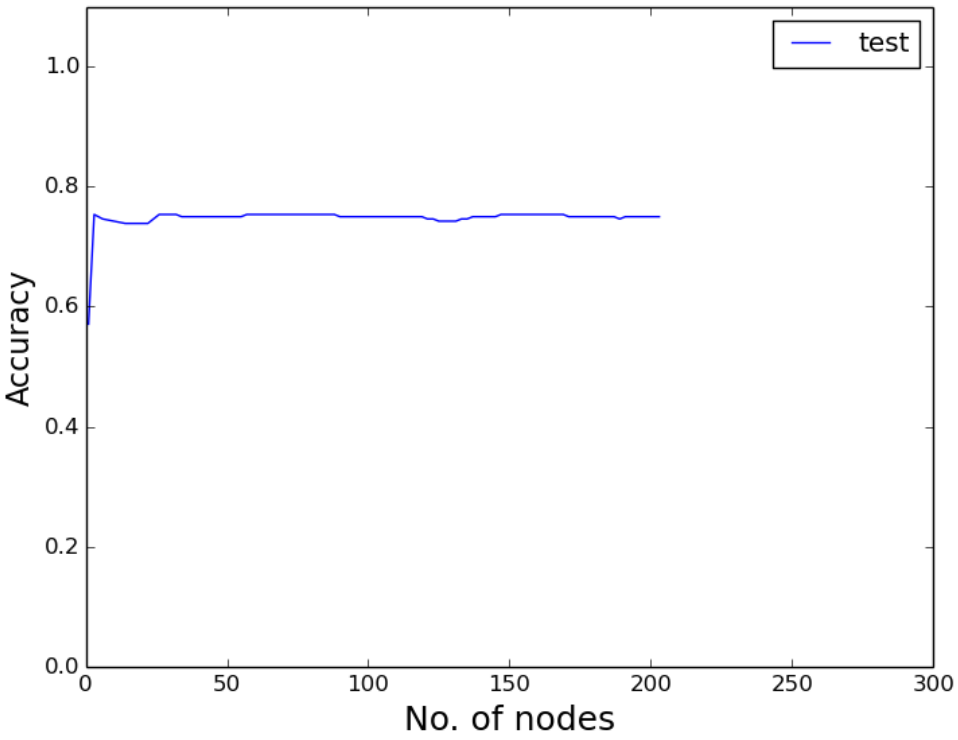
Machine Learning - COL774 Assignment 3

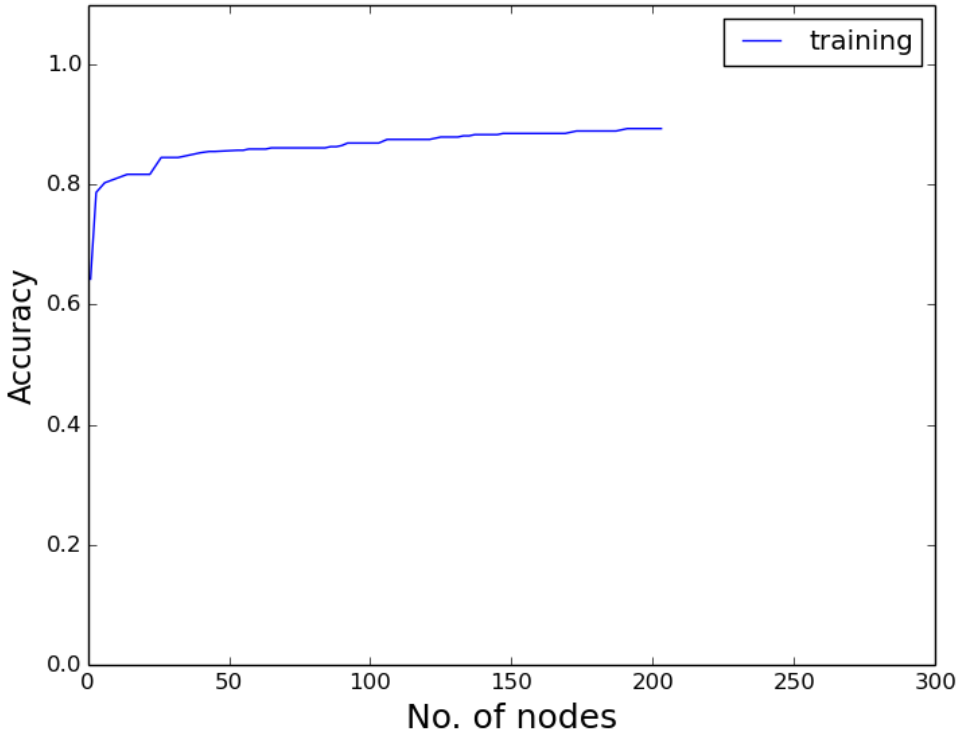
Tanuma Patra
2015MCS2356

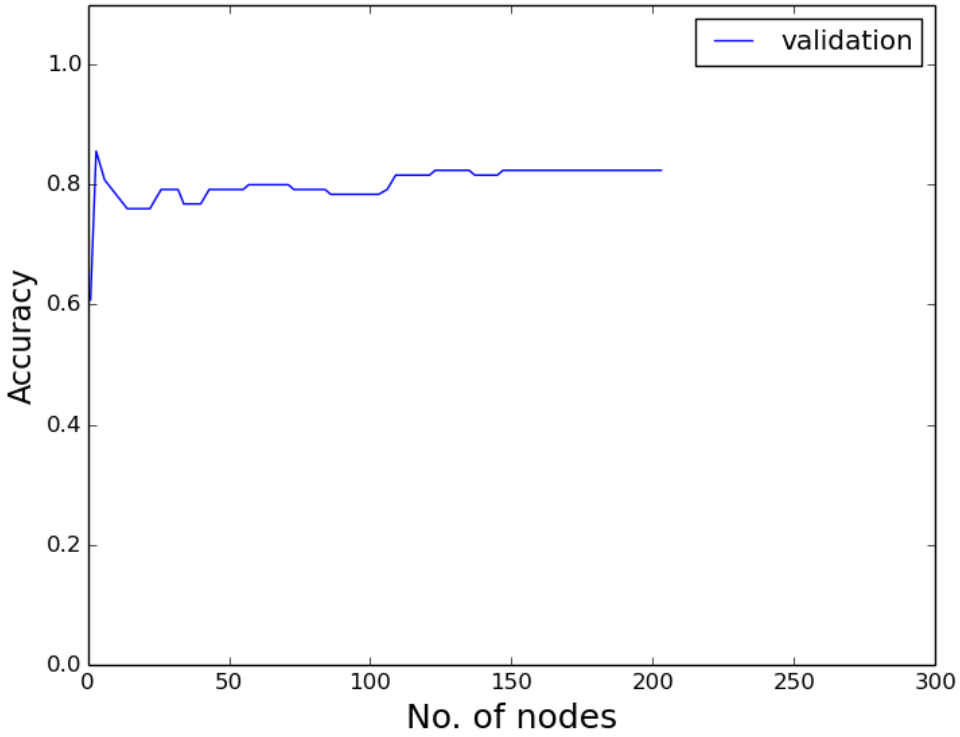
April 13, 2016

Question 1. Implementation of Decision Tree algorithm for the Kaggle Problem - Tree-Building with constant medians, Post Pruning implementation and Tree with dynamic median

Solution (. a) The Accuracies obtained are as follows:
Train Accuracy: 89.35 %
Test Accuracy : 75 %
Validation Accuracy: 82.4%
The plot for train test and validation accuracy is obtained as follows

