CS57300: Data Mining

Assignment 4

Name: Mohammad Haseeb
Purdue ID: mhaseeb@purdue.edu

Due: March 31, 2019

Note: Figures appear at different locations in the document due to position issues of LaTeX.

1 Preprocessing

The files 'testingSet.csv' and 'testSet.csv' are stored in the same directory from which the script 'preprocess-assg4.py' was run.

2 Decision Trees, Bagging and Random Forests

(i) Figure 1 shows the output after training and testing Decision Tree. It takes approximately 25 seconds to run this script on data.cs.purdue.edu.

```
data 81 $ python3 trees.py trainingSet.csv testSet.csv 1
Training Accuracy DT: 0.78
Testing Accuracy DT: 0.73
```

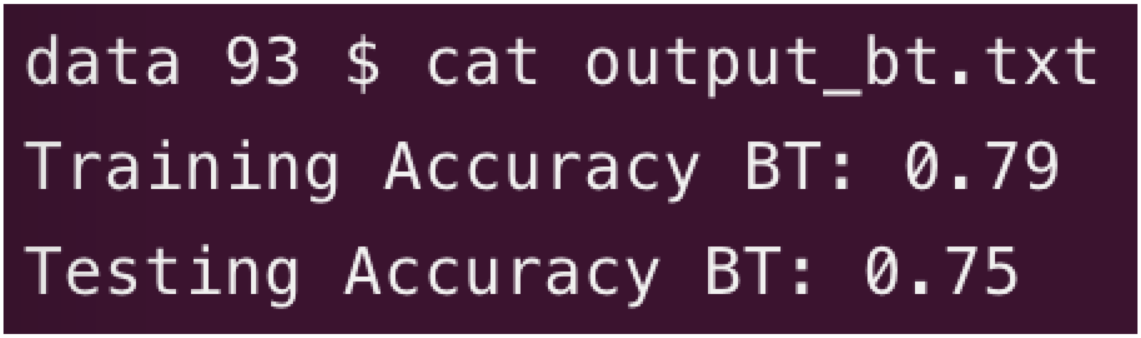
Figure 1: Output of trees.py for DT

(ii) Figure 2 shows the output after training and testing Bagging. It takes approximately 11 minutes to run this script on data.cs.purdue.edu.

(iii) Figure 3 shows the output after training and testing Random Forests. It takes approximately 3 minutes to run this script on data.cs.purdue.edu.

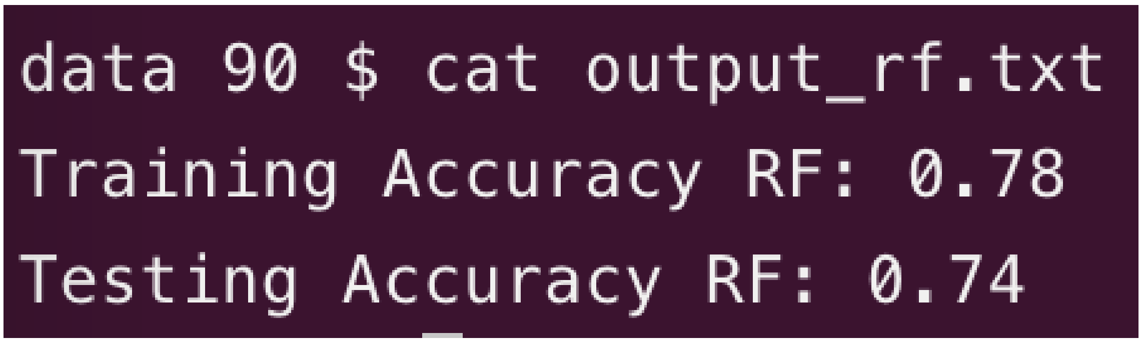
3 Influence of Tree Depth on Classifier Performance

(a) The learning curves for the algorithms are shown in Figure 4. It takes approximately 3 hours and 20 minutes to run this script.



```
data 93 $ cat output_bt.txt
Training Accuracy BT: 0.79
Testing Accuracy BT: 0.75
```

Figure 2: Output of trees.py for BT



```
data 90 $ cat output_rf.txt
Training Accuracy RF: 0.78
Testing Accuracy RF: 0.74
```

Figure 3: Output of trees.py for RF

(b) The output of my hypothesis testing script (hyp_testing.py) if shown in Figure 5.

