COE 379L Project 02: Breast Cancer

This project deals with a breast