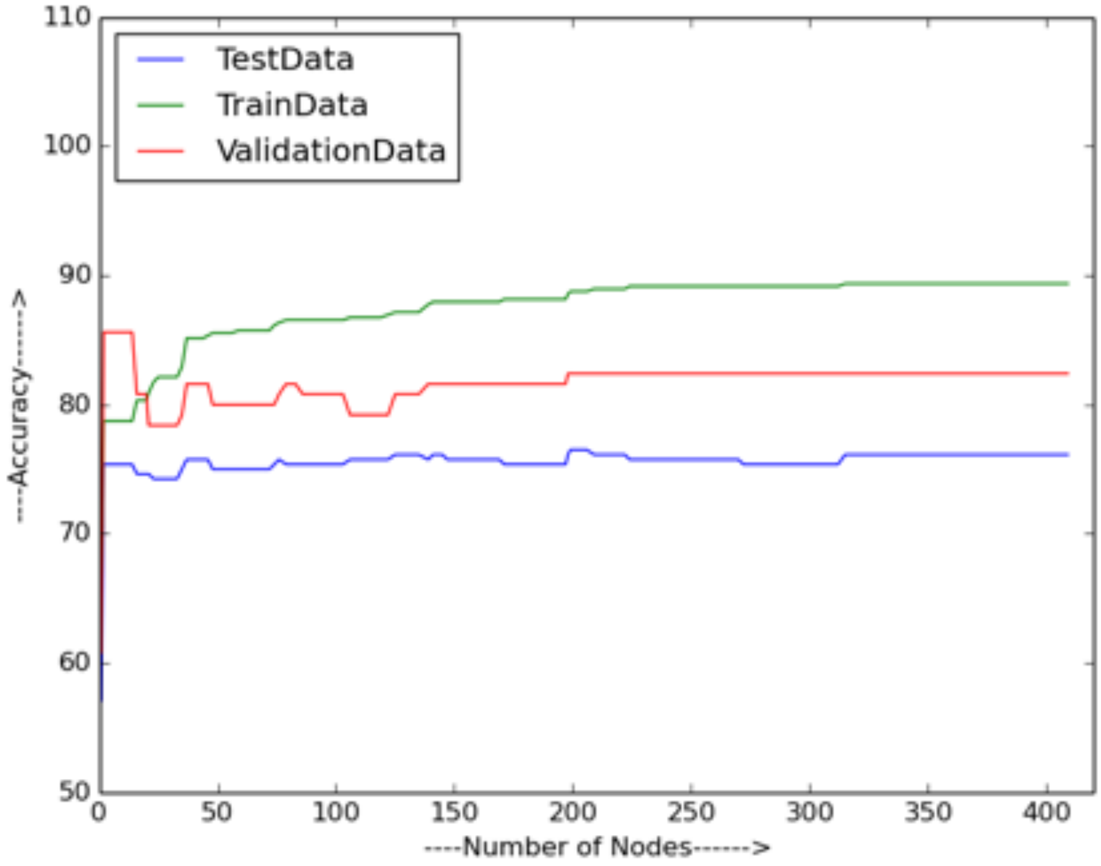
