CS57300: Assignment 4
Name: Tunazzina Islam, PUID: 0031294421

1. Preprocessing:

```
python preprocess-assg4.py dating-full.csv trainingSet.csv testSet.csv
```

2. Implement Decision Trees, Bagging and Random Forests:

(i)

```
python trees.py trainingSet.csv testSet.csv 1
```

Output:

```
Training Accuracy DT: 0.76
Test Accuracy DT: 0.73
```

ii)

```
python trees.py trainingSet.csv testSet.csv 2
```

Output:

```
Train Accuracy BT: 0.78
Test Accuracy BT: 0.75
```

iii)

```
python trees.py trainingSet.csv testSet.csv 3
```

Output:

```
Train Accuracy RF: 0.76
Test Accuracy RF: 0.72
```

3. The Influence of Tree Depth on Classifier Performance:

```
python cv_depth.py trainingSet.csv testSet.csv
```

a) Plot the learning curve with error bars:

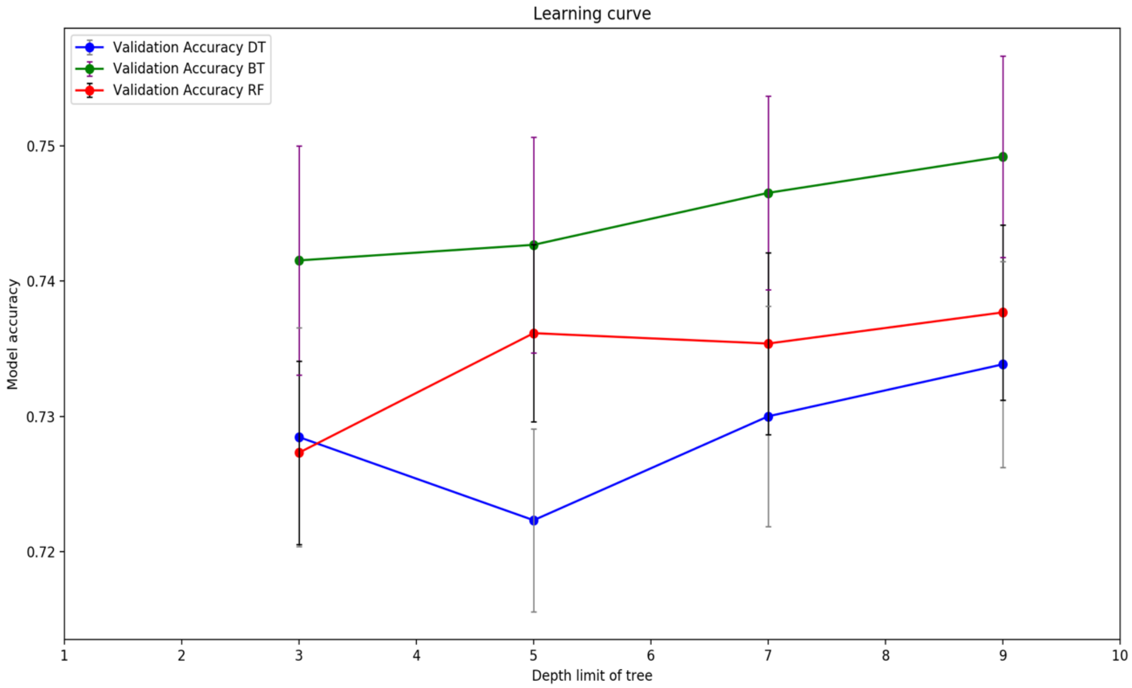

Figure 1: Learning curves for 3 models (DT, BT, RF) tree depth vs. accuracy including error bars.

b) Hypothesis Testing:
Given, significance level = 0.05

H0 and corresponding t-statistics, p-values, as well as reject or not are as follows:
Depth: 3 H0 for BT and RF: t-statistics = 1.83, p-value = 0.10
Reject with significance level of 0.05? False

Depth: 5 H0 for BT and RF: t-statistics = -0.12, p-value = 0.91
Reject with significance level of 0.05? False

Depth: 7 H0 for BT and RF: t-statistics = 1.46, p-value = 0.18
Reject with significance level of 0.05? False

Depth: 9 H0 for BT and RF: t-statistics = 0.89, p-value = 0.96
Reject with significance level of 0.40? False

4. Compare Performance of Different Models:

```
python cv_frac.py trainingSet.csv testSet.csv
```

a) Plot the learning curve with error bars:

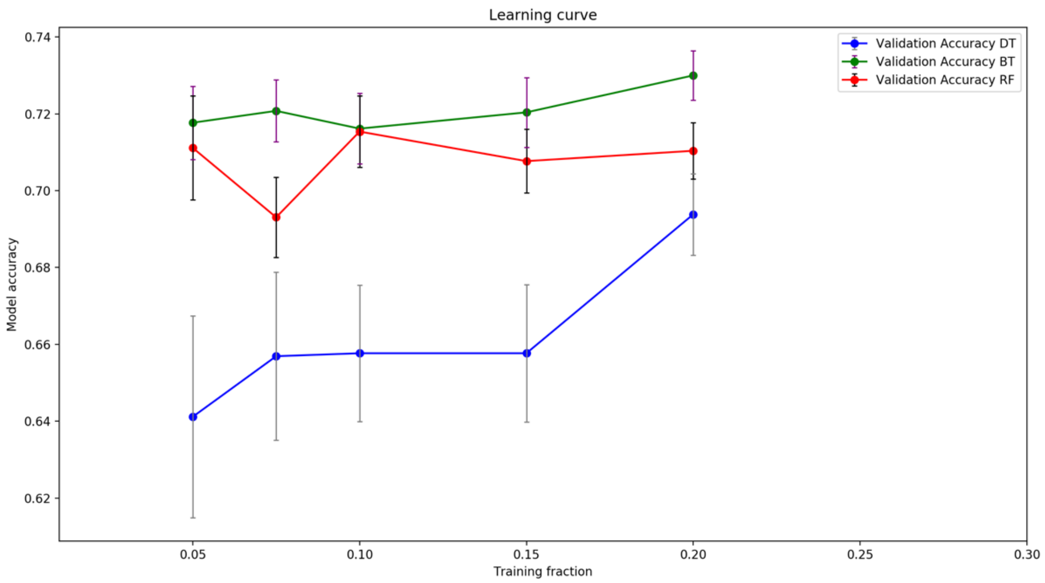

Figure 2: Learning curves for 3 models (DT, BT, RF) training fraction vs. accuracy including error bars.

