Decision Trees / Random Forest

Decision trees

Professional School
ISO 9001 : 2000 Organisation

What is a Decision Tree?

Decision tree is a type of supervised learning technique (having a pre-defined target /dependent variable) that is mostly used in **classification problems**.

It works for both categorical and continuous input and output variable

Types of Decision trees

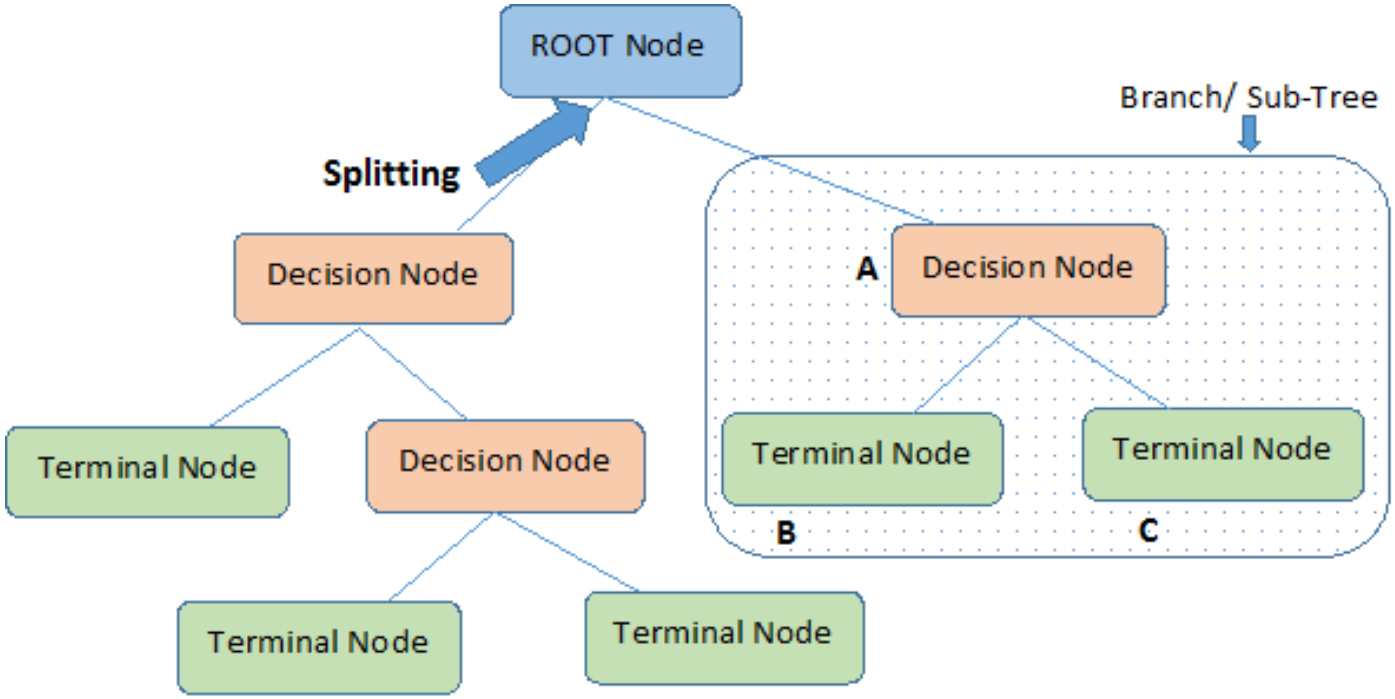
1. Categorical Variable Decision Tree (CHAID):
Decision Tree which has categorical target variable then it called as categorical variable decision tree. For example, when the target variable is “Whether a Student will play football or not” i.e. YES or NO.

2. **Continuous Variable Decision Tree (CART):**
Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Key Terms in Decision Trees

1. **Root Node: I**t represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.

Key Terms in Decision Trees



Note:- A is parent node of B and C.

Decision tree- Example

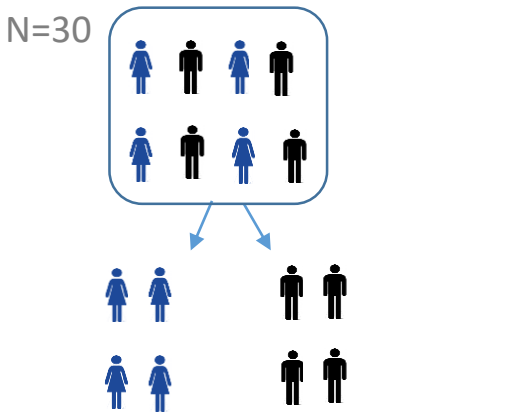
Let's say we have a sample of 30 students with three variables:

- Gender (Boy/ Girl),
- Course (Statistics/Economics)
- Height (5 to 6 ft)

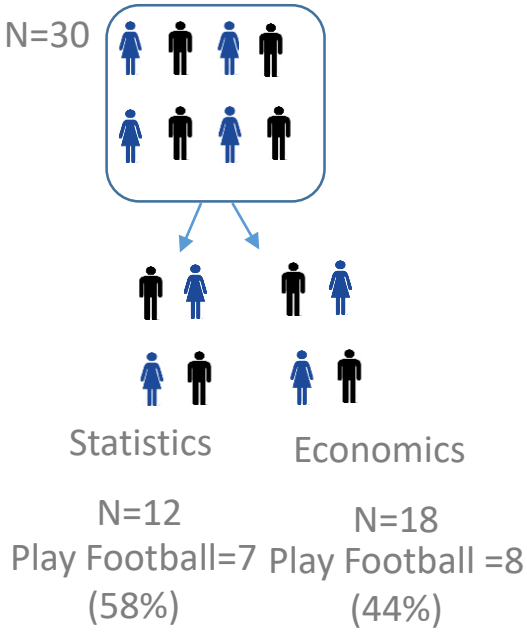
15 out of these 30 play Football in leisure time.

Now, I want to create a model to predict *who will play football during leisure period*?

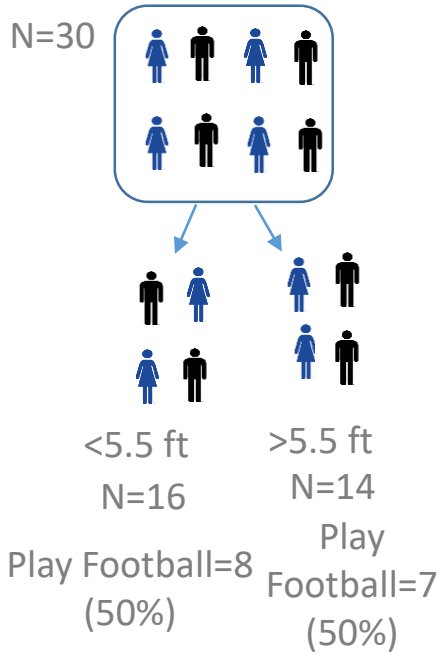
Split on Gender



Split on Course



Split on Height

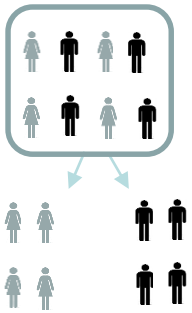


Decision tree- Example (Continued)

Decision Tree identifies the most significant variable and its value, that gives homogenous set of population.

Split on Gender

N=30



Female
N=10

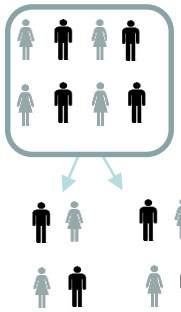
Play Football=2
(10%)

Male
N=20

Play Football=13
(65%)

Split on Course

N=30

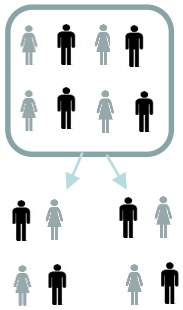


Statistics
N=12
Play Football=7
(58%)

Economics
N=18
Play Football=8
(44%)

Split on Height

N=30



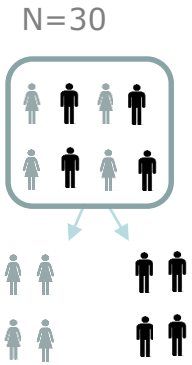
<5.5 ft
N=16
Play Football=8
(50%)

>5.5 ft
N=14
Play Football=7
(50%)

Decision tree- Example (Continued)

Decision Tree identifies the most significant variable and its value, that gives homogenous set of population.

