COL774 Assignment 3

Aniruddha Deb
2020CS10869

October 2022

1 Decision Trees, Random Forests, Gradient Boosted Trees

1.1 Dataset 1

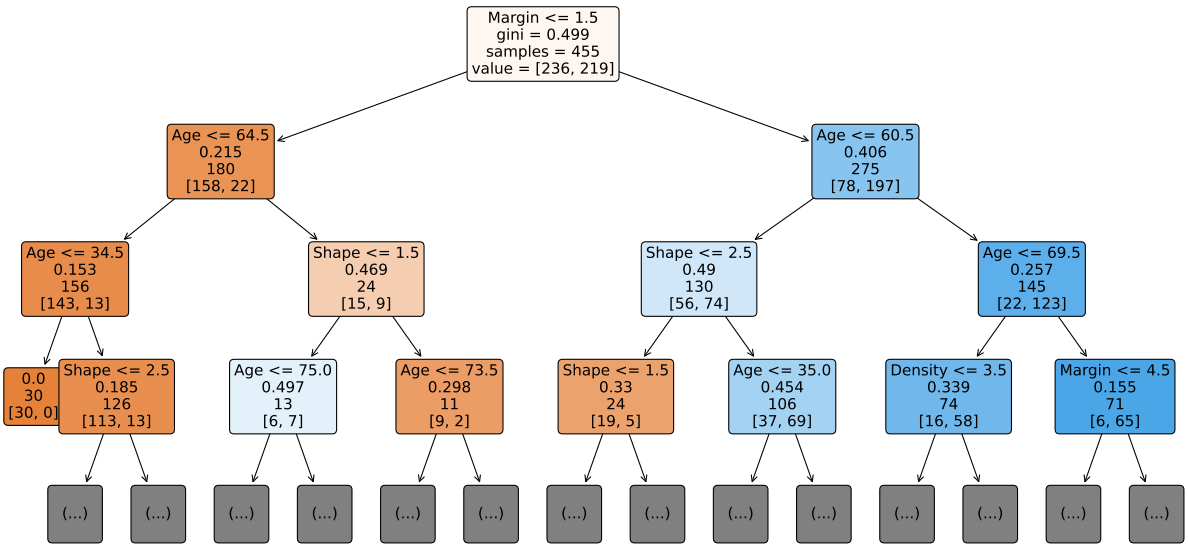
(a) After removing the missing values, we obtain the following accuracies:

(i) **Training set:** 92.53 %

(ii) **Validation set:** 76.03 %

(iii) **Test set:** 69.17 %

The decision tree we obtain is as follows (drawn upto a depth of 3):



(b) With GridSearchCV and the grid of parameters max_depth $= \{4, 6, 8, 10\}$, min_samples_split $= \{2, 3, 4, 5\}$ and min_samples_leaf $= \{1, 2, 3, 4, 5\}$, the best parameters (obtained via five-fold cross validation) are:

(i) max_depth $= 4$

(ii) min_samples_leaf $= 5$

(iii) min_samples_split $= 2$

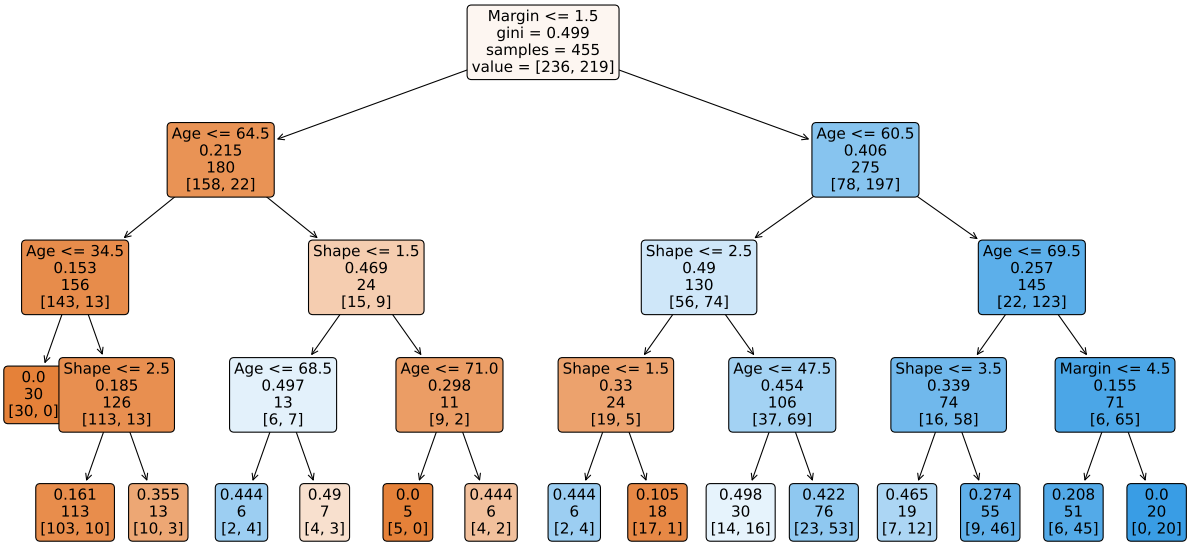
With them, we obtain the following accuracies:

