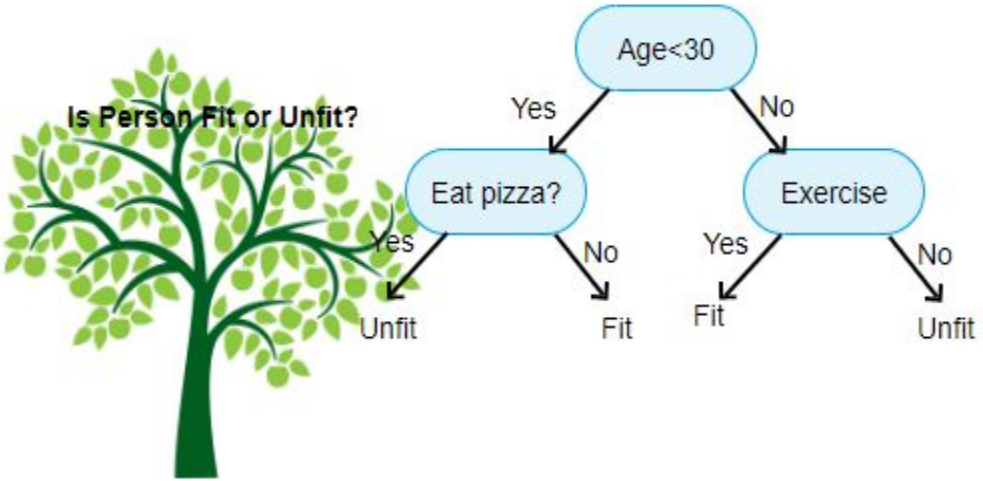
5)Decision Tree (C/R):

--> Decision tree is a supervised type machine learning algorithm and it builds classification or regression models in the form of a tree structure.
--> The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches.
--> Leaf node represents a classification or decision.
--> The topmost decision node in a tree which corresponds to the best predictor called root node.



Decision tree algorithm uses following 2 method to create the decision tree:

1. ID3 (Iterative Dichotomiser 3) → uses Entropy function and Information gain as metrics.

2. Gini Index or Gini impurity

ID3

Process of selecting nodes in decision tree:

1. Find the Information gain of the target attribute.

2. Find the Entropy of rest attributes.

3. Find gain for each and every rest attribute.

Gain = IG - Entropy

4. The attribute that has highest gain will be selected as root node.

Information gain - Information gain is used for determining the best features/attributes that render maximum information about a class.

Entropy - Entropy is the measurement of impurities or randomness in the data points. It basically ranges between 0-1.

0- Pure, 1-Impure
Gini Index

Gini Index - It measures the impurity of the nodes. ranges between 0-1.

Gini Index/Gini Impurity = 1-Gini

1 - (/Sum of squared probability of each class)

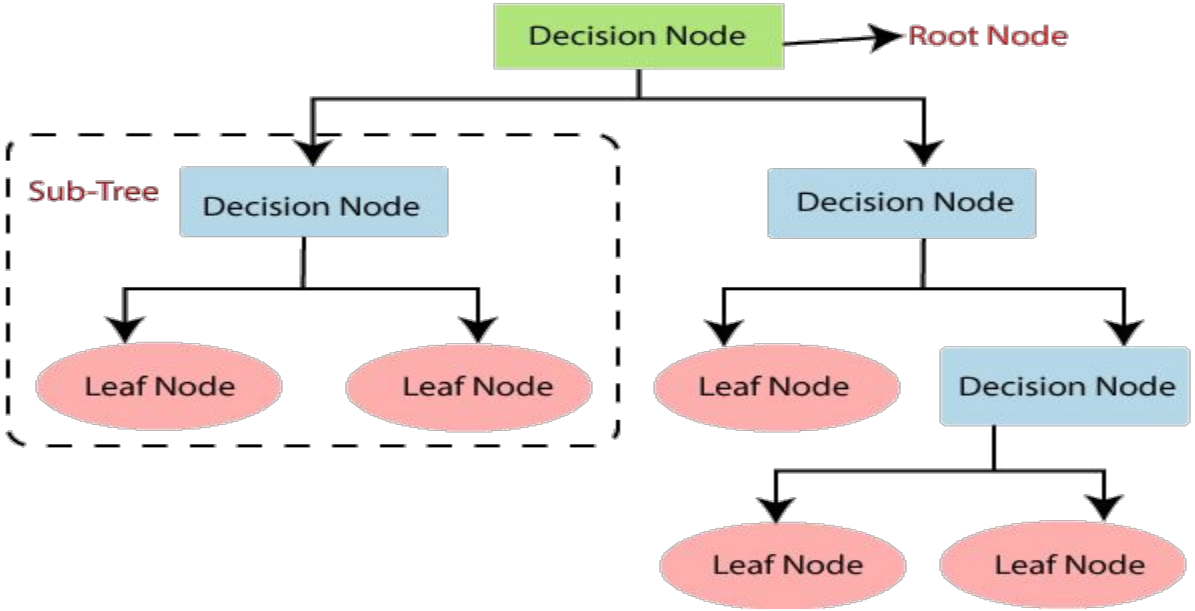
0 - Data points/feature are pure. Eg. All Yes

