# ASSIGNMENT-3

# Q1- PART A

# Q1. Part A(Dataset 1)

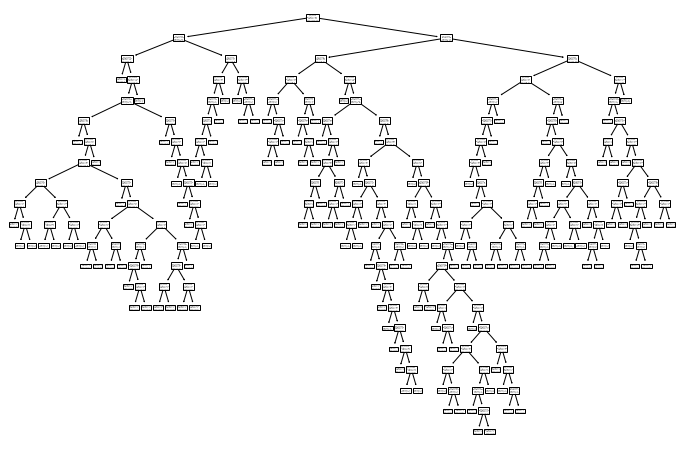
1. **Decision Tree Construction and Visualization:**

After dropping the samples with any missing attribute, a decision tree was trained with t training accuracy, validation accuracy, and test accuracy:



The really low accuracy of training set on the test set indicates overfitting in the model.

Tree visualization:

****

we can see that the tree is not an ideal tree as it has a really long depth.

1. **Decision Tree Grid Search:**

The optimal parameters obtained after the grid search:

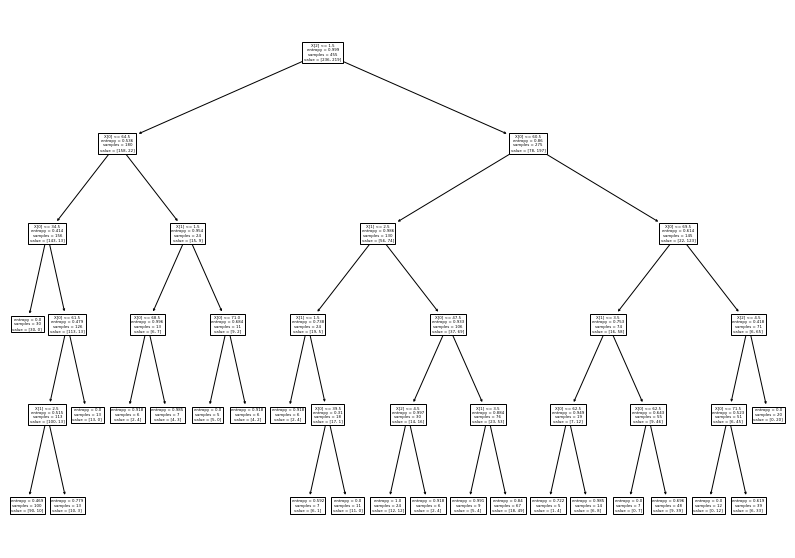
****

**The** training accuracy, validation accuracy, and test accuracy obtained after training the model using these parameters:

****

The training accuracy has reduced whereas the validation and test accuracy have improved. The comparatively low accuracy of training set on the test set still indicates overfitting in the model.

**Tree visualization:**

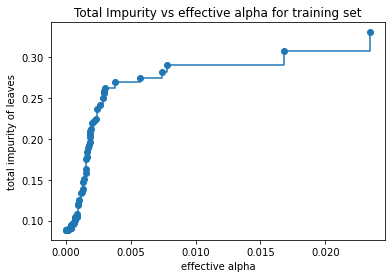
****

After using the right parameters we are able to achieve a tree still the max depth is 5 which can be further reduced and there is overfitting in the model.

