DATA SCIENCE WITH PYTHON : DECISION TREE ALGORITHM #1126

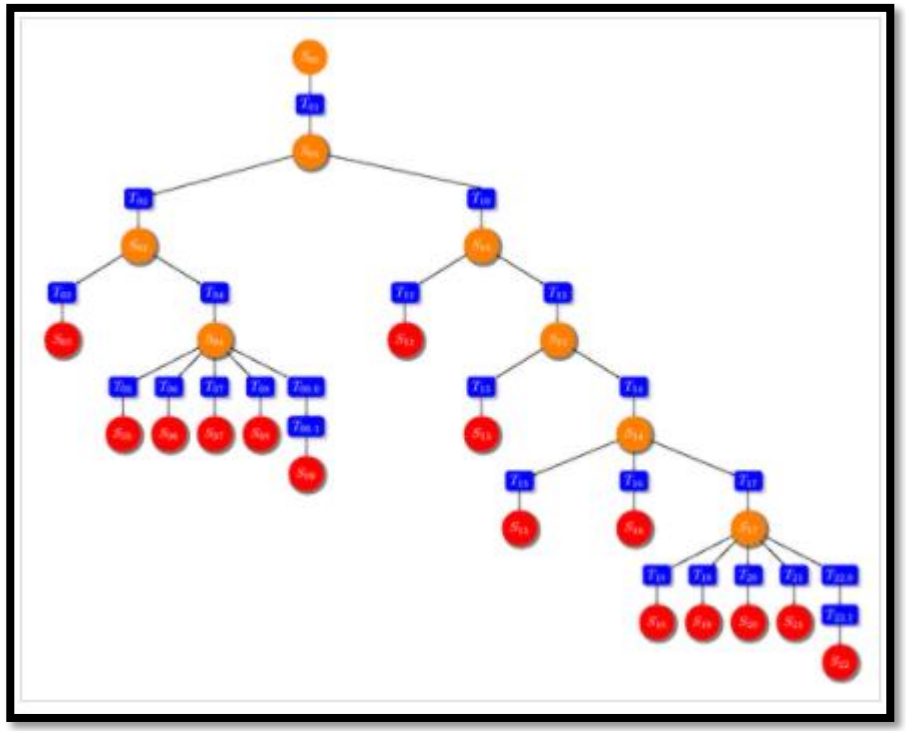
Presented by: Deepthi M

Batch Number: 05

Serial Number: 172

Decision Tree

A decision tree is a supervised machine learning model. It is called a decision tree because it takes decisions that help in splitting the dataset and it is called a tree because resembles a tree structure.

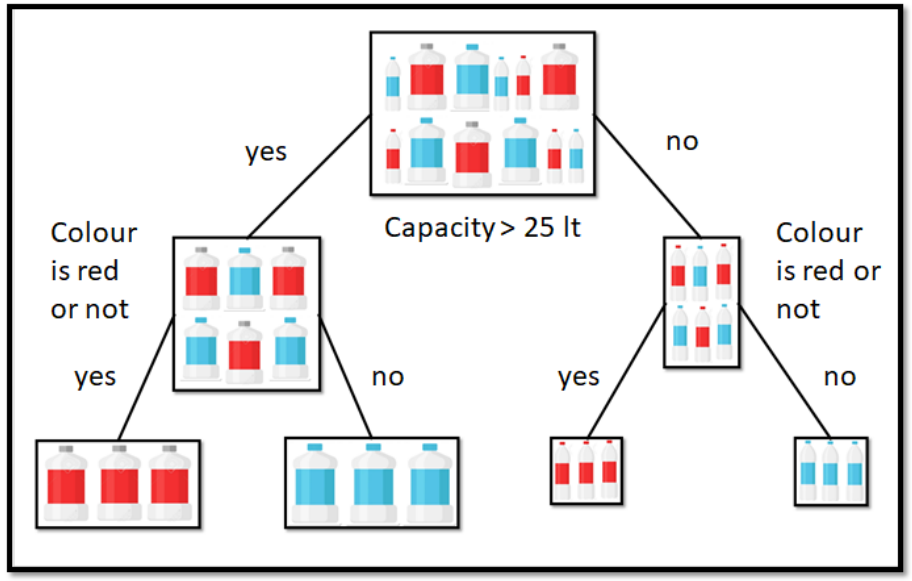


Example:

Now let's look into an example of how the decision tree algorithm train and splits the data set.

