Chris DiBello

CSC570R-Advanced Validation Assignment

1. Python Code is Implemented at github :

https://github.com/dibello99/cdibe01s/ChrisDiBelloAdvancedValidation.py

writeup at https://github.com/dibello99/cdibe01s/ChrisDiBelloAdvancedValidationWriteUp.docx.

2 My K-Fold score was slightly higher than AUC at .961 for K-Fold as opposed to

.96 for my AUC score. AUC is restricted to binary classification task or multilabel classification.

It is possible that we could get a different score if we used a different data format. AUC is area under a curve and represents the area from a dataset X and dependent variable y - in this case malignant.

The K-fold score is a cross validation split of model data set in this case 10 folds. You would think the K-Fold score is higher because it has more precision- in my case it is slightly higher but similar to AUC. The K-fold should produce a more ccurate model using a k-1 split into more learning modules to analye.

3. Precision is the percentage of data classified which are true positives. Recall represents the percentage of identified positives. Accuracy is a measure of based on the proportion of correct predictions from our model. AUC is the area under a curve as we discussed in number 2. In our model our values for accuracy, precision, recall and AUC are very high. I think this is due to the fact that our input dataset is very clean - we have no missing values or alphanumeric data and all data is in a floating point format with decimals. This is the easiest data to work with. In real life data is much more messy.