b) Hypothesis Testing:
Given, significance level = 0.05

H0 and corresponding t-statistics, p-values, as well as reject or not are as follows:
Fraction: 0.05 H0 for BT and RF: t-statistics = 1.14, p-value = 0.28
Reject with significance level of 0.05? False

Fraction: 0.075 H0 for BT and RF: t-statistics = 1.70, p-value = 0.12
Reject with significance level of 0.05? False

Fraction: 0.1 H0 for BT and RF: t-statistics = 1.93, p-value = 0.09
Reject with significance level of 0.05? False

Fraction: 0.15 H0 for BT and RF: t-statistics = 0.05, p-value = 0.96
Reject with significance level of 0.05? False

Fraction: 0.2 H0 for BT and RF: t-statistics = 2.91, p-value = 0.02
Reject with significance level of 0.05? True

5. The Influence of Number of Trees on Classifier Performance:
python cv_numtrees.py trainingSet.csv testSet.csv

a) Plot the learning curve with error bars:

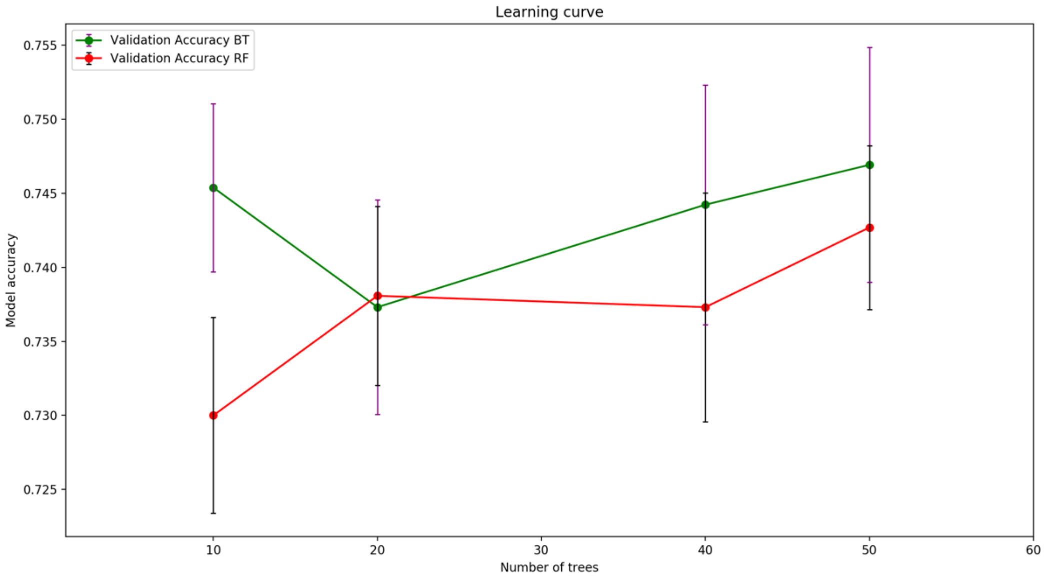

Figure 3: Learning curves for 2 models (BT, RF) Number of Trees vs. accuracy including error bars.

b) Hypothesis Testing:
Given, significance level = 0.05

H0 and corresponding t-statistics, p-values, as well as reject or not are as follows:
Number of trees: 10 H0 for BT and RF: t-statistics = 1.75, p-value = 0.11
Reject with significance level of 0.05? False

Number of trees: 20 H0 for BT and RF: t-statistics = -0.11, p-value = 0.91
Reject with significance level of 0.05? False

Number of trees: 40 H0 for BT and RF: t-statistics = 1.15, p-value = 0.28
Reject with significance level of 0.05? False

Number of trees: 50 H0 for BT and RF: t-statistics = 0.71, p-value = 0.50
Reject with significance level of 0.05? False

……………………………………… Bonus Points…………………………………..

i) I implemented Neural Network using bias. The I chose the regularization parameter reg_lambda = 0.0001, step size (learning rate) = 0.0001 and number of iterations = 1000 Run the code:

```
python neural_network_bonus.py trainingSet.csv testSet.csv
Training Accuracy NN: 0.75
Testing Accuracy NN: 0.73
```

I did 10-fold cross validation and I got following validation accuracy:

```
10-fold Validation Accuracy NN: 0.75
10-fold Validation Accuracy NN: 0.77
10-fold Validation Accuracy NN: 0.74
10-fold Validation Accuracy NN: 0.72
10-fold Validation Accuracy NN: 0.8
10-fold Validation Accuracy NN: 0.77
10-fold Validation Accuracy NN: 0.75
10-fold Validation Accuracy NN: 0.78
10-fold Validation Accuracy NN: 0.77
10-fold Validation Accuracy NN: 0.76
```

Average validation accuracy after 10-fold cross validation:

```
Average 10-fold cross validation accuracy of NN:  0.76
```