Jefferson Parker

CS677

April 18, 2022

Homework 5, Written Answers

**Question 1.**

* 1. – See code in japarker\_hw5\_1.py for loading of Excel data file to Pandas data frame.
  2. – See code in japarker\_hw5\_1.py for mapping of NSP labels to new column with two groups.

**Question 2.**

2.1 – See code in JparkerHw5Helper.py for the 50/50 split of data into training and test. See code in japarker\_hw5\_2.py for Naïve Bayes Classifier prediction and execution.

2.2 – The accuracy of the Naïve Bayes classifier was 0.805.

2.3 – The confusion matrix for the Naïve Bayes Classifier is:

Chart, treemap chart

Description automatically generated

**Question 3.**

3.1 – See japarker\_hw5\_3.py for Decision Tree prediction and execution.

3.2 – The accuracy of the Decision Tree Classifier was 0.886.

3.3 – The confusion matrix for the Decision Tree Classifier is:

Chart, treemap chart

Description automatically generated

**Question 4.**

4.1 – The Random Forest classifier for N = 1-10 and d = 1-5 was coded and executed in japarker\_hw5\_4.py.

4.2 – The Random Forest error rates for N = 1-10 for each level of d = 1-5 is plotted below:

Chart, line chart

Description automatically generated

4.3 – The best combination of N and d changes with each run of the Random Forest model. Past combinations were N = 2, d = 1 and N = 6, d = 1. As of this run, the best combination is N = 3, d = 1.

4.4 – The confusion matrix for the Random Forest classifier with N = 3 and d = 2 is:

Chart, treemap chart

Description automatically generated

**Question 5.**

The summary metrics table for the three models is:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TP** | **FP** | **TN** | **FN** | **accuracy** | **TPR** | **TNR** |
| Naïve Bayes | 139 | 98 | 717 | 109 | 0.805 | 0.560 | 0.880 |
| Decision Tree | 183 | 56 | 759 | 65 | 0.886 | 0.738 | 0.931 |
| Random Forst (3,2) | 32 | 82 | 733 | 216 | 0.720 | 0.129 | 0.899 |

In general, I think the data may have contained more negatives than positives in general (as indicated by the high true negative rates). Second, I have my annotations for my confusion matrices backwards somehow.

In terms of performance, I am mildly surprised that the Decision Tree had the highest performance over both Naïve Bayes and Random Forest. However, the Random Forest models all had significant increases in error at higher depths so maybe this is just a feature of the data rather than the models.

Altogether, I think this particular combination of features is not particularly well suited to prediction by these modeling algorithms and a different method would need to be explored to use it in any real application.