4 Compare Performance of Different Models

(a) The learning curves for the algorithms are shown in Figure 6. It takes approximately 2 hours and 12 minutes to run this script.

(b) The output of my hypothesis testing script (hyp_testing.py) if shown in Figure 7.

5 Influence of Number of Trees on Classifier Performance

(a) The learning curves for the algorithms are shown in Figure 8. It takes approximately 4 hours and 50 minutes to run this script.

(b) The output of my hypothesis testing script (hyp_testing.py) if shown in Figure 9.

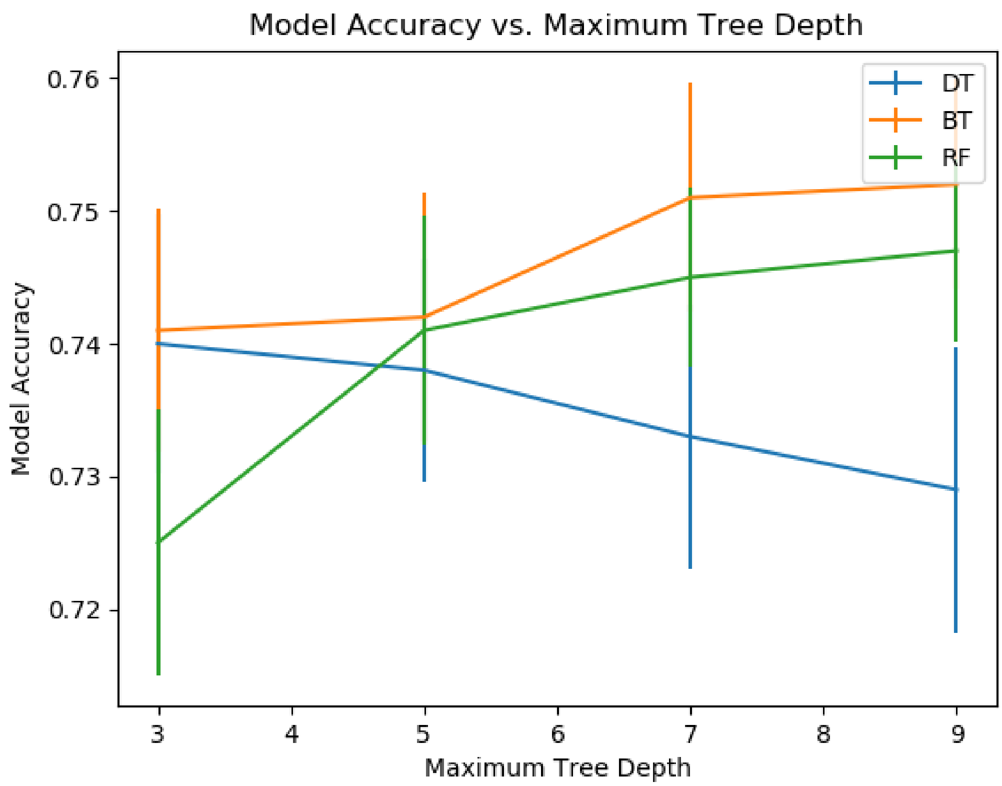


Figure 4: Average Model Accuracy v.s. Maximum Tree Depth for DT, BT and RF

```
data 275 $ python3 hyp_testing.py
H0: As the tree depth increases, the mean accuracies for both DT and BT remain the same i.e. their performance does not change with respect to each other.
H1: As the tree depth increases, the mean accuracies of DT != mean accuracies of BT i.e. their performance differs with respect to each other.
Depth: 3 H0 for DT and BT: t-statistics = -0.5570860145311556, p-value = 0.5910512317836047 Reject with significance level of 0.05? False
Depth: 5 H0 for DT and BT: t-statistics = -0.8017837257372732, p-value = 0.44333185016966015 Reject with significance level of 0.05? False
Depth: 7 H0 for DT and BT: t-statistics = -2.4246715773614604, p-value = 0.03831594659876352 Reject with significance level of 0.05? True
Depth: 9 H0 for DT and BT: t-statistics = -3.2857142857142865, p-value = 0.009442721748495789 Reject with significance level of 0.05? True
```