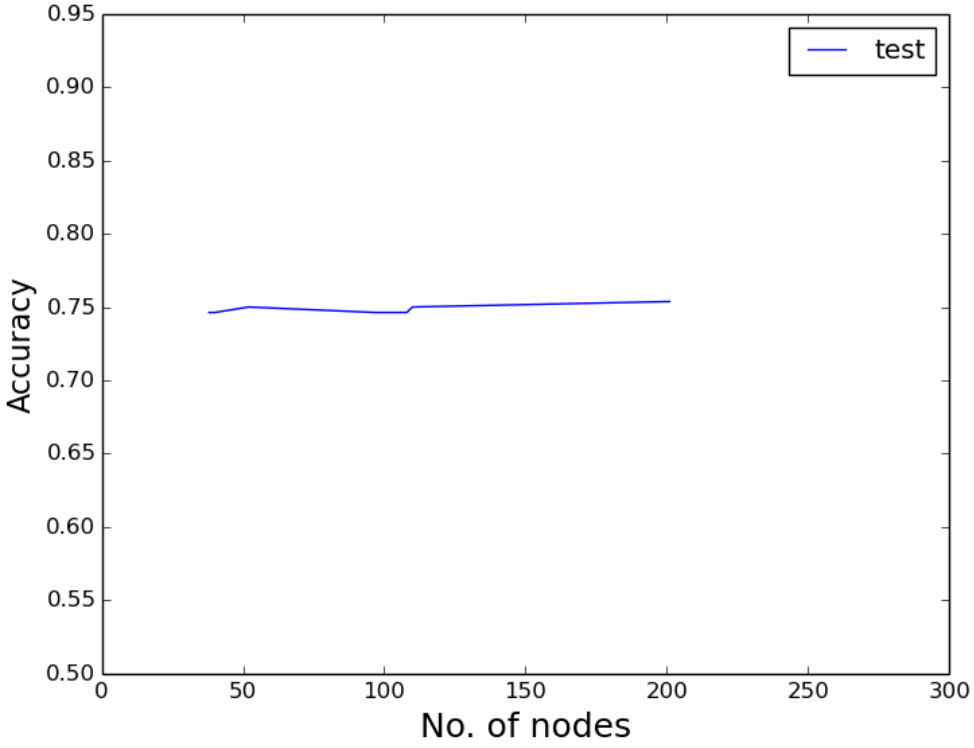


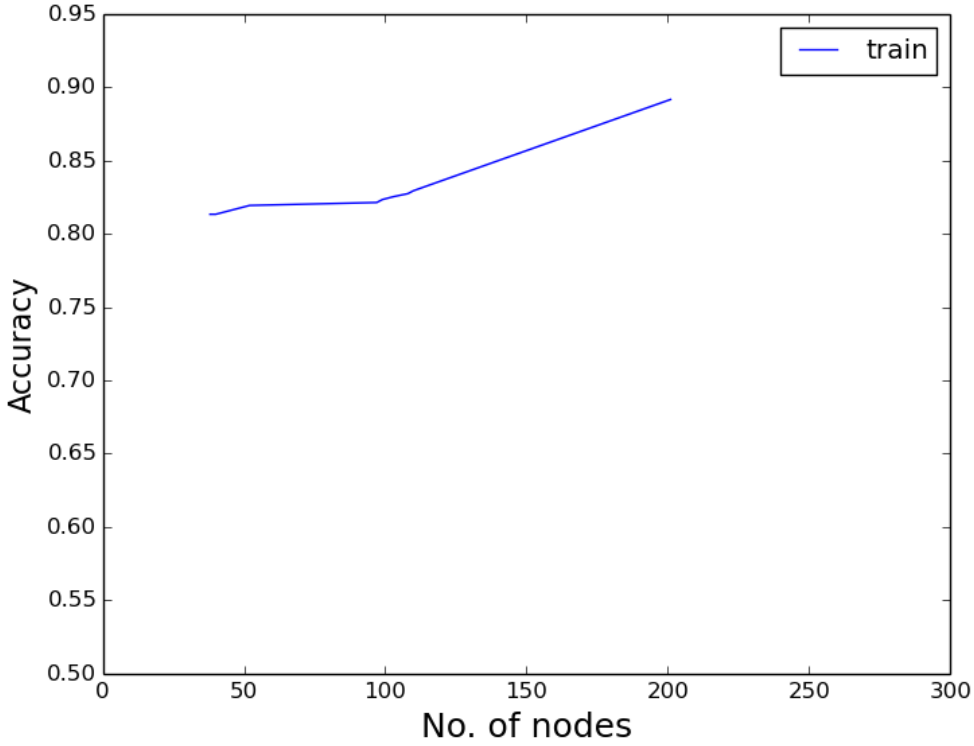

training
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
50
100
150
200
250
300
No. of nodes

