

Decision Tree & Random Forest Assignment

### **Problem Statement:**

Sam is done with the regression algorithms. Now, he decides to build some classification algorithms on top of the dataset

Sam starts off by building Decision Tree algorithm on top of the data

## Questions on decision tree:

1. Building a decision tree model:
   1. Start off by dividing the ‘customer\_churn’ data into train & test sets in 70:30 ratio. The split-criteria would be determined by the ‘Dependents’ column
   2. Build a decision tree model on top of the ‘train’ set, where the dependent variable is ‘Dependents’ & the independent variable is ‘Partner’. Store the result in ‘mod\_tree1’
   3. Plot the tree and add text
   4. Predict the values on the test set and store the result in ‘result\_tree1’
   5. Build a confusion matrix for the actual values & the predicted values
   6. Calculate the accuracy from the confusion matrix
2. Building 2nd decision tree model on same ‘train’ & ‘test’ sets:
   1. In this case the dependent variable is ‘Dependents’ & the independent variables are ‘Partner’ & ‘InternetService’. Store the result in ‘mod\_tree2’
   2. Plot the tree & add text
   3. Predict the values on the test set & store the result in ‘result\_tree2’
   4. Build a confusion matrix for the actual values & the predicted values
   5. Calculate the accuracy from the confusion matrix

After building a single decision tree model, Sam decides to implement ensemble of trees. So, he goes ahead & applies the random forest algorithm on the data

## Questions on random forest:

1. Building the first “Random Forest” model:
   1. Start off by dividing the ‘customer\_churn’ data into train & test sets in 65:35 ratio. The split-criteria would be determined by the ‘gender’ column
   2. Build a random forest model on top of the ‘train’ set, where the dependent variable is ‘gender’ & the independent variables are ‘MonthlyCharges’ & ‘tenure’. The number of decision trees in the random forest would be 35. Store the result in ‘mod\_forest1’
   3. Find the importance of the independent variables and also plot it
   4. Predict the values on top of the test set & store the result in ‘result\_forest1’
   5. Build a confusion matrix for the actual values & the predicted values
   6. Find out the accuracy from the confusion matrix
2. Build a 2nd ‘Random forest’ model on the same train & test sets:
   1. The dependent & the independent variables would be same. The number of decision trees would be 350. Store the result in ‘mod\_forest2’
   2. Find the importance of the independent variables & also plot it
   3. Predict the values on top of test set & store the result in ‘result\_forest2’
   4. Build a confusion matrix for the actual values & predicted values
   5. Find out the accuracy from the confusion matrix