(i) **Training set:** 81.98 %

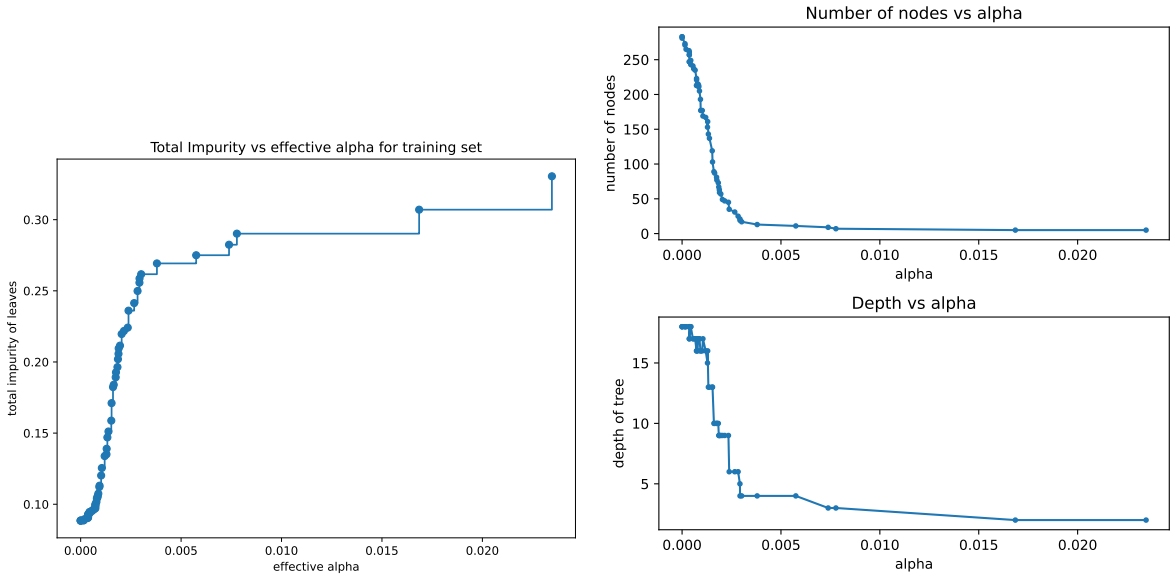
(ii) **Validation set:** 87.6 %

(iii) **Test set:** 75.1 %

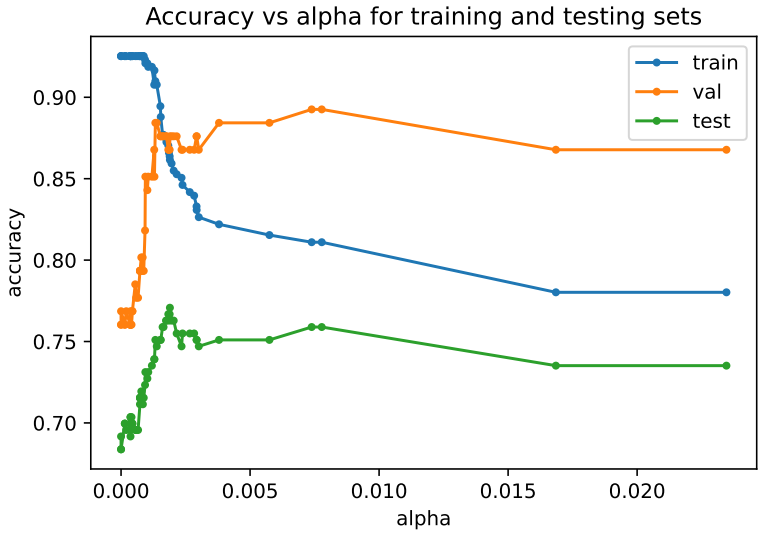
The decision tree has a much shallower depth than the previous tree, and is also simpler