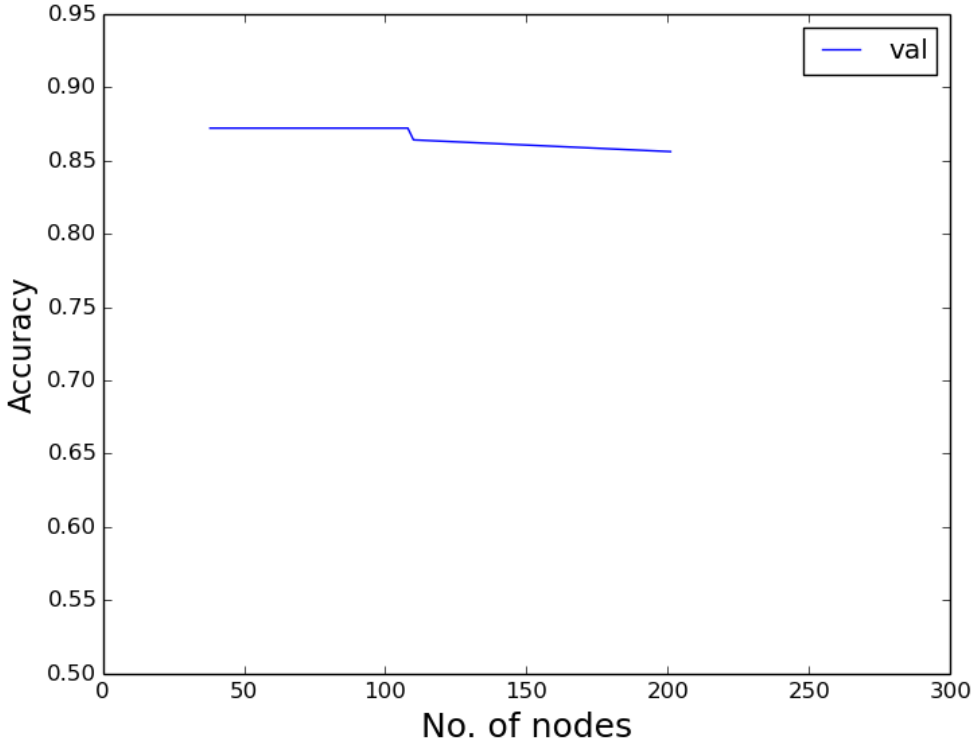


Observation : It is observed that the train set accuracy increases with the number of nodes. The Validation set accuracy hikes up and then suffer a huge decrease then increase gradually. The Test set accuracy increases till a certain number of nodes and then remains constant.

(b) Pruning is implemented and the train, test and validation accuracies are obtained as follows