1. Decision Tree Post Pruning (Cost Complexity Pruning):

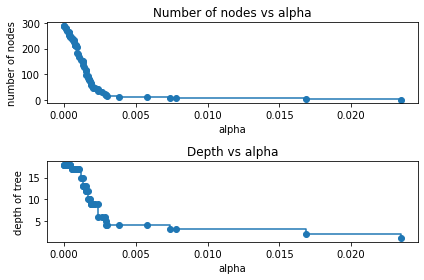
I chose the optimal ccp\_alpha: 0.005

* plot total impurity of leaves vs effective alphas of pruned tree:



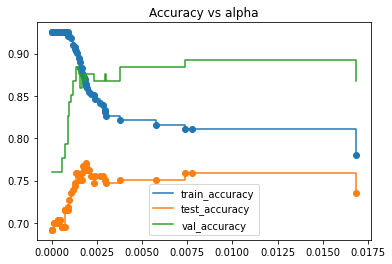
As the alpha increases the total impurity of the leaves increases too.

* Plot the number of nodes vs alpha and the depth of tree vs alpha



As alpha increase the number of nodes decrease and as alpha increase the depth of the tree also reduces.

* Plot training accuracy, validation accuracy, and test accuracy vs alpha



We can choose alpha = 0.005

In [16]:

What are your observations?

As the alpha increases the number of nodes and depth of the tree decreases.

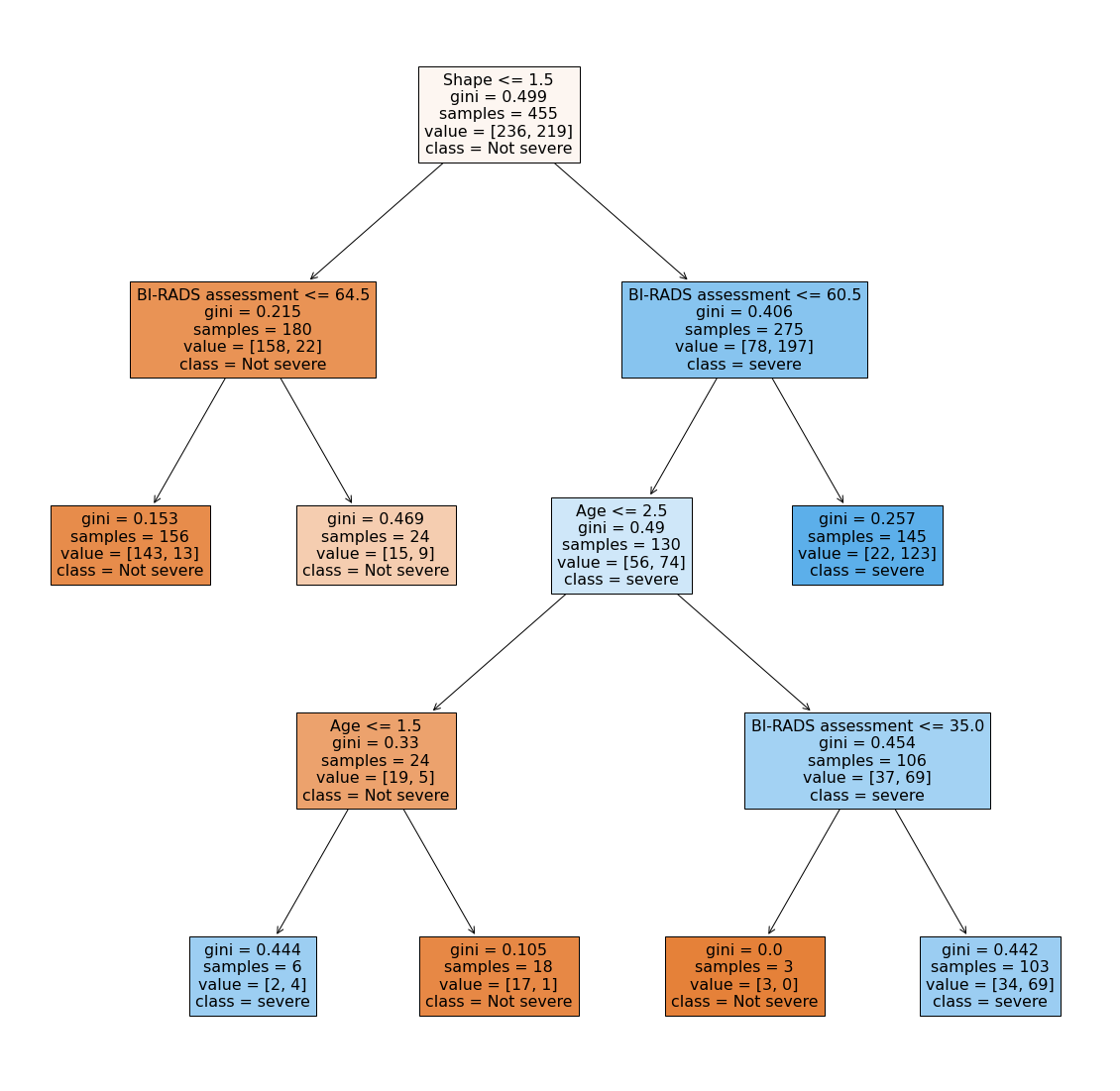
The low amount of test accuracy indicates that there are meaningful differences between the kind of data you trained the model on and the testing data you're providing for evaluation. The model after post pruning performs better on validation set thus giving a better accuracy.

