Ethan Colley

Project 2-Classification

For this project we were tasked with taking a dataset with breast cancer patients and using classification to find out which patients had recurrence issues. To do this, we had to take the datasets and make sure all datatypes were useable and then use at least three methods to perform classification on the dataset and find out if they can accurately predict the recurrence events.

Part 1:

The first step in getting to our classification is performing all necessary preliminary tasks on the data. After looking over the csv file, I imported pandas and numpy and found that there are 286 rows and 10 columns(class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat). When viewing the data in the csv I was able to see some "?" in certain data spot so the first thing I did was replace those. The columns "node-caps" and "breast-quad" had these question marks and the way I replaced them was with the mode of each column using the ".replace" function. Mean and median do not make sense here since they are object columns, but by choosing mode the question marks are just replaced with the most frequent variable in the column. Each column must be changed into a bool variable type using one hot encoding, but before doing this I did the univariate analysis using count plots for each object variable. Using "astype()" I converted each object variable to categorical variables and then created the countplots with seaborn. The first countplot is for the "age" column where we find that there is a largely normal distribution in the age ranges for the patients. Then I moved on to the "menopause" column for a count of menopause status. Here it shows that most patients are either premeno status or ge40 status, with very few being lt40. After this I ran count plots for "class" and "irradiat" columns, and I am grouping these together because I noticed the plots have almost the exact same shape, with irradiation “yes” almost directly corresponding to recurrence. Finally I made a histogram on “deg-malig”, the only numerical variable and it had a normal distribution with most patients being degree 2. Once all of these plots were realized, I was able to perform one hot encoding to create bool variables for classification, and move onto part 2.

Part 2:

To do actual classification in part 2, I first separated the data into test and training sets with a thirty-seventy ratio using “train_test_split”. The data had been set to predict recurrence events and the three methods I chose were decision trees, logistic regression, and k nearest neighbors. I output the statistics of these models after running them to get all different metric data for each model. I also output confusion matrices in the case of decision trees and logistic regression so I could easily see the numbers for false positives and false negatives in the models. The first model I ran was the decision trees model which had an accuracy of 71% for the test set and a recall of 44% with a precision of 50%. The model had 11 false positives and 14 false negatives which is not great. This model only having middling performance could not bode well for the model, but it must be compare to other examples first. The second model I ran was the logistic regression model, which had an accuracy of 70% and a recall of 32% with a precision of 47% on the test set. This model has 17 false negatives and 9 false positives. The final model used was the k nearest neighbors model, and it had an accuracy of 73% and a recall of 44% with a precision of 55%, meaning that this model has high numbers of false positives and false negatives. After performing gscv to find the most optimal hyperparameter, the knn model has an accuracy of 77%. No model is ideal for this situation, but given the improvement made in the knn model I would recommend it for any future use.

All model performance metrics are important in finding an acceptable model, but I believe the most important metric for this particular case is the recall. This is because in the case of breast cancer, a false negative would be much more detrimental since the patient would believe they are okay when they actually have breast cancer. Precision would also be important because a false positive would also be quite detrimental, but I believe a false negative would be just a bit worse in this case. Based on this decision, the best model for prediction would either be the base knn with a 0.44 recall or the logistic regression model which also has 0.44 recall. Either model would be acceptable in this case but once again I would recommend the knn model simply because it could be modified for a better recall value which will be explored in the bonus.

Bonus

For the bonus, I used the GridSearchCV method to improve the recall within the knn classifier. After running the function, I see that k=1 optimizes the recall and returns a recall value of 0.56, which is much better than any other models. In this case, this means that this would be the most optimal model for recall.