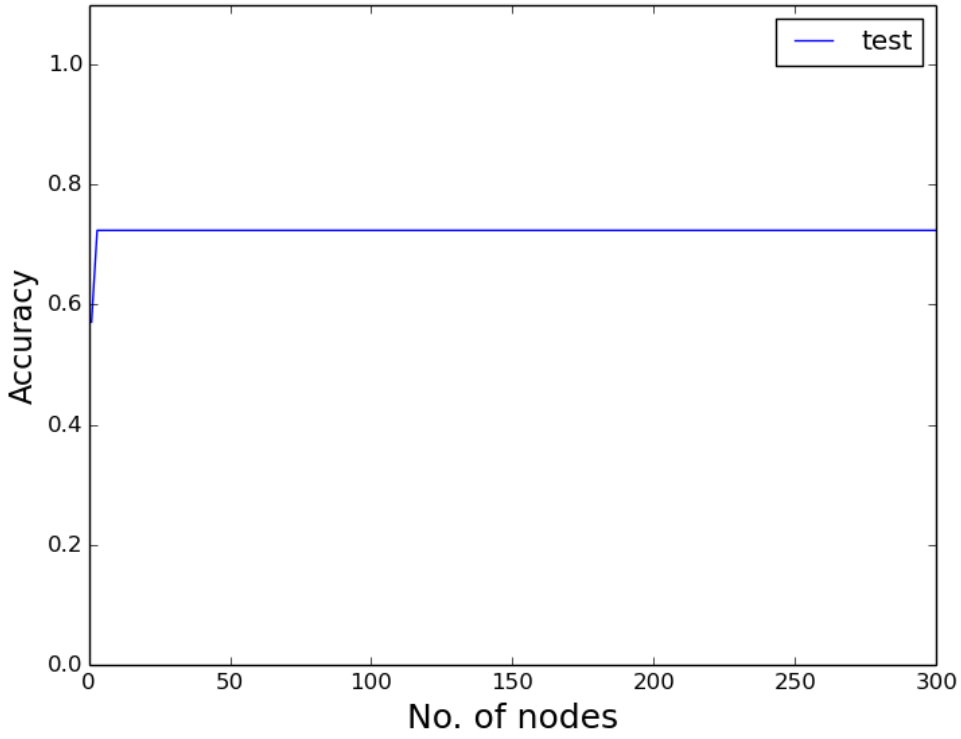

test
Accuracy
No. of nodes
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
50
100
150
200
250
300

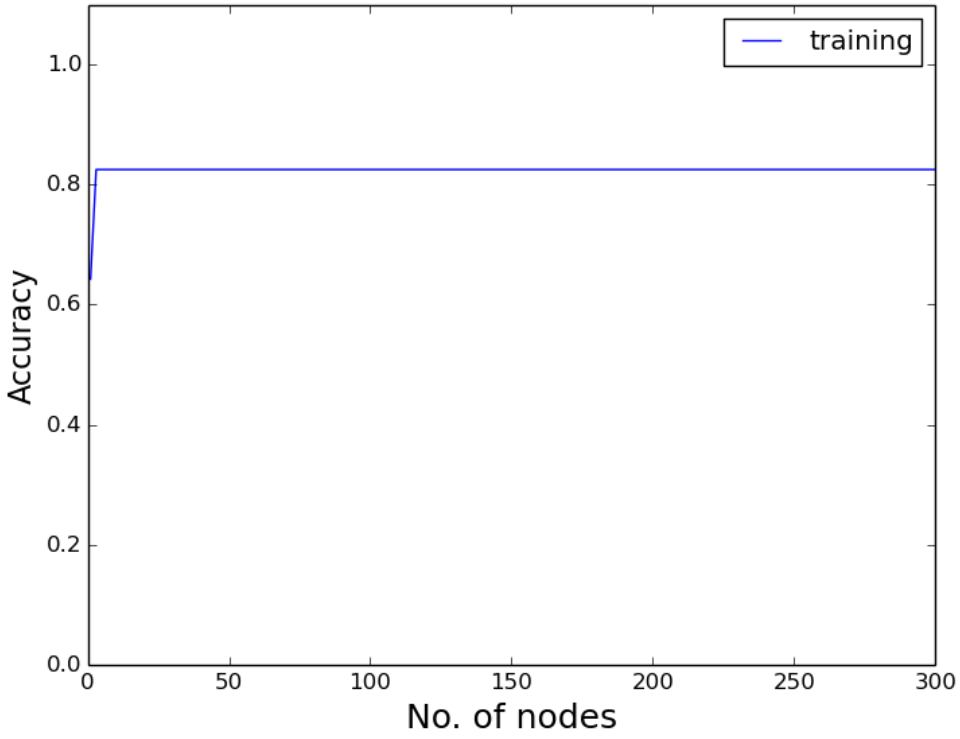

train
Accuracy
No. of nodes
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
50
100
150
200
250
300

