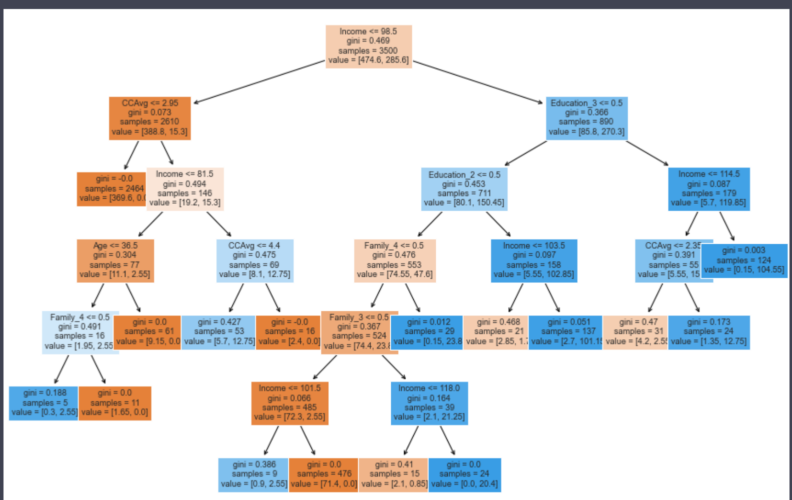
DECISION TREE

The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data.A decision tree is built in the top-down fashion. The decision nodes represent the question based on which the data is split further into two or more child nodes. The tree is created until the data points at a specific child node is pure . The criteria for creating the most optimal decision questions is different for different algorithms (CART,C4.5,ID3,CHAD)



ROOT NODE

The top-most node of the tree from where the tree starts.The first split which decides the entire population or sample data should further get divided into two or more homogeneous sets.

LEAF NODE

The node representing the data segment having the highest homogeneity (purity).The leaf node is reached when no significant information gain happens.

DECISION NODE

One or more decision nodes that result in the splitting of data in multiple data segments. The goal is to have the children nodes with maximum homogeneity (purity).

ENTROPY

Entropy can be defined as a measure of impurity of data. The higher the entropy, the higher impure the data is. The maximum value for entropy is 1. The value of entropy close to 0 represents the fact that data is pure.ID3 &C4.5 uses entropy.

$$Entropy = \sum_j -p_j \log_2 p_j$$

GINI IMPURITY

Gini impurity is the probability of a random sample being classified correctly if you randomly pick a label according to the distribution in the branch.
CART uses gini impurity.

$$Gini\ Index = 1 - \sum_j p_j^2$$

INFORMATION GAIN

Information gain is the difference between the entropy of a data segment before the split and after the split. Higher the difference implies the lower entropy, higher information gain, and better the feature used for the split. ID3 & C4.5
Information Gain = Entropy(parent) - Weighted Sum of Entropy(Children).

ADVANTAGES

- Decision Trees are simple to understand.
- Requires less data preparation as it is not influenced by outliers.
- Can handle both continuous and categorical variables.
- Works with both linearly and nonlinearly related variables.

DISADVANTAGES

- Decision Trees can easily overfit.
- Large trees can become over-complex and difficult to interpret.
- Decision Trees can be unstable with small variations in data.
- A decision tree can be computationally expensive to train.
- Can create biased learned trees if some classes dominate.

Created by www.linkedin.com/in/yogitadarade