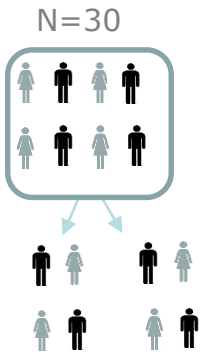
Split on Gender



Female
N=10
Play Football=2 (10%)

Male
N=20
Play Football=13 (65%)

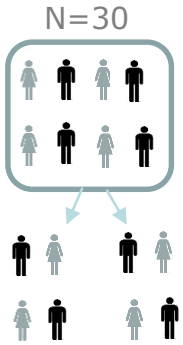
Split on Course



Statistics
N=12
Play Football=7 (58%)

Economics
N=18
Play Football=8 (44%)

Split on Height



<5.5 ft
N=16
Play Football=8 (50%)

>5.5 ft
N=14
Play Football=7 (50%)

CART (Chi-square Automatic Interaction Detection

Decision tree- CART

CART-Classification and Regression Tree

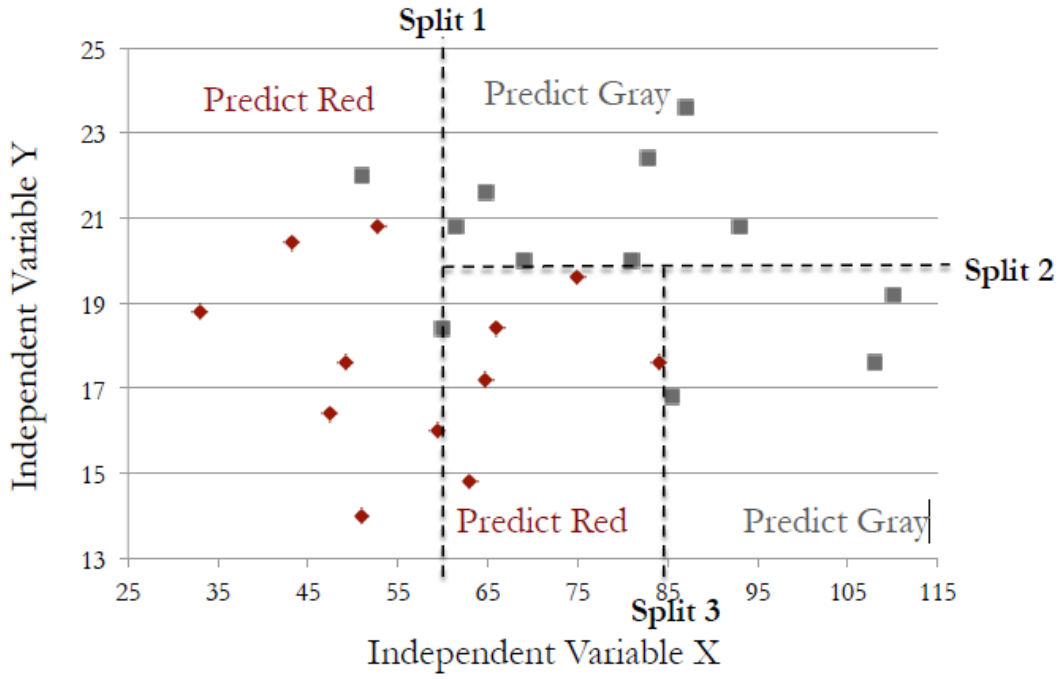
Decision Tree, which has continuous target/dependent variable then it is called as Classification and Regression Tree (CART).

-

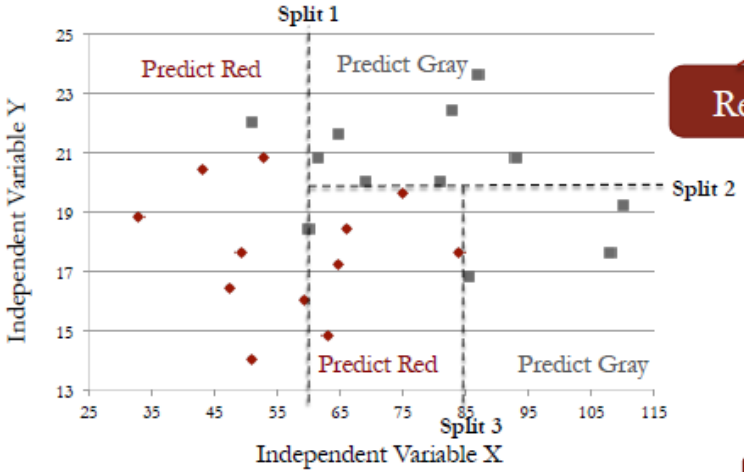
The CART model, usually follows the below steps

