6.0 Decision Trees

1. A Decision Tree is a branching flow diagram or tree chart which helps in making decisions based on previous experience.
2. It uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.
3. Decision trees are effective techniques, particularly for classification problems and also regression tasks.
4. It comprises of the following:

a. A target variable and its initial distribution.

b. Root node:

This is the node that begins the splitting process by finding the variable that best splits the target variable.

c. Node purity:

Decision nodes are typically impure, or a mixture of both classes of the target variable.

Pure nodes are those that have one class — hence the term ‘***pure*** ’ .

d. Decision nodes:

The nodes we get after splitting the root nodes are called Decision Node
They are subsequent or intermediate nodes, where the target variable is again split further

by other variables.

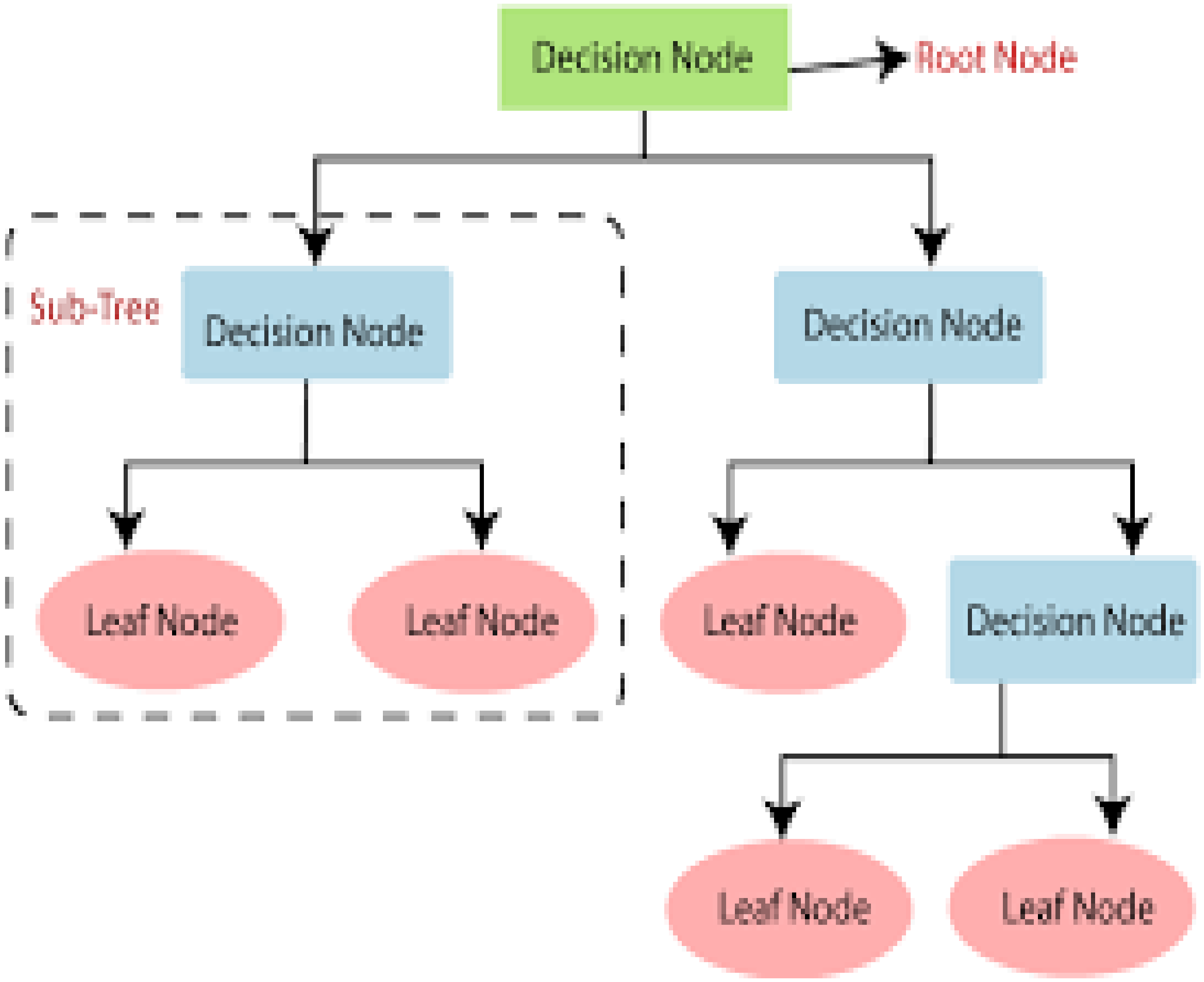
e. Leaf nodes:
These are nodes where further splitting is not possible. Also referred to as terminal nodes.

f. Sub-tree / Branch :
This is a sub-section of the decision.

g. Parent and Child Node:
A node, which is divided into sub-nodes is called a <u>parent</u> node of sub-nodes whereas sub-nodes are the <u>child</u> of a parent node.

h. Pruning: This is the cutting down of some nodes to stop overfitting.

i. Splitting: It is a process of dividing a node into two or more sub-nodes.



5. A decision tree takes a statement or condition and then makes a decision on whether the condition holds or does not.

6. The <u>conditions</u> are shown along the <u>branches</u> ; and the <u>outcome</u> of the condition, as applied to the <u>target</u> variable, is shown on the <u>node</u>.

7. Arrows leading away from a node indicate a condition which is being applied to the node.

8. Arrows pointing to a node indicate a condition that is being satisfied.

<u>Note:</u>

This is the first level of the Decision Tree — understanding the flow of splitting the decision space into smaller spaces which ultimately become more and more homogenous in the target variable which ultimately leads to a prediction.

9. Decision Trees offer tremendous flexibility in that we can use both numeric and categorical variables for splitting the target data.

10. The key points for a data scientist to observe include:

    a) Flow of information through the Decision Tree.

    b) How does the decision tree select which variable to split on at decision nodes?

    c) How does it decide that the tree has enough branches and that it should stop splitting?

11. Types of Decision Trees

Based on the type of target variable and can be of two types viz:

I. **Categorical Variable Decision Tree:** Decision Tree which has a categorical target variable.

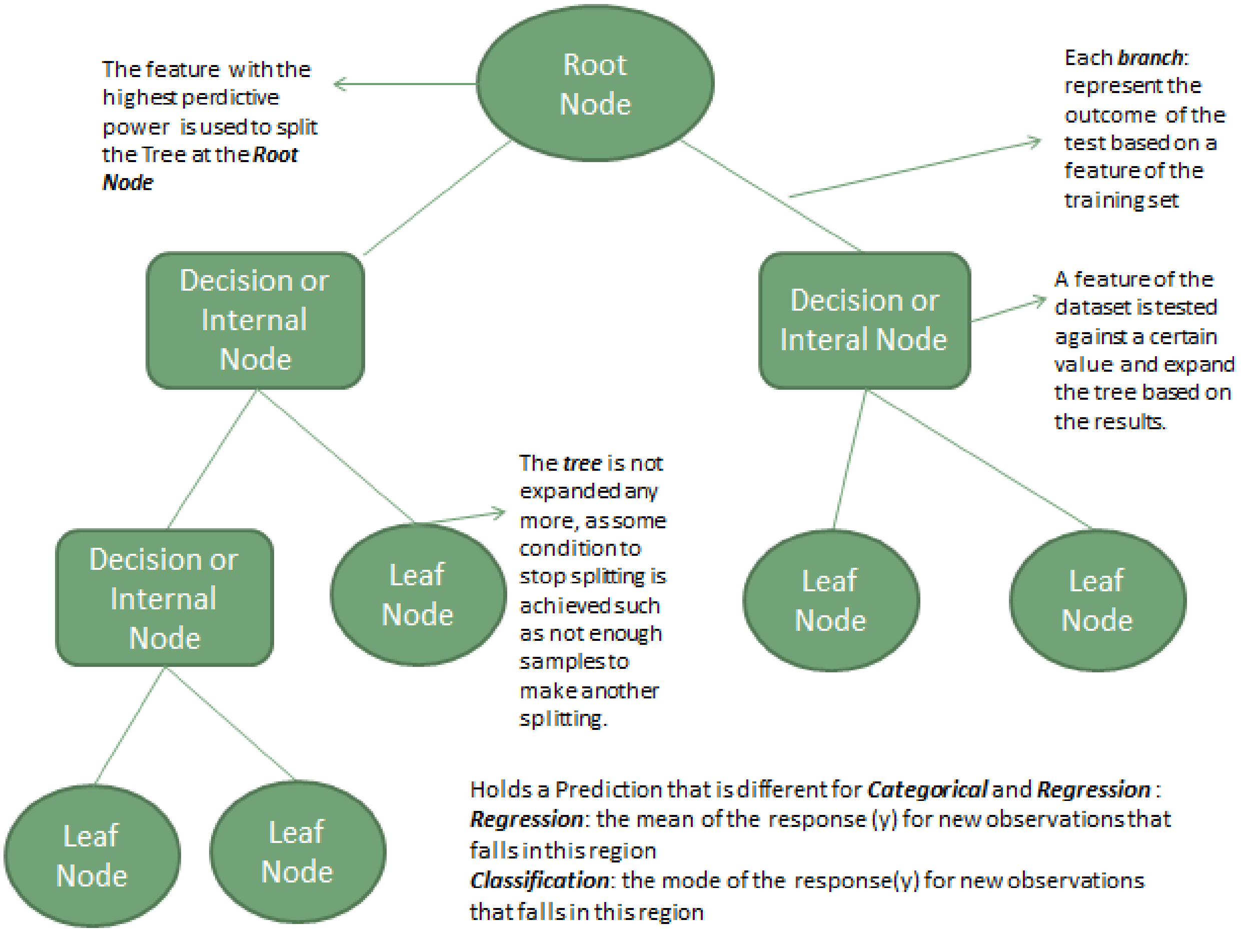
II. Continuous Variable Decision Tree: Decision Tree has a continuous target variable.

12. Overfitting:
This is a condition where the model learns the data too well and hence performs well on training dataset but fails to perform on testing dataset.

13. Underfitting:
This is a condition where the model is too simple for it to learn the dataset effectively.

14. Entropy:
This is a measure of disorder or impurity in a node.

15. Gini:
The Gini Index or Impurity measures the probability for a random instance being misclassified when chosen randomly. The lower the Gini Index, the lower the likelihood of misclassification.



The feature with the highest perdictive power is used to split the Tree at the ***Root Node***

Each ***branch***: represent the outcome of the test based on a feature of the training set

A feature of the dataset is tested against a certain value and expand the tree based on the results.

The ***tree*** is not expanded any more, as some condition to stop splitting is achieved such as not enough samples to make another splitting.

Holds a Prediction that is different for ***Categorical*** and ***Regression*** :
Regression: the mean of the response (y) for new observations that falls in this region
Classification: the mode of the response(y) for new observations that falls in this region