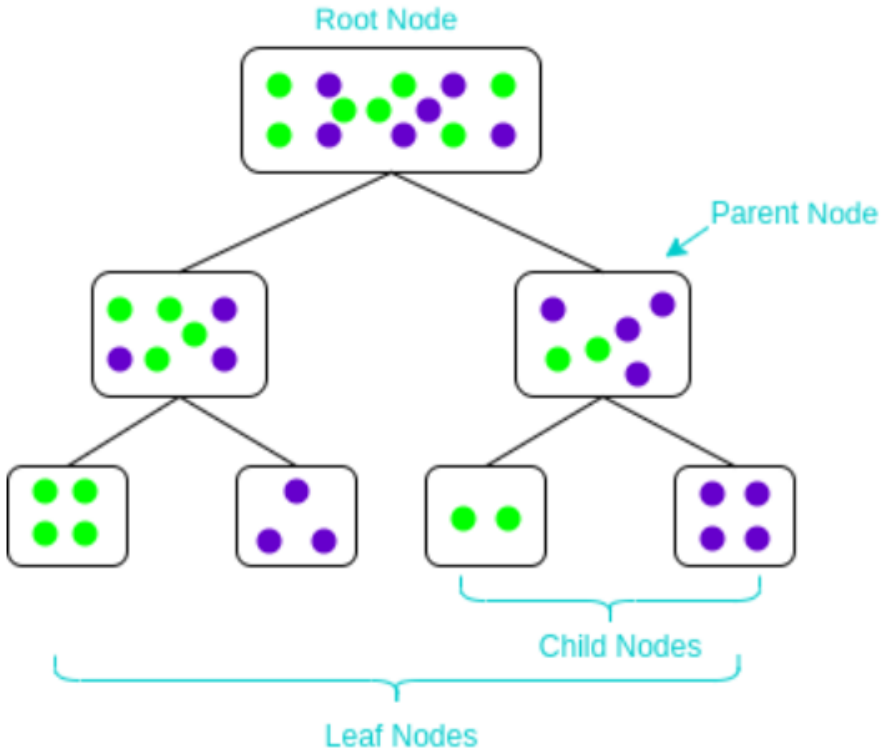
Growing a Decision Trees

Decision Tree Recap



Node splitting in a Decision Tree?

- **Node splitting is the process of dividing a node into multiple sub-nodes to create relatively pure nodes.**

- Node splitting can be broadly classified into 2 types, based on target variables.

Node splitting Types :

- Regression problems (Target variable is Continuous)
  - Reduction in Variance

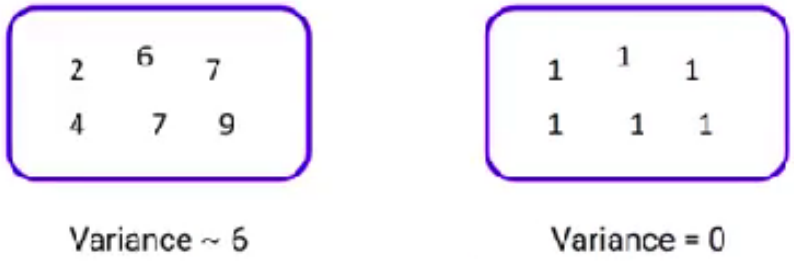
- Classification problems (Target variable is Discrete)
  - Entropy / Information Gain
  - Gini Impurity

Reduction in Variance:

- Method of splitting the node when the target variable is continuous.

$$Variance = \frac{\sum (X - \mu)^2}{N}$$

- If a node is entirely pure, then the variance is zero.



Steps to calculate variance of a split:

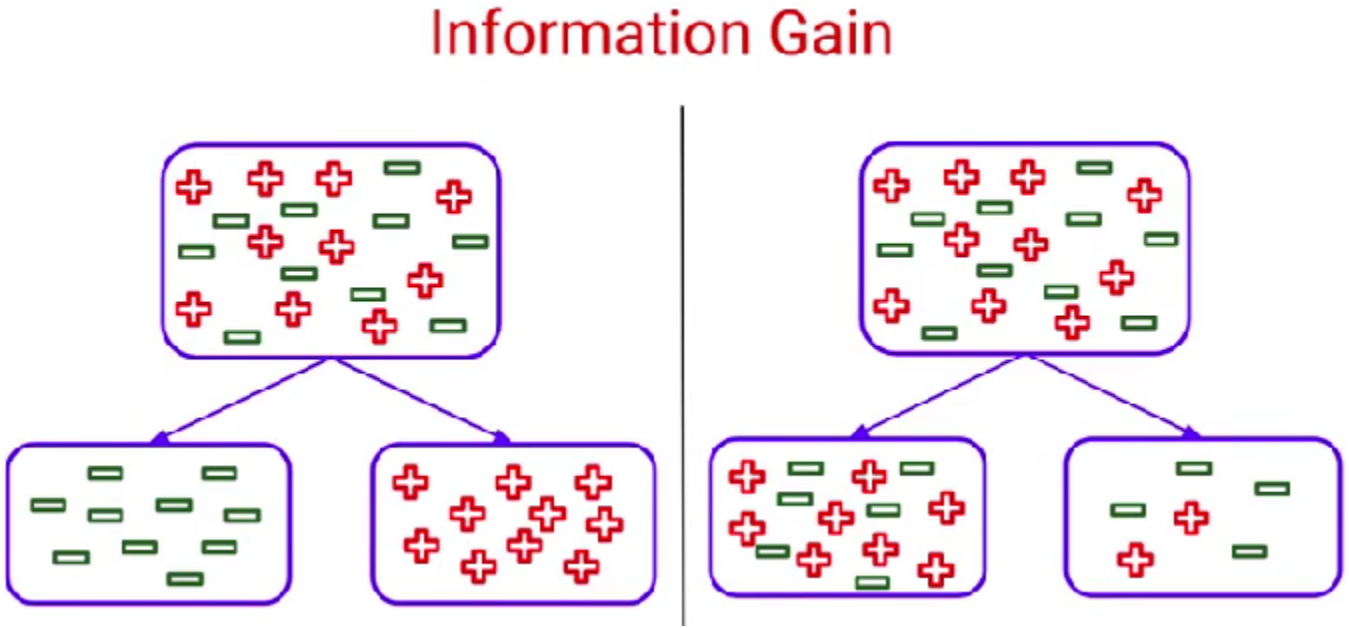
- Calculate the variance of each child node.
- Calculate the variance of each split as weighted average variance of each child node.
- Select the split with the lowest variance.
- Perform steps 1-3 until completely pure nodes are achieved.

Information Gain

- Information Gain is used for splitting the nodes when the target variable is categorical.
- IG = 1 - Entropy
- Entropy is used for calculating the purity of a node.
- **Lower the value of entropy, higher is the purity of the node.**

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Information Gain



Steps to calculate Entropy:

- Calculate the entropy of each child node.
- Calculate the entropy of each split as weighted average entropy of each child node.
- Select the split with the lowest entropy.
- Perform steps 1-3 until completely pure nodes are achieved.

Gini Impurity

- Gini impurity is used for splitting the nodes when the target variable is categorical.
- GI = 1 - Gini
- Gini says that if we pick 2 points at random then both the pts belong to the same class.

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2$$

- **Lower the Gini Impurity, higher the purity of the node.**

Steps to calculate Gini Impurity:

- Calculate the gini impurity of each child node.
- Calculate the gini impurity of each split as weighted average gini impurity of each child node.
- Select the split with the lowest gini impurity.
- Perform steps 1-3 until completely pure nodes are achieved.