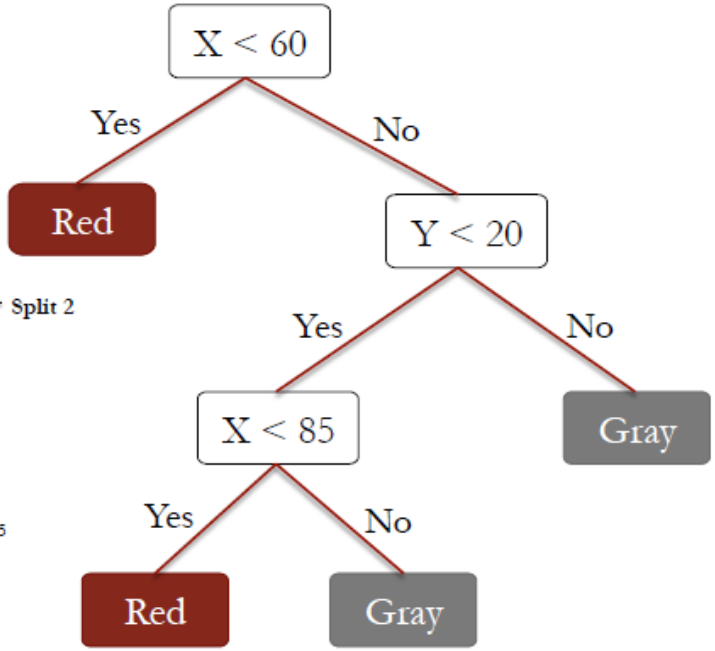
- Build a tree by splitting on variables
- To predict the outcome for an observation, follow
- the splits and at the end, predict the most frequent
- Outcome
- Does not assume a linear model
- Interpretable

Splits in CART



Final Tree





Components of Nodes

All the nodes, except the leaf nodes (colored terminal nodes), have 5 parts:

1. **Question asked about the data based on a value of a feature**. Each question has either a True or False answer that splits the node. Based on the answer to the question, a data point moves down the tree.

2. **Gini:** The Gini Impurity of the node. The average weighted Gini Impurity decreases as we move down the tree.

3. **samples:** The number of observations in the node

4. **value**: The number of samples in each class. For example, the top node has 2 samples in class 0 and 4 samples in class 1.

5. **class:** The majority classification for points in the node. In the case of leaf nodes, this is the prediction for all samples in the node

How to Split Nodes?

Making strategic splits drastically affects a **tree's accuracy.**

The decision criteria are different for classification and regression trees. Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of the resultant sub-nodes.

In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

How to Split Nodes – GINI INDEX

Gini index says, if we randomly select two items from a population, they must be of the same class and probability for this is 1 if the population is pure.

It works with the categorical target variable “Success” or “Failure”. It performs only binary splits. Higher the value of Gini, higher the homogeneity. CART (Classification and Regression Tree) uses the Gini method to create binary splits.

Steps to Calculate Gini for a split

Calculate Gini for sub-nodes, using formula sum of the square of probability for success and failure (p^2+q^2).

Calculate Gini for split using weighted Gini score of each node of that split..

How to Split Nodes – INFORMATION GAIN

We can derive information gain from entropy as 1 – Entropy. Entropy is a way of measuring the amount of impurity in a given set of data. It is represented by a formula:

$$H = -\sum_i p_i(\log_2 p_i)$$

The ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided it has the entropy of one.

Overfitting: Or Why a Forest is better than One Tree

The reason the decision tree is prone to overfitting when we don't limit the maximum depth is because it has unlimited flexibility, meaning that it can keep growing until it has exactly one leaf node for every single observation, perfectly classifying all of them.

If you go back to the image of the decision tree and limit the maximum depth to 2 (making only a single split), the classifications are no longer 100% correct. We have reduced the variance of the decision tree but at the cost of increasing the bias.

As an alternative to limiting the depth of the tree, which reduces variance (good) and increases bias (bad), **we can combine many decision trees into a single ensemble model known as the random forest.**

Random Forests

Random Forests: Introduction

Why Random Forest?

- Designed to improve prediction accuracy of CART
- Can be used to perform both regression and classification tasks
- Falls in the category of ensemble methods, where weak models are aggregated to form a reliable model

What is Random Forest?

- Random Forest is a substantial modification of a bagging technique (based on random samples of data)

How does it works?

- Works by building a large number of CART trees
- To make a prediction for a new observation, each tree "votes" on the outcome, and we pick the outcome that receives the majority of the votes and in case of regression (continuous variable) it takes the average of output

Random Forests

Let there be N no. of cases in the data and K features (variables)

1. Each tree is split based on random sample of n from N but with replacement. This sample become the training set for growing the tree

For example:

Original Data: 1,2,3,4,5

New ''data'' for growing tree: 2,4,5,3,1 (1st random sample), 3,1,5,4,2 (2nd random sample) and so on…

2. Minimum no. of random selected k from K input variable, which are again used at each node of the tree

3. Each tree is grown as much as possible and prediction happens by aggregating the predictions of 'n' trees

Thank you

Professional School
ISO 9001 : 2000 Organisation