Sample dataset:

Name	Capacity in lt	Colour	No. of Bottles
Bottle A	25	Red	3
Bottle B	2	Blue	3
Bottle C	25	Blue	3
Bottle D	2	Red	3



Some of the questions that arise are:
What exactly decision tree is?
What are the measures used in the decision tree while splitting the dataset?
When decision tree is used?
What are the terminologies used in the decision tree?
Why it is known as a greedy algorithm?
When to stop splitting of decision tree?
What are the advantages?
What are the disadvantages?

Definition:

The decision tree splits the dataset based on a decision/condition/question. It continuously splits the data until it reaches a threshold value which is upfront decided or naturally stops.

Measures used:

•Information gain
•Gini index

Information gain:

•It is calculated based on entropy. Information gain is the difference in entropy before and after splitting of the data set.
•The high value of information gain is preferred in the splitting of the data set.
Formula:
•On a board basis:

Entropy(before) - Entropy(after)

•Exactly:

Entropy= -P(yes)log2 P(yes)- P(no) log2 P(no)
Where, P(yes)= probability of yes, P(no)= probability of no.

Gini Index:

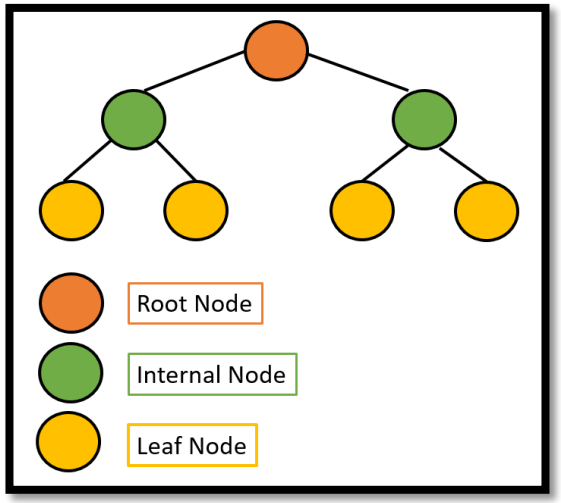
•It is based on the purity or impurity of the data set.
•It only splits into binary nodes.
•The low value of the Gini index is preferred in the splitting of data.
•It is used in the CART algorithm (classification and regression tree algorithm)
Formula:
•subtracting the sum of squared probabilities of each class from one.
•1 - ∑ ((p(x))^2)

When it is used:

•It is used in both classification and regression problems. But mostly used for classification problems.

Terminologies to know:

1.The decision tree has root nodes, internal nodes, and leaf nodes.



Root node:

•It consists of the input data which gets divided into further nodes.

Internal node:

•It resembles the decision of the root node.

Leaf node:

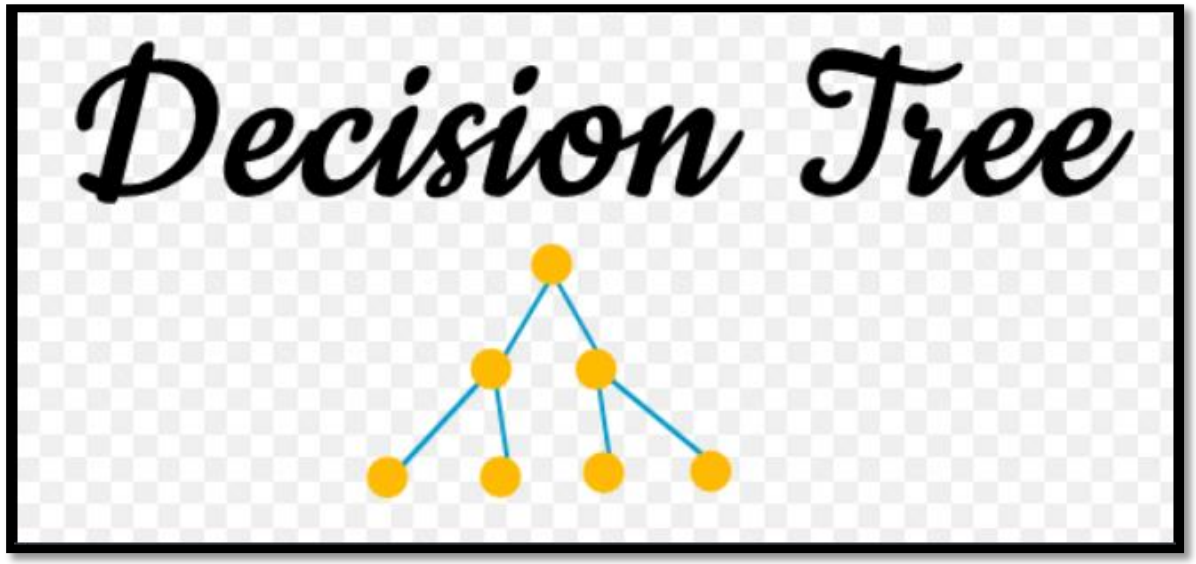
•It represents the output of the data which can be numeric or categorical.

2.Entropy:

•Randomness in the data set.

A decision tree is a greedy algorithm:

•It is called a greedy algorithm because at every node it takes a decision and splits into further nodes which is a characteristic of a greedy algorithm. Hence, it is called a greedy algorithm.



Decision tree stopping criteria:

•Setting a threshold value for the entropy, so when values are less than the threshold value the growth is stopped.

•Pre Setting the depth of the tree.

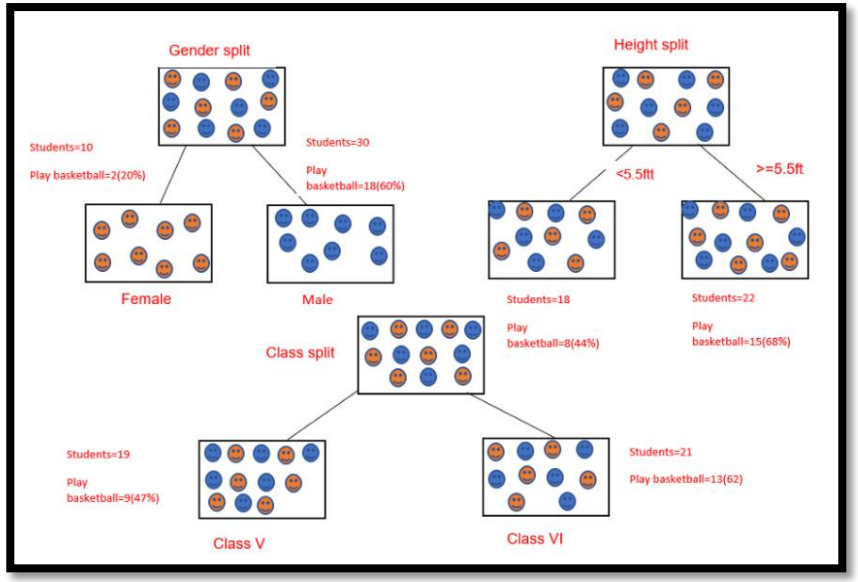
•Leftover data points after the split are less than the present data points.

Advantages:

•It takes the decision which imitates human's decision-making possibilities that makes it easily understandable.
•It requires less preprocessing of data.
•It is helpful in decision-making problems.

Disadvantages:

•It is prone to overfit which can be reduced by using a random forest algorithm or pruning technique.
•It is complex as it involves several layers.



I hope now we have an overview on the topic:

- Decision Tree

Thank you I hope you have enjoyed!!!