(c) The total impurity vs the alphas is plotted below, along with the depth and the number of nodes v/s alpha



We see that the training accuracy is very high for low values of alpha, at the expense of the validation and test accuracy i.e. the model overfits for low values of alpha. The best trees are the ones with fewer nodes



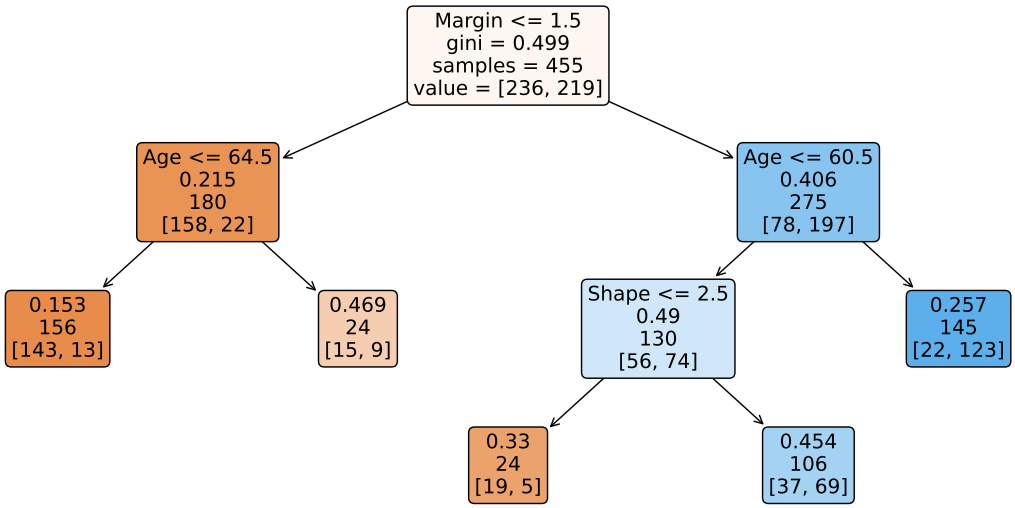
The accuracies we obtain for the various datasets, using the best pruned tree are:

(a) **Training set:** 81.1 %

(b) **Validation set:** 89.26 %

(c) **Test set:** 75.89 %

The best pruned tree, as discussed before, has only 9 nodes compared to the trees we obtained in part (a) and part (b).



(d) With GridSearchCV and the grid of parameters n_estimators $= \{50, 100, 150, 200\}$, max_features $= \{1, 2, 3, 4\}$ and min_samples_split $= \{2, 3, 4, 5\}$, the best parameters (obtained using out of bag score as the metric) are:

(i) n_estimators $= 200$

(ii) min_samples_split $= 5$

(iii) max_features $= 3$

With them, we obtain the following accuracies:

(i) **Training set:** 90.11 %

(ii) **Out of Bag set:** 76.26 %

(iii) **Validation set:** 86.78 %

(iv) **Test set:** 76.68 %

(e) The results are summarized in the following table:

Default decision tree:	Median	Mode
Train	91.81 %	90.69 %
Validation	74.07 %	75.56 %
Test	74.31 %	71.18 %

Grid Searched Decision tree:	Median	Mode
max_depth	4	4
min_samples_split	2	2
min_samples_leaf	3	2
Train	81.56 %	81.38 %
Validation	87.41 %	86.67 %
Test	80.90 %	77.43 %

ccp_alpha optimized Decision tree:	Median	Mode
n_nodes	11	137
ccp_alpha	0.0055	0.0012
Train	80.26 %	88.83 %
Validation	87.41 %	88.15 %
Test	79.17 %	76.04 %

Random Forest classifier:	Median	Mode
n_estimators	200	100
min_samples_split	5	5
max_features	3	2
Train	88.45 %	88.08 %
Out of Bag	75.23 %	73.37 %
Validation	83.70 %	83.70 %
Test	77.78 %	77.43 %

What do I see? Lot of gruntwork without any point

(f) After implementing an XGBoost classifier and searching for parameters using GridSearchCV in the given parameter space, we obtain the following best parameters:

(i) max_depth = 10

(ii) n_estimators = 10

(iii) subsample = 0.5

With them, we obtain the following accuracies:

(i) **Training set:** 83.61 %

(ii) **Validation set:** 84.44 %

(iii) **Test set:** 77.08 %

1.2 Dataset 2

(a) blah