Ryan Christopher CS677

Assignment 5

-The Fetal Cardiogram Strikes Back-

Group chosen: 2
Features worked with: ASTV, MLTV, Max, Median

Question 1:

Part 1 - assignClass.py loads the excel file worksheet titled “raw data” and creates a dataframe using the read_excel method in pandas.

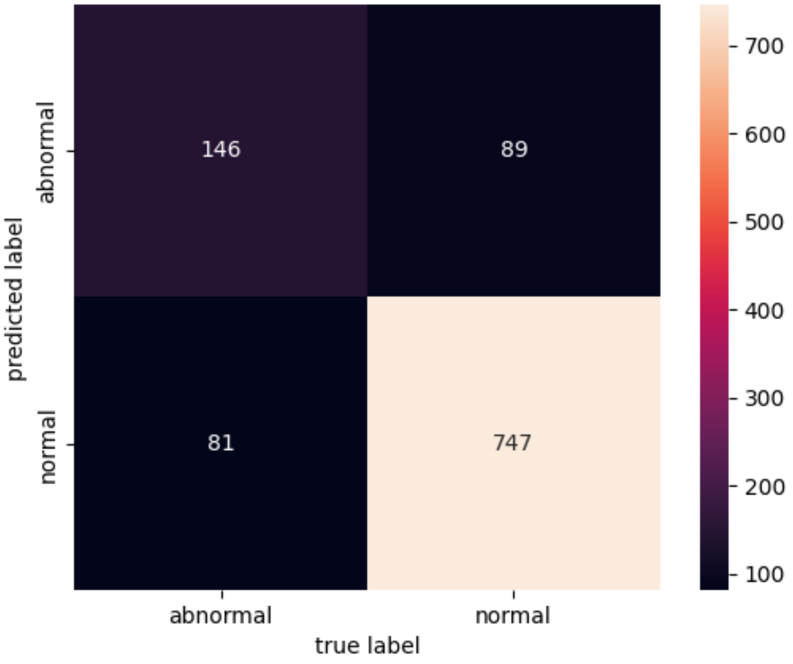
Part 2 - Within assignClass.py, the function assignClass drops all extra columns, and separates all entries into either normal or abnormal. Then, the NSP label is set to 1 for all normal entries and 0 for all abnormal entries. The two groups are then combined again and the function returns the dataframe for use in the following questions.

Question 2:

Part 1 - The function nb within nbClassify.py takes the dataset, splits it 50/50, then trains the GaussianNB model with the training set and predicts the class labels for the test set.

Part 2 - The accuracy for the NB classifier is 84.01%.
(Accuracy rounded to 2 decimal places, unrounded value can be seen when nbClassify.py is run)

Part 3 -

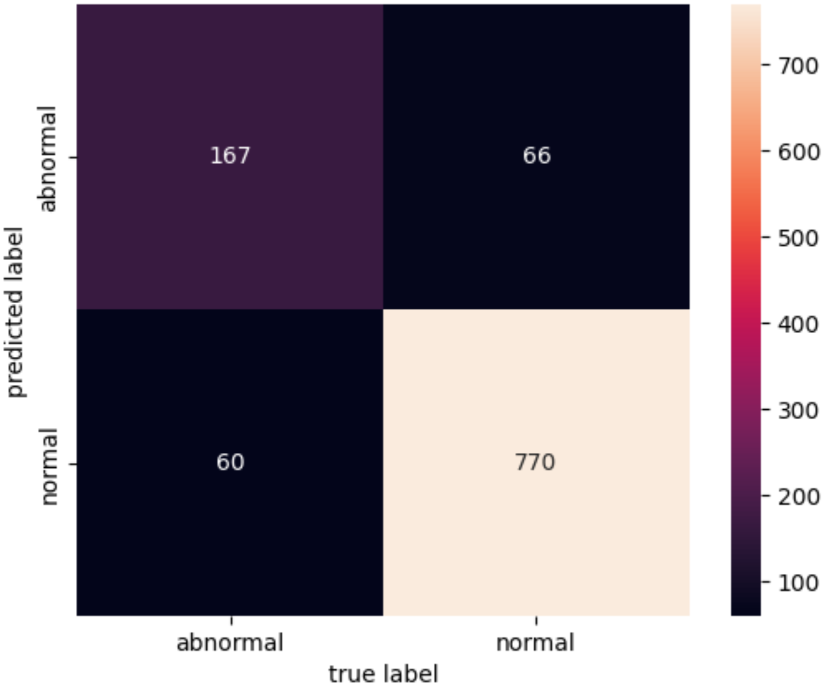


Question 3:

Part 1 - The function decisionTree within decisionTree.py takes the dataset, splits it 50/50, then trains the Decision Tree model with the training set and predicts the class labels for the test set.

Part 2 - The accuracy for the Decision Tree classifier is 88.15%.

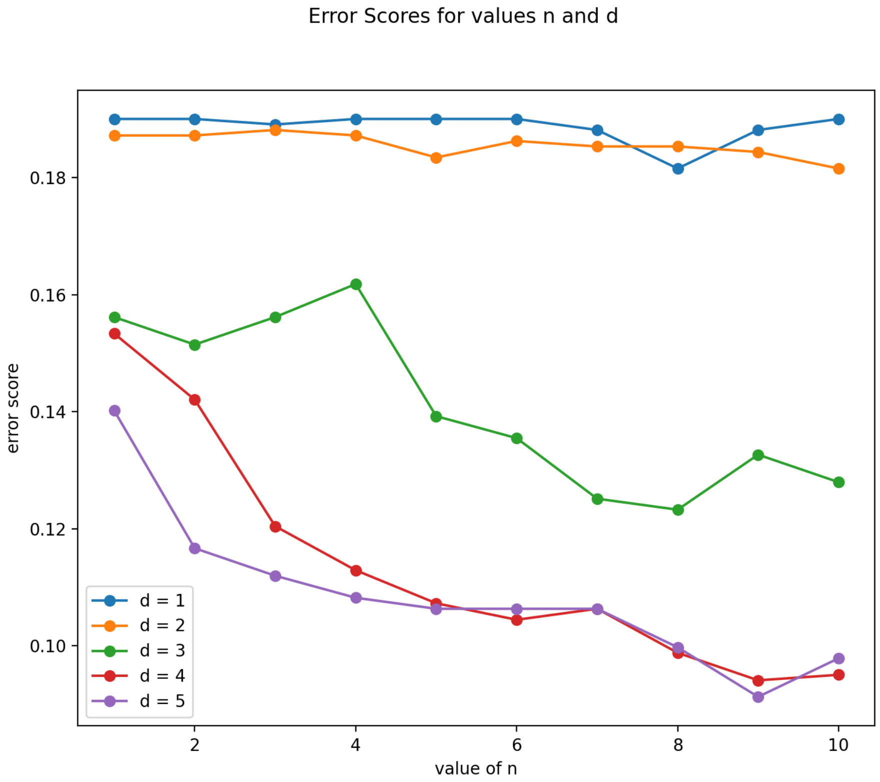
Part 3 -



Question 4:

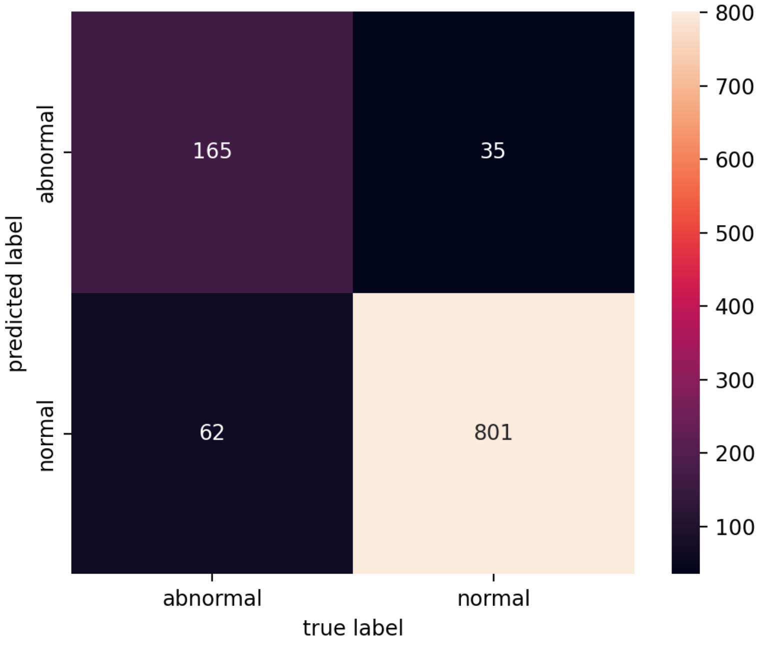
Part 1 - randomForest in randomForest.py uses the random tree classifier and applies n = 1 through 10 for number of subtrees and d = 1 through 5 for max depth of each subtree. While applying the classifier to each iteration, it calculates the error score and stores the best combination of n and d.

Part 2 -



Part 3 - The best accuracy, 90.87%, is when n = 9 and d = 5.

Part 4 - Confusion matrix for n = 9 and d = 5:



Question 5:

*note - accuracy, TPR, and TNR are displayed rounded to 2 decimal places.
In this case, normal (class 1) is considered positive and abnormal (class 0) is considered negative

Model	TP	FP	TN	FN	Accuracy	TPR	TNR
Naive Bayesian	747	89	146	81	84.01%	90.22%	62.13%
Decision Tree	770	66	167	60	88.15%	92.77%	71.67%
Random Forest	801	35	165	62	90.87%	92.82%	82.50%

The best combination of n and d for the random forest classifier yielded the highest accuracy at 90.87%, with the decision tree being the second most accurate at 88.15% and the naive bayesian having the lowest accuracy at 84.01%. Each of the models had similar TPR values that ranged from 90.22% to 92.82%, however the TNR values show large differences as the naive bayesian has a TNR of 62.13%, the decision tree has a TNR of 71.67%, and the random forest has a TNR of 82.50%. Overall the three models were noticeably better at predicting normal entries than abnormal entries.