1 - Data points/feature are impure(Randomness) eg. Yes, No both

0.5 - Equal distribution of classes. 5 yes, 5 no

--> Gini ranges from zero to one, as it is a probability and the higher this value, the more will be the purity of the nodes.
--> lesser value means lesser pure nodes.
--> 1 - pure, 0 - randomness

-->Gini works only in those scenarios where we have categorical targets. It does not work with continuous targets.attribute having least gini impurity value will be selected as root node.

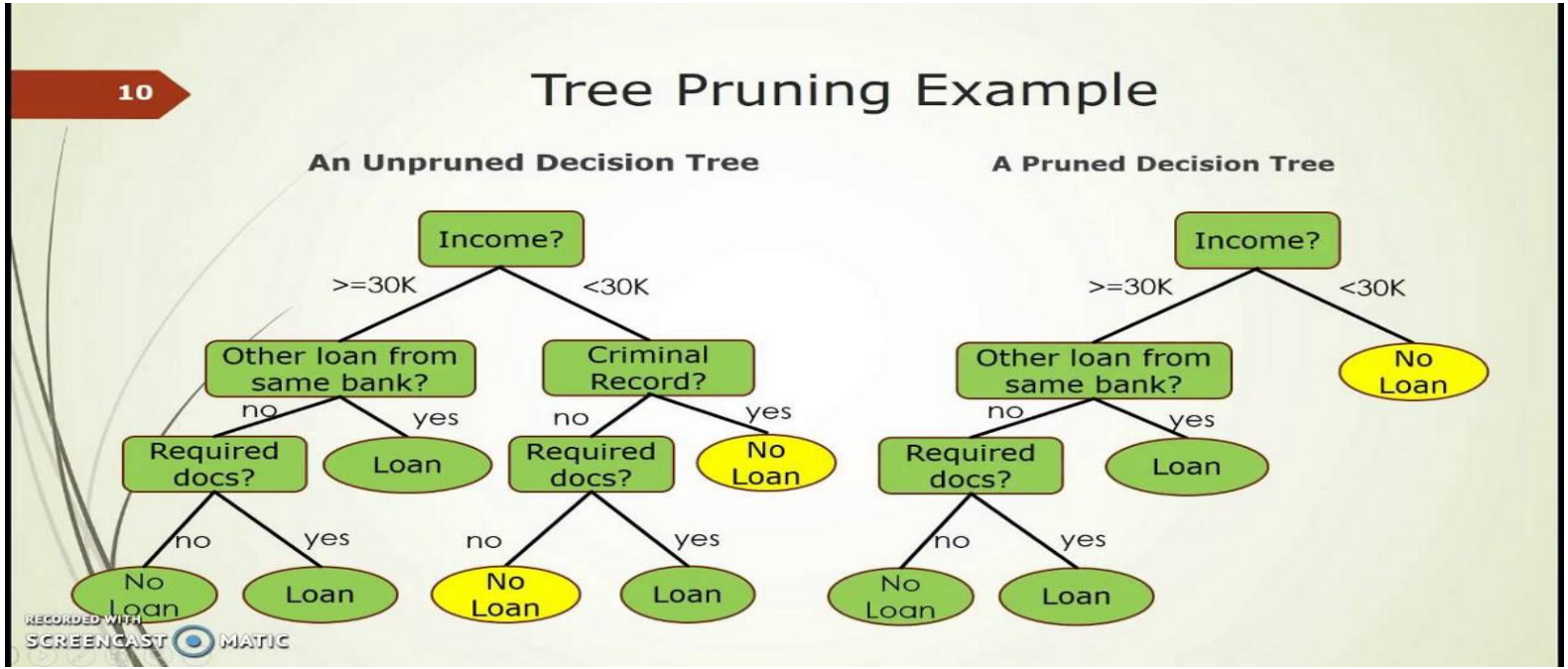


Pruning in decision tree:

--> Pruning is one of the techniques that is used to overcome problem of overfitting in decision tree.

--> Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch.

--> It reduces the size of a Decision Tree which might slightly increase training error but drastically decrease your testing error,hence making it more adaptable.



Types of pruning:
--> Pre-pruning that stop growing the tree earlier, before it perfectly classifies the training set.
--> Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.

Advantages:

1. Easy to read and interpret.

2. Easy to prepare.

3. Less data cleaning required.

Disadvantages:

1. Unstable nature

2. Less effective in predicting the outcome of a continuous variable