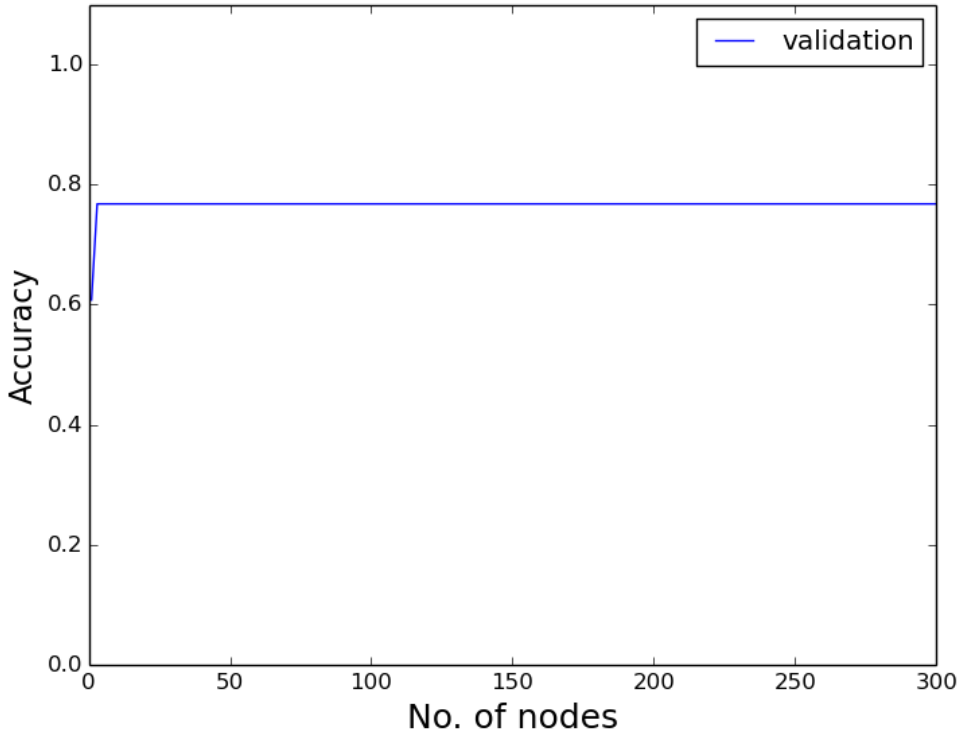


Findings: The validation accuracy on the original tree is less initially. It increases as we prune the tree. It reaches an optimum with respect to the validation accuracy.The train accuracy decreases on pruning whereas the test accuracy doesn't change much.

(c)The accuracies obtained for this case are:
Train Accuracy: 82.5301204819 %
Test Accuracy: 72.3880597015 %
Validation Accuracy: 76.8 %
The plots obtained are as follows:


test
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
50
100
150
200
250
300
No. of nodes


training
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
50
100
150
200
250
300
No. of nodes



The attributes that were taken multiple times to split on were the Age, Ticket and Fare. The exact counts are :
Age: **5 times**
Median values are: 25 35 29 27.5 26
Ticket: **4 times**
Median values are: 244335.5 30669.5 228665.5 224270
SibSp: **2 times**
Median values are: 1.0 0.0
Fare: **2 times**
Median values are: 8.05 16.1

Observations: 1 There is a substantial increase in the training accuracy. However the test and validation accuracies remain the same, rather deteriorate. So it can be concluded that this method just overfits the train data. It does not give good performance on the validation and test set.