* Train, validation, and test accuracy with respect to the best tree:



The model runs very well on the validation set.

Visualize the best pruned tree.



We can see that the size of decision tree significantly got reduced. When compared to the trees in part a and b, the post pruned tree gives better accuracy results and the overfitting in model has also been reduced.

1. **Random Forest:**

Different trees constructed by changing parameters:

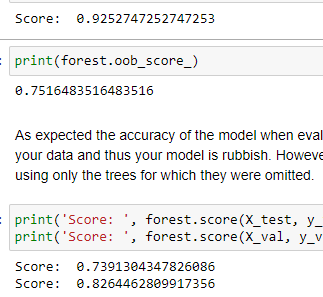
1. n\_estimators = 100

train accuracy: 92.52

test: 73.91

validation: 82.64

OOB score: 75.16



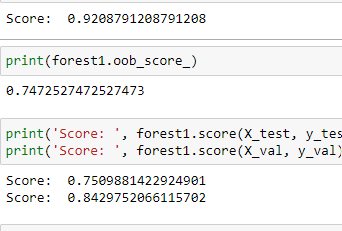
ii) n\_estimators = 100, min\_samples\_split=3

train accuracy: 92.08

test: 75.09

validation: 84.29

OOB score: 74.72



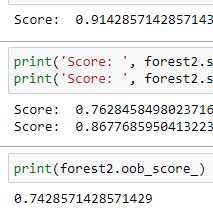
iii) n\_estimators = 150, min\_samples\_split=5

train accuracy: 91.42

test: 76.28

validation: 86.77

OOB score: 74.28



We observe as the number of n\_estimator increases the training accuracy reduces but the model runs well on the validation set, improving validation accuracy.

* Grid search over parameters to find the best parameters:



* Train, validation, and test accuracy achieved with respect to the best forest:



The high training, test and validation accuracy states that random forest constructed using the best parameters gives the best result, better than post pruned trees.

* OOB SCORE:



Conclusion: the random forest model trained using the best parameters gives better results than post pruned trees. With default parameters the training accuracy comes out good but the test and validation accuracy are low. After using the correct set of parameters the training and validation set accuracy have much improved.

1. **Missing data imputation**

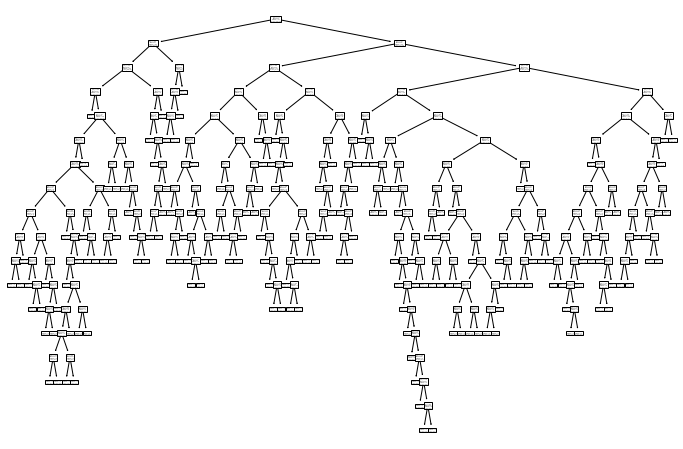
Rather than dropping the rows with NAN values I will replace these values with median and mode and analyse which metric gives better results.