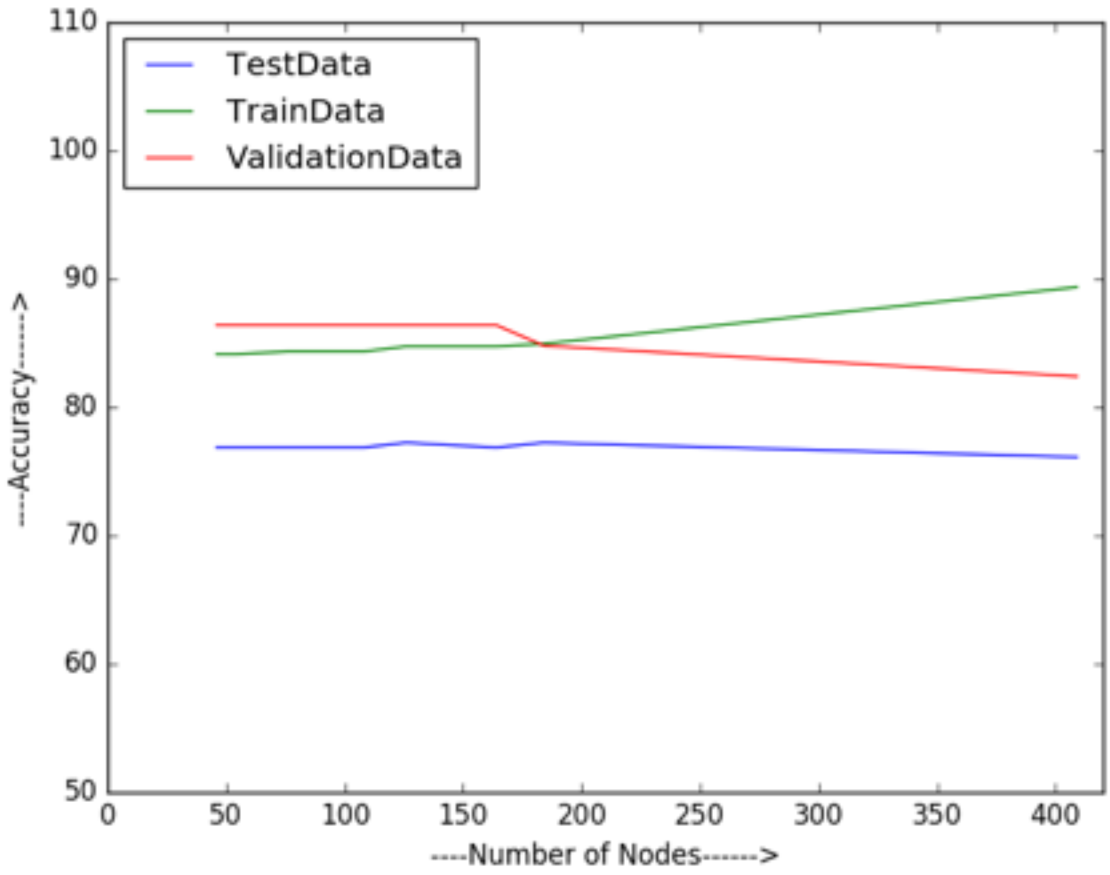
ASSIGNMENT-3 REPORT