Figure 5: Hypothesis Testing

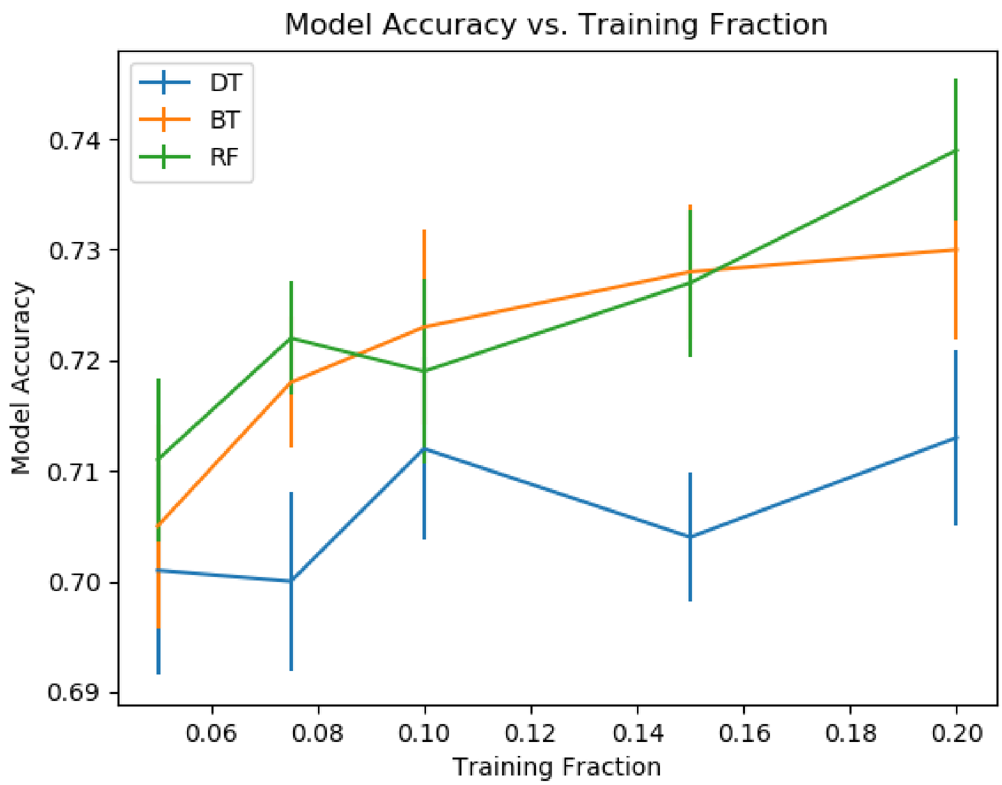


Figure 6: Average Model Accuracy v.s. Training Fraction for DT, BT and RF

```
data 293 $ python3 hyp_testing.py
H0: As the training fraction increases, the mean accuracies for both DT and RF changes i.e. their performance changes with respect to each other.
H1: As the training fraction increases, the mean accuracies of DT and RF do not change i.e. their performance remains the same with respect to each other.
Fraction: 0.05 H0 for DT and BT: t-statistics = -1.5932550136313814, p-value = 0.14556709184183406 Reject with significance level of 0.05? False
Fraction: 0.075 H0 for DT and BT: t-statistics = -3.1151495115351384, p-value = 0.012415333159229299 Reject with significance level of 0.05? True
Fraction: 0.1 H0 for DT and BT: t-statistics = -1.6329931618554518, p-value = 0.13690412558075216 Reject with significance level of 0.05? False
Fraction: 0.15 H0 for DT and BT: t-statistics = -3.851204448083039, p-value = 0.0038991282005206144 Reject with significance level of 0.05? True
Fraction: 0.2 H0 for DT and BT: t-statistics = -2.850765804418486, p-value = 0.019065356473909133 Reject with significance level of 0.05? True
```