1. **Median:**

* Accuracies of Tree with basic parameters:



* Tree visualization with basic parameters:



* Grid search:

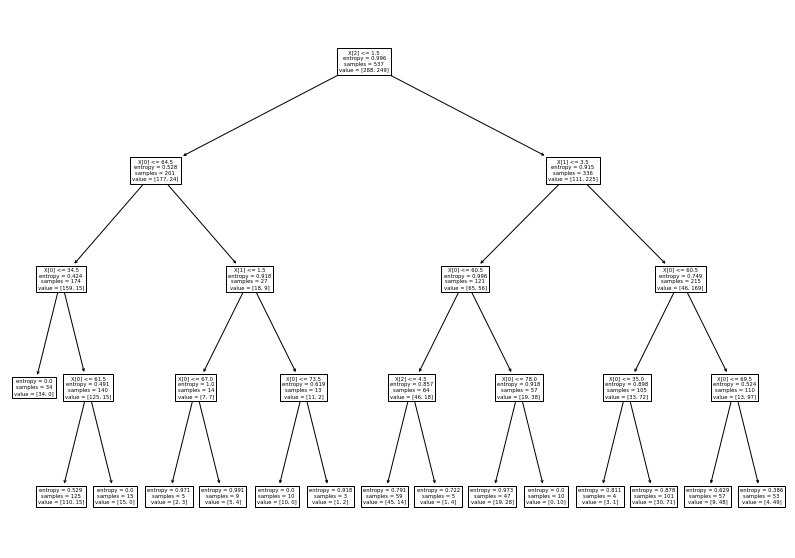


* Accuracies achieved after grid search:



The training accuracy has been reduced but the model runs really well on validation set thus improving the validation accuracy.

* Tree visualization after grid search



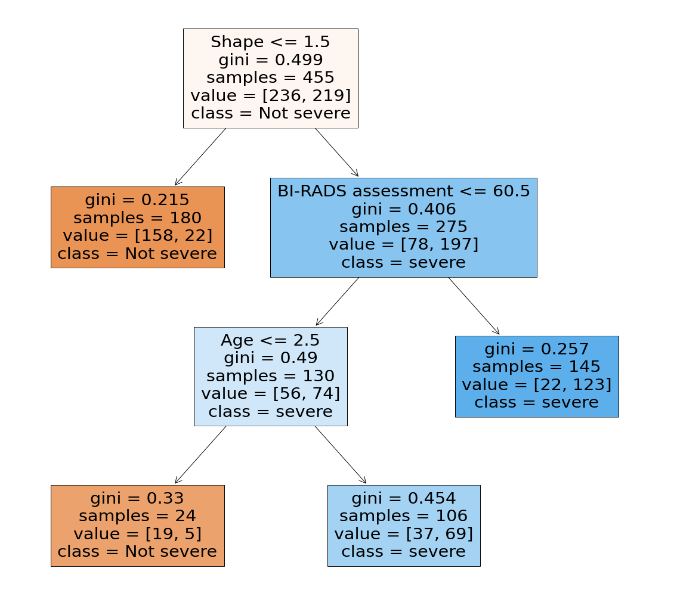
The size of the tree has significantly reduced.

* Post Pruning:

The training and the test accuracy has reduced but the validation accuracy has improved stating that the model word really well on the validation data.

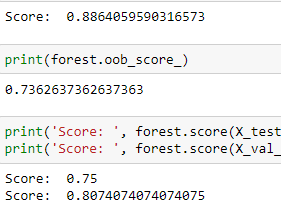


* Tree visualization :



The size of the tree is reduced to depth= 3 giving us the best tree till now.

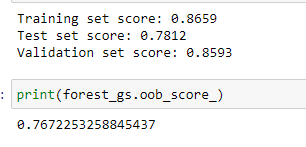
* Random Forest:



The random forest gives good training accuracy but not a very good test and validation accuracy. So lets evaluate the best parameters:



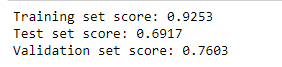
* Accuracies after gridsearch:



Here the results given by random forest with best parameters and post-pruned trees are almost the same with minimal difference. But the training and test set accuracies are far better in random forest grid search as compared to post pruned trees.