Q1a) The number of nodes obtained: 409
Number of Internal nodes: 165
Train Data Accuracy: 89.3574%
Validation Data Accuracy: 82.4%
Test Data Accuracy: 76.12%



The accuracies are dependent on the number of nodes of tree. If the number of nodes increases then accuracy increases to a certain limit and then accuracies fall due to overfitting of data. This excess number of nodes needs to be pruned to give better accuracies.

Q1b) The number of nodes post pruning are: 46
The validation set accuracy is increased to : 86.4 %



The Train set accuracy is decreased to 84.14 %
The Test set accuracy is increased to : 76.9 %

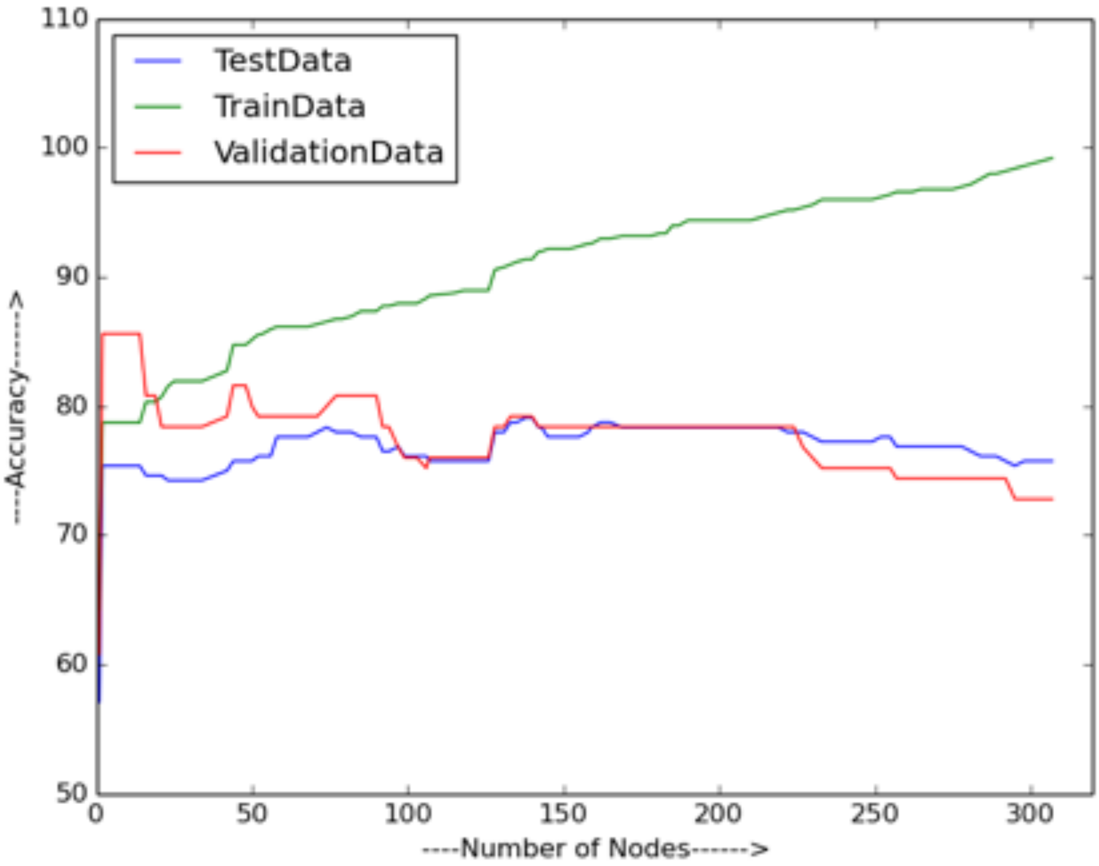
Finally we can conclude that on pruning not only we can reduce the size of the decision tree but also avoid overfitting and thereby increase accuracies.

Q1c) Number of nodes obtained: 307
Number of Internal nodes: 135
Train Data Accuracy: 99.197 %
Validation Data Accuracy: 72.8 %
Test Data Accuracy: 75.75 %

The results of the first part is better than that of the second part because the attribute is split again and again causing overfitting in the second case. And hence only the training set accuracies are high and both the validation and test set accuracies are very low.

Table 1

	MAX SPLITS	TRESHOLDS
Pclass	0	
Sex	0	
Age	5	25.0, 32.0, 29.0, 27.5, 26.0
SibSp	2	1.0, 0.0
Parch	1	0.0
Ticket	4	244335.5, 30669.5, 228665.5, 244270
Fare	2	8.05, 16.10
Embarked	0	
Cabin_a	0	
Cabin_b	1	62.0

