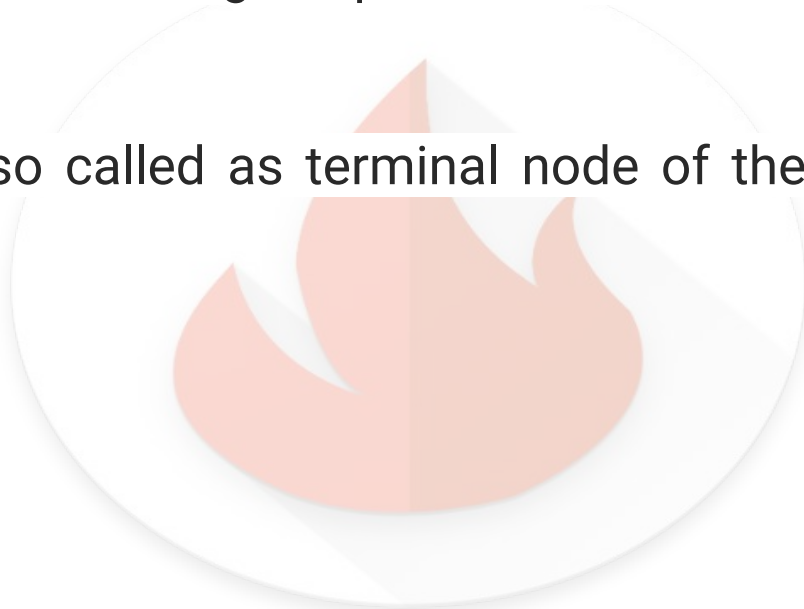
Pruning in Decision Tree

1. Pruning is the process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

2. A technique that decreases, the size of the learning tree without reducing accuracy is know as 'pruning'.

3. **Two types**

   a. Cost Complexity Pruning

   b. Reduced Error Pruning

Pruning in Decision Tree

1. Pruning can start at either root or the leaves. The simplest method of pruning starts at leaves and removes each node with most popular class in that leaf,

2. A technique that decreases, the size of the learning tree without reducing accuracy is know

Representation

1. A node represents is a single input variable X and a split point on that variable.

2. The leaf node also called as terminal node of the tree contain output variable Y.

Thank you