Divya Thomas

Class: CS 677

Date: 4/15/2023

Non-code answers to assignment.

**Question 1.** Pandas dataframe is printed in question1.py file.

**Question 2-1.** Prediction dataframe printed in question2.py file.

**Question 2-2.** Accuracy: 75.16%

**Question 2-3.**

Confusion Matrix:

[[ 39 195]

[ 69 760]]

**Question 3-1.** Prediction dataframe printed in question3.py file.

**Question 3-2.** Accuracy: 87.39%

**Question 3-3.**

Confusion Matrix:

[[179 55]

[ 79 750]]

**Question 4-1.** Dataframe holding all combinations’ error rates found in output for question4.py

**Question 4-2.** Please see error rate plot below.

Chart, line chart

Description automatically generated

Based off this plot, the best combination of N and d (lowest error rate) is N=9, d=5.

**Question 4-3.** Accuracy of Random Forest (N=9, d=5) is 90.03%

**Question 4-4.**

Confusion Matrix for Random Forest (N=9, d=5):

[[168 66]

[ 41 788]]

**Question 5.** See result summary below for all classifiers.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TP** | **FP** | **TN** | **FN** | **Accuracy (%)** | **TPR** | **TNR** |
| **NB** | 760 | 195 | 39 | 69 | 75.16 | 0.92 | 0.17 |
| **DT** | 750 | 55 | 179 | 79 | 87.39 | 0.9 | 0.76 |
| **RF (N=9, d=5)** | 788 | 66 | 168 | 41 | 90.03 | 0.95 | 0.72 |

While all models tested seem to be significantly accurate with predicting Normal (1) values, as reflected from the equally high TPR values for all three models, the Naïve Bayesian method has a significantly lower accuracy with Abnormal (0) values. This low TNR is the main factor in bringing this method down to the lowest accuracy of the three. The other two methods, Decision Tree and Random Forest, are pretty close in all measures. That being said, Random Forest does score slightly higher than Decision Tree overall, making it the most accurate method of the three.