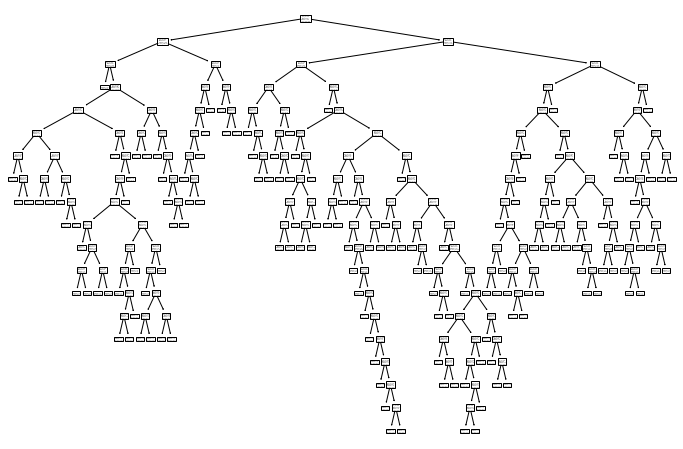
1. **Mode:**

* Accuracy of tree with default parameters:



The train accuracy is really good but the test set and validation accuracy are really poor.

* Tree visualization:



* Best parameters after grid search

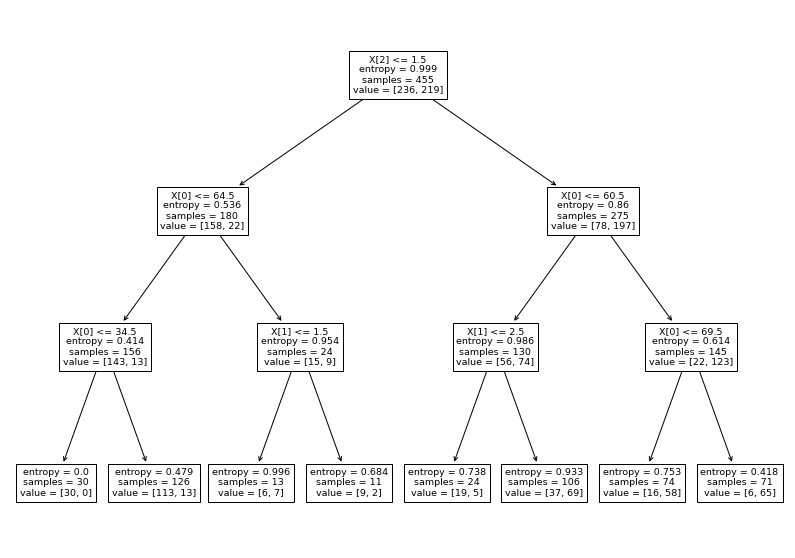


* Accuracy obtained by using the best parameters:



With the best parameters the model runs really well on the validation set giving 87% accuracy which is good.

* Tree visualization:



The size of three is also reduced to a lot of extent, better than the one achieved with median.