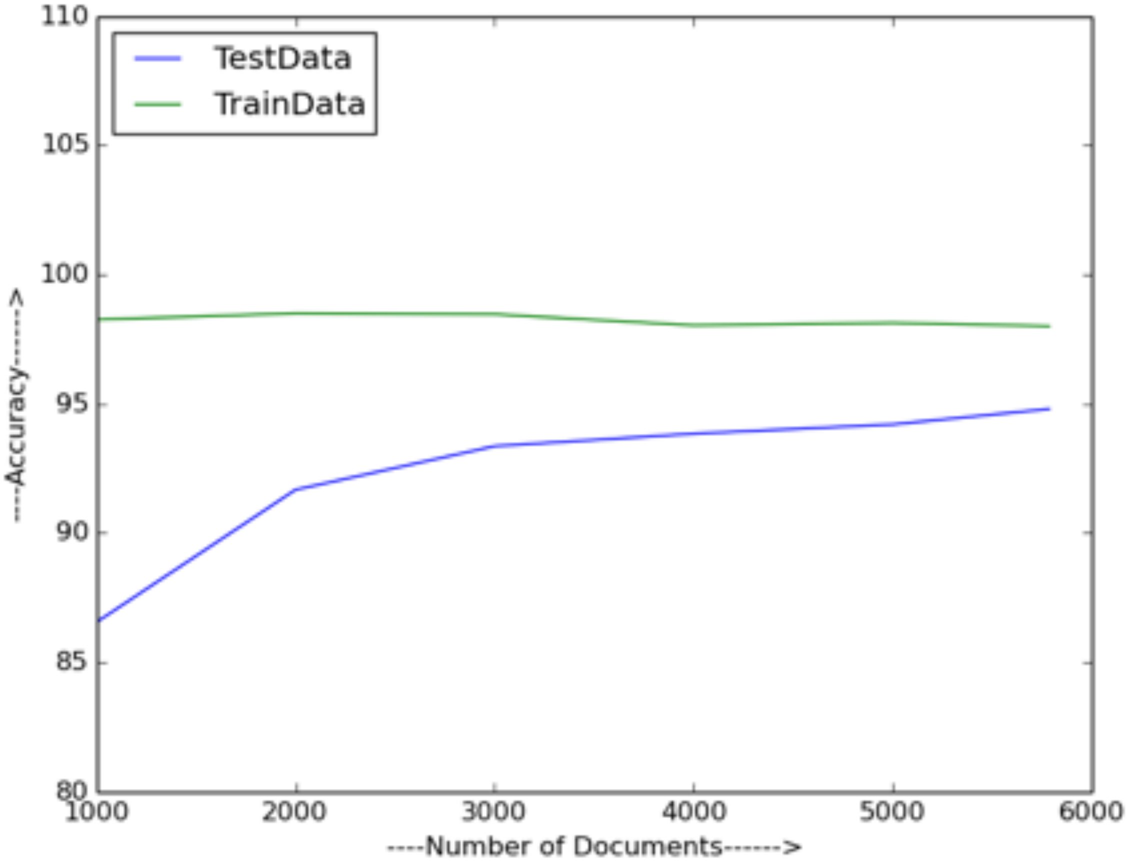


Q1d) The setting considered is

criterion = “entropy”, splitter=“best”, max_depth=6, min_samples_split=4, min_samples_leaf=3

The training accuracy = 90.3 %
The validation set accuracy = 81.6 %
The test set accuracy = 80.6 %

As the max_depth is limited to 6, the overfitting is avoided. And min_samples_split and min_samples_leaf ensure that a decision is taken considering fairly sufficient number of examples.



Q2a) The average test set accuracy obtained: 94.799 %
The average accuracy for Train set: 98.04 %

Test accuracies for 5-fold validation :
[94.05 %, 95.09 %, 95.02 %, 94.4 %, 95.44 %]

Q2b) The random accuracy obtained: 12.96 %
Improvement = (94.8 - 12.96)/12.96
= 6.3 times

Q2c) It does not create much problem when the data is very huge because the learning algorithms considers very less probability of those articles belonging to a particular group.It depends on the amount of data and their corresponding probabilities calculated.

Q2d) The below graph shows how the accuracy varies with the number of examples. With very few examples the accuracies very low.

From the learning curve we can clearly notice the improvement in accuracies because the probability of a word to be in a particular group also increases.

NUMBER OF EXAMPLES TAKEN	AVERAGE ACCURACIES OBTAINED
1000	86.54%
2000	91.67%
3000	93.36%
4000	93.85%
5000	94.2%
5784	94.8%

Q2e) The confusion matrix is shown below.

	re.autos	rec.motorcycles	rec.sport.baseball	rec.sport.hockey	talk.politics.guns	talk.politics.mideast	talk.politics.misc	talk.religion.misc
rec.autos	956	20	4	0	4	0	4	1
rec.motorcycles	20	971	1	0	2	0	1	1
rec.sport.baseball	3	6	972	12	0	0	0	1
rec.sport.hockey	3	7	19	966	0	0	4	0
talk.politics.guns	1	5	1	0	873	1	25	3
talk.politics.mideast	1	9	2	2	7	878	26	15
talk.politics.misc	4	3	4	1	67	7	668	21
talk.religion.misc	1	8	0	0	30	6	13	570

The group “re.sport.baseball” has the highest value of diagonal entry in the confusion matrix.

The groups “talk.politics.misc” and “talk.politics.guns” are the most confused groups with respect to each other.

This matrix tries to depict the dependency between each of the two groups. And also establishes which two groups have many similar words in vocabulary.