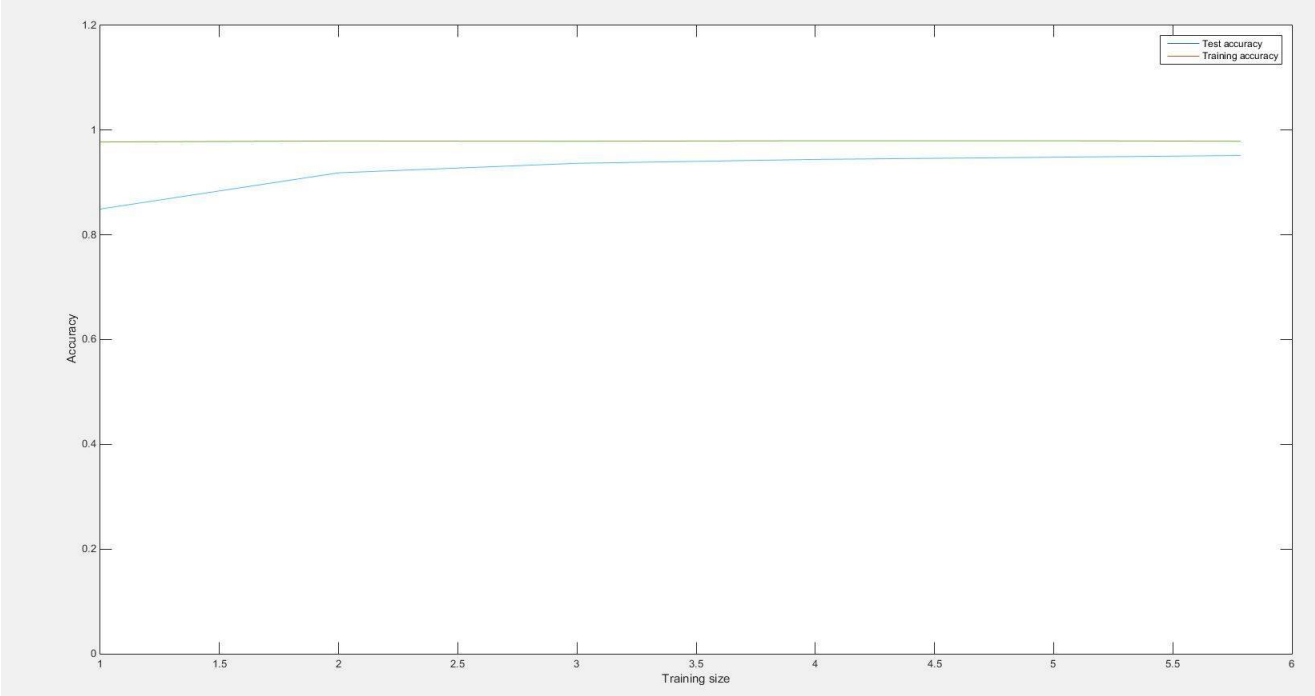
criteria = entropy and gini both work fine max_leaf_nodes = 10 max_depth

works best for depth = 1
min_samples_split = 2-8 work fine
min_samples_leaf = 4-8 will restrict the creation of pure class nodes to avoid an overfit
Test Accuracy = 75.37 % Validation Accuracy = 85.6 %

Question 2. Implementation of Naive Bayes Algorithm for text classification of Articles belonging to 8 Newsgroups

Solution (. a) The average test set accuracy for matlab implementation obtained is **95.186 %**
(b) On randomly guessing the category, the accuracy obtained was only **12.4 %**. The improvement in the prediction accuracy is by **82.71 %**
(c) If articles are posted on more than one newsgroups,the probability is uniformly distributed for all the categories in which they are posted. For the naive bayes classifier we are determining the class by relatively considering the probabilities. It will affect the algorithm if the amount of cross-posting is very high resulting in confusion amongst classes. So the classifier will not be affected as the percentage of cross posting is low.

(d The plot of learning curve is shown below:



Observations: The accuracy is less initially and with number of training

examples it increases.

(e) The confusion matrix for all types of splits ins consolidated as below

191.2	2.4	0.4	0	1.6	0.2	2	0
3.4	193.2	0.2	0.4	1	0	1	0
0.4	0.4	193.6	0.6	0.6	0	0.6	0
0.4	1.2	1.6	195.2	0.2	0	1.2	0
0	0.6	0	0	175.2	0.4	5	0.6
0.2	0.6	0.2	0.4	1	182	2.2	1.4
0.6	0	0.2	0	14.6	2.8	135.4	1.4
0.4	1.2	0	0	7.2	1.8	4.4	110.6

Observations: The category for maximum diagonal entry is rec.sport.hockey. The maximum non-diagonal entries are for talk.politics.guns and talk.politics.misc