Figure 7: Hypothesis Testing

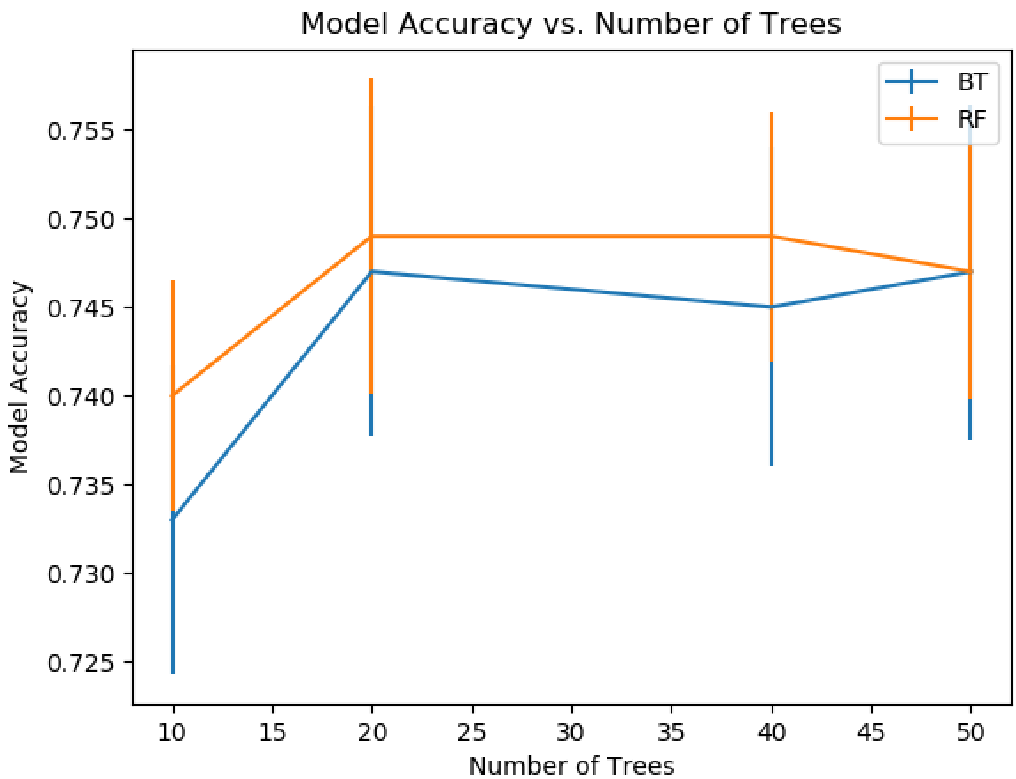


Figure 8: Average Model Accuracy v.s. Number of Trees for BT and RF

```
data 342 $ python3 hyp_testing.py
H0: As the number of trees increases, the mean accuracies for both RF and BT remain the same i.e. their performance does not change with respect to each other.
H1: As the number of trees increases, the mean accuracies of RF != mean accuracies of BT i.e. their performance changes with respect to each other.
Number of Trees: 10 H0 for DT and BT: t-statistics = 0.9780675089485068, p-value = 0.35359758396894503 Reject with significance level of 0.05? False
Number of Trees: 20 H0 for DT and BT: t-statistics = 0.23076923076923075, p-value = 0.8226545408271235 Reject with significance level of 0.05? False
Number of Trees: 40 H0 for DT and BT: t-statistics = 0.5827715174143585, p-value = 0.5743560190941981 Reject with significance level of 0.05? False
Number of Trees: 50 H0 for DT and BT: t-statistics = 0.0, p-value = 1.0 Reject with significance level of 0.05? False
```

Figure 9: Hypothesis Testing