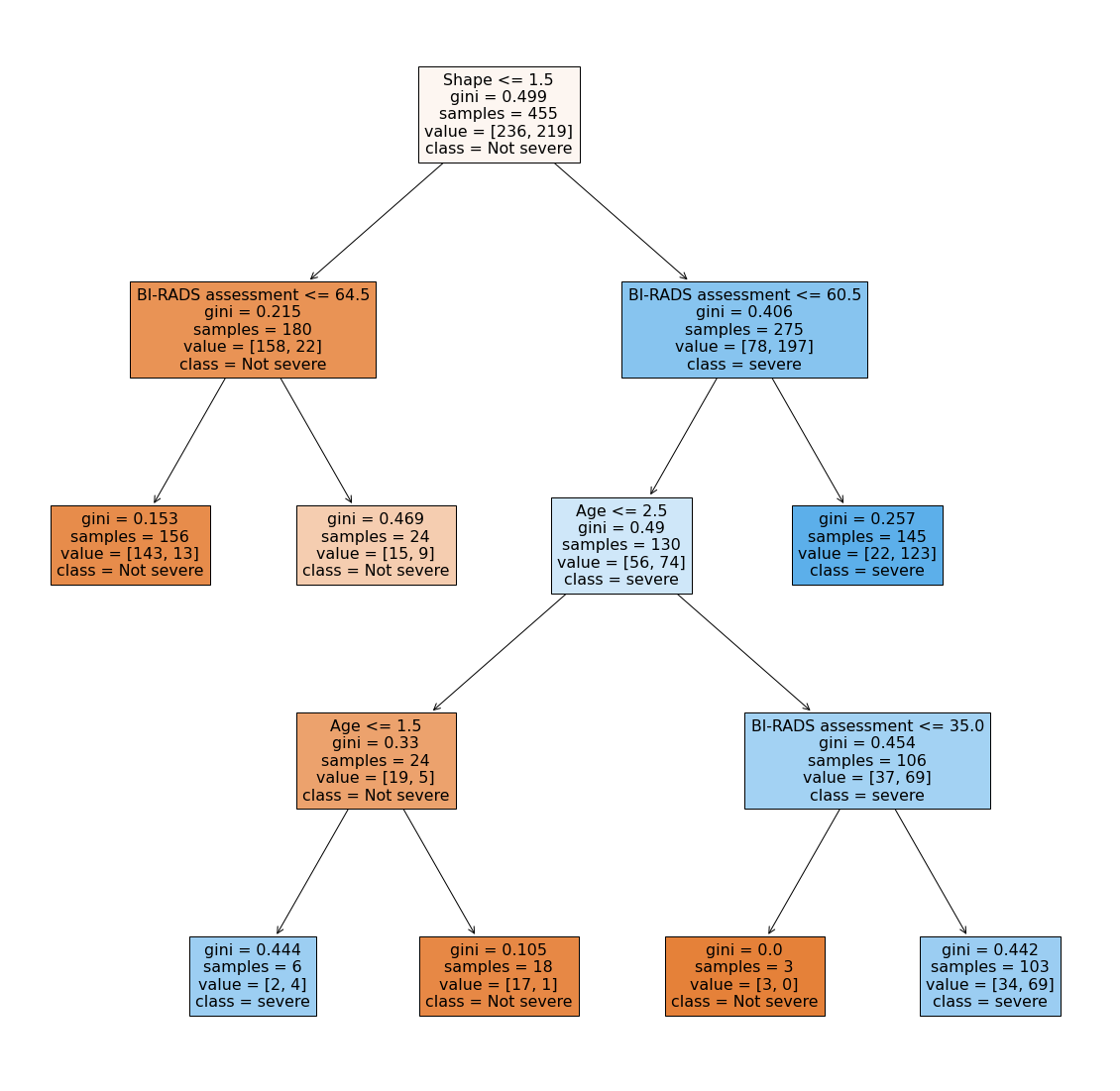
* Post pruning:



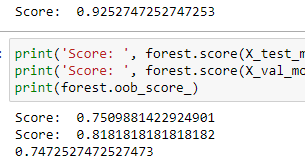
Gives poor accuracy results. The ccp\_alphas = 0.10

* Tree visualization:

The tree obtained is good and reduced in size but the performace of this model is not really good, will not give very accurate results.



* Random forest:

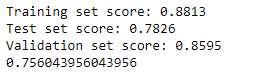


The train accuracy is good but the test and validation accuracy are comparatively low.

* Best parameters obtained after grid search:



* Accuracy obtained :



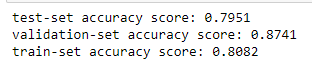
The random forest with optimal parameters gives really good results better than decision tree or post pruned trees.

Conclusion: the medians imputed values give better results as compared to mode imputed tree thus median is a better metric for treating null values.

1. **Gradient boosted trees:**



As the gradient boosted tree deal with missing values on their own and gives a model with less bias, and better predictive performance.



We can see that the accuracy on the test and validation set are coming out to be really good and the low difference between test and train accuracies means that the overfitting in the model has been dealt with giving better results.