**Decision Trees**

* Belongs to the family of supervised learning algorithm.
* Goal is to create a training model that can used to predict the class or value of the target.

**Important Terminology related to Decision Trees**

1. **Root Node:**It represents the 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 the decision node.
4. **Leaf / Terminal Node:**Nodes do not split is called Leaf or Terminal node.
5. **Pruning:**When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
6. **Branch / Sub-Tree:**A subsection of the entire tree is called branch or sub-tree.
7. **Parent and Child Node:**A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.



Assumptions while creating Decision Tree

* In the beginning, the whole training set is considered as the **root.**
* Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
* Records are **distributed recursively** on the basis of attribute values.
* Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

Decision Trees follow **Sum of Product (SOP) r**epresentation. For a class, every branch from the root of the tree to a leaf node having the same class is conjunction (product) of values, different branches ending in that class form a disjunction (sum).

**Multiple Regression**

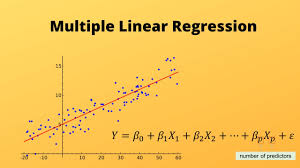
Multiple linear regression is a statistical method of predicting or explaining a continuous variable (sometimes called the dependent variable) as a linear combination of one or more variables (sometimes called independent variables).

The model is of the form

Y = ∑ bi xi+e

It makes several assumptions:

* Independent errors (e)
* Linearity - that is, that the model is linear in its parameters (the B)
* Additivity of terms (although you can add interaction terms)
* Errors have mean 0 and constant variance across all levels of x
* normality of errors (necessary only for some aspects of how the model is used)



Assumptions in Multi Regression

* Regression residuals must be normally distributed.
* A linear relationship is assumed between the dependent variable and the independent variables.
* Absence of multicollinearity is assumed in the model, meaning that the independent variables are not too highly correlated.

Why Multiple Regression –

* This model is directly interpretable. It’s easy to get ideas from a regression model about how to change strategies.
* Regression model is easily explainable.
* Regression models are computationally *cheap* and extremely *fast*. Especially when you're running data through a model millions (or more) times a day, these differences in speed add up to real dollars and cents.