Constantinos Skevofilax
cgs2528

Project 2 Report - Breast Cancer Analysis

Preparing the data

To prepare the data, I first checked the size and shape of the raw data to see how large of a dataset I was working with. With 286 rows and 10 columns, there was a fair bit of data to analyze but a manageable size none the less. I checked to see the kind of data types these columns were by using the .info(), but this wasn't clear at first because majority were of type 'object'. After looking at the data frame and seeing why generally the data types were that of 'objects', I was able to see that the columns were more like categories as to describe the different ranges different health states take. After this, I checked to see if there were any missing values in the dataframe by using the .unique() function for each of the columns. I found that the 'node-caps' and 'breast-quad' columns had '?' values in 9 rows. I decided to use the mode, or the most frequently seen value, statistical value to handle these missing values because I cannot predict if a column case will be 'yes' or 'no' based off the available data trends. After this, I decided to get a better understanding of what the spread was on the data by plotting it. I created histogram and count plots to see the distribution of the age, menopause, tumor size, and invasive nodes for breast cancer patients. From my analysis, more people have breast cancer starting at age 30 up to 59, with the more frequent range between 40-59 years old. More people tend to have premeno and ge40 menopause compared to It40, which may be more predisposed to breast cancer. There is a higher frequency of people who have 'large' tumor sizes, with the tail ends being 10-14 and 40-44, with a higher bell curve around size 20-34. Finally, it seems that these large tumor sizes come within 0-2 counts, with more tumors such as 3 and above being in lesser frequency but still occuring amongst people.

Techniques to train the model

To train the model, I used the sklearn train_test_split function to split the model. But to be more specific, I needed to make the category 'type' columns into category types. This is because it splits up these

categories into booleans, which is better classification when splitting the data into training and testing sets. After this process, I noting that the dependent variables are all the columns, such as age, number of tumors, etc., except the recurrence of breast cancer, which is the independent variable. After dropping the recurrence columns from the dependents and setting said column to the Y value, I used the train_test_split() function to split the data into training and testing data. I decided to use K-Nearest Neighbor (KNN), Decision Tree, and Random Forest models to look for trends in the dataset to see if they could predict the recurrence of breast cancer.

KNN is a method that looks at a sample class, or data subset, on how close it is to all other data points to see if the sample class is the same class of the "nearest neighbor".

Decision Tree is a method that uses a tree structure of "decision rules" to model a dataset and predict the future values based on previous decisions made. However this is a relatively weak method as it builds one tree.

Random Forst similar to Decision Tree creates a decision tree structure but multiple based on random sampling of features. It then creates a single model to average all the trees together.

Model Performance against Dependent Prediction

KNN

There was an accuracy of 69%, with the training data having 86%. This means that with more data, the model was able get more accurate clustering with a K = 3. There was a recall score of 38.46%, 47.62% precision score, and 42.55% F1 score on the test data. There was a 15% improvement, which shows that there was a better prediction made than what was initially predicted on data it had not seen. However, with less than 50% precision, recall, and F1 scores, it shows that the model is accurate around ~48% of

the time. This is bad, as this is showing there were higher occurrences of false positives and false negatives, which are critical in medical diagnosis such as breast cancer assessment.

Decision Tree

There was an accuracy of 65% on the model's test performance, with a precision of class 0 of 72%, class 1 of 39%, 82% on recall on class 0 and 27% on class 1. The training performance was much better, with 97% accuracy, with a 96% precision of class 0, 100% on class 1, 100% recall on class 0 and 90% on class 1, 98% F1 score on class 0 and 95% on class 1. Overall, there was a drop in performance between the training and testing datasets, with overall low recall scores of 27% in the testing performance. This suggests that there was overfitting on the training data, which is not good for the accuracy of our model.

Random Forest

There was an accuracy of 29% on the model's test performance, with a precision of class 0 of 0%, class 1 of 29%, 0% on recall on class 0 and 96% on class 1. The training performance was barely better, with 30% accuracy, with a 100% precision of class 0, 30% on class 1, 1% recall on class 0 and 100% on class 1, 3% F1 score on class 0 and 46% on class 1. There was a consistent holding between the training and testing model performances, of 29% and 30%, but the model overall performance was rather poor. This suggests that there were many missed positive cases, which is very severe in the case of cancer diagonsis.

Overall, the most important metric to optimize for model performance would be recall. We want to reduce the number of false negatives that are released to the public, as false positives are much better in the state of assessing life scenarios than false negatives. Therefore, optimizing to reduce the number of false negatives is key to optimizing model performance in this scenario. I would recommend KNN, as both the overall model accuracy was high and three was ways to optimize the recall score to find the best nearest neighbor. Also in the field of finding the predictable factors of cancer, finding the trends between N neighbors is helpful for finding outliers in cancer diagnosis.