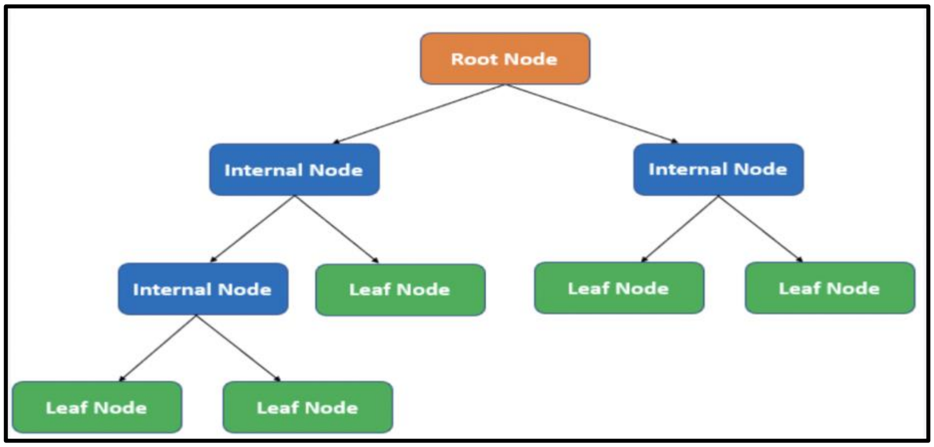
Decision Tree

- It is a tree like model used to make predictions.
- We begin with the root node where the split is done and when we cannot split on any node that node is called the leaf node.
- Intermediate nodes are called internal node.
- **Before fitting the decision tree we need to perform feature engineering i.e., for categorical variables need to do one hot encoding and for continuous variables we need to normalize them.**



- The difference between decision tree classification and decision tree regression is that in regression, each leaf represents the average of all the values as the prediction as opposed to a class label in classification trees.

- For classification problems, the prediction is assigned to a leaf node using majority voting but for regression, it is done by taking the average value. This average is calculated using the following formula:

Splitting and Homogeneity

- Split will try to achieve homogenous subsets belonging to one class.
- For **classification**, a data set is completely homogenous if it has only single class label.
- For regression, a data set is completely homogenous if its variance is as small as possible.
- If an attribute is nominal categorical (2 or more categories) then we have $2^{(k-1)} - 1$ possible splits. where k is the no. of classes.
- If an attribute is ordinal categorical (where there are natural, ordered categories) then we have n-1 different possible splits.

Methods that quantify homogeneity

1. **Classification error**: If everything is assigned to majority class then how much error i make is called classification error.
   Formula: 1-max(p_i) where p_i is the probability of class

2. **Gini Index**: It is the degree of randomly selecting a data point classified incorrectly.
   - **When GI is 0 then there is complete homogeneity**
   - **Formula:**

   $$G = \sum_{i=1}^{k} p_i(1 - p_i)$$

   - k is the number of classes

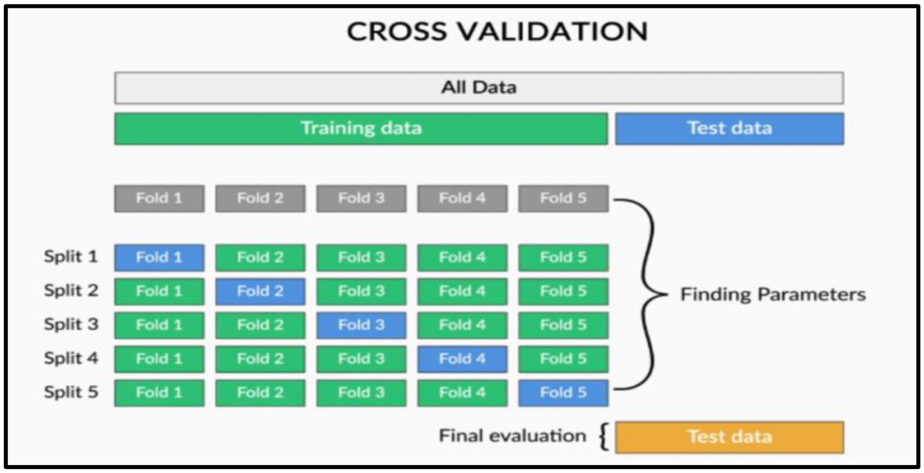
3. **Entropy:** It is the degree of disorder present in the data.
   - **When entropy is 0 then there is complete homogeneity**
   - **Formula:**

   - $D = -\sum_{i=1}^{k} p_i . log_2(p_i)$.

K-Fold Cross Validation

All Data = Training + Testing

- Will divide my training data into no. of fold (i.e. K=3/4/5).
- In 1st split, my 1st block becomes the target and the other 4 become the train.
- 2nd time 2nd block will be my target and the rest will be train.
- This will continue till the no. of splits specified.
- We calculate the accuracy on each of the target blocks(Fold1/Fold2..) We take their average and we will get to know how our hyperparameters have performed on the unseen data.
- Once this is complete, we evaluate or model against the test data.

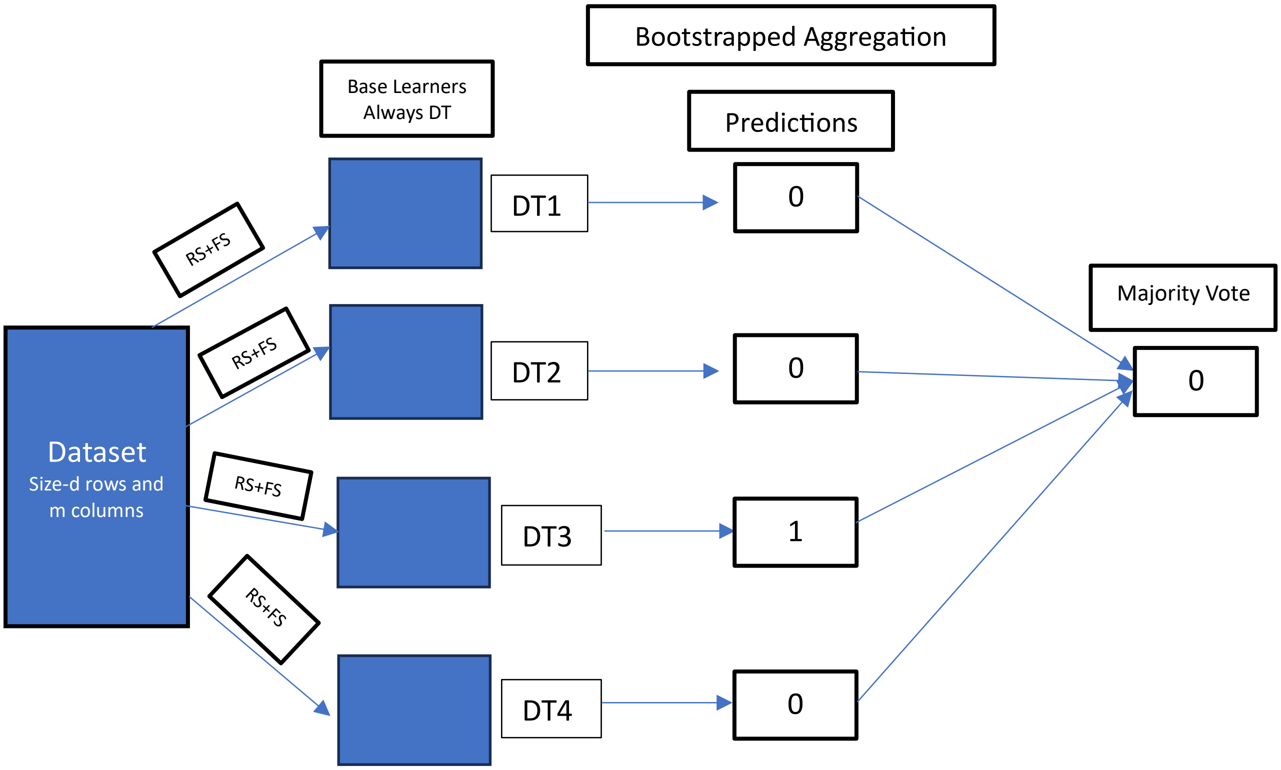


Ensembles

- It is a collection of models used to make predictions.
- **Random Forest** is an ensemble made by the combination of large number of decision tree.
- Random forest use a technique known as **bagging** which is an ensemble.

Bagging method

- **Goal**: To reduce the variance of the algorithm



- **Row Sampling+ Feature Sampling** (RS+FS) with replacement certain rows and columns will be passed to the base learner from our data set and some records could get repeated in other base learners as well.
- Models M1, M2…M4 can be only DT.
- If it is a classification problem then we do a majority vote for regression we take the average.
- We use random forest because DT are prune to over fitting if hyperparameter tuning is not done i.e., means model has performed well on the training data but not so well on the test data. **This means the model has low bias (high training acc.) and low variance (low test accuracy).** We get low variance when we use RF.

OOB (Out of Bag Evaluation)

- There are certain data points from the main dataset that never get used because they never get selected in the sample.
- It is never used in any decision tree.
- We can use that data for validation.
- In RF if we mark **OOB=True** then it takes those points as validation data.
- Train data will have 2/3rd of total data points and OOB will have 1/3rd of total data points.
- It will give us the accuracy of training data and validation data.

Advantage:

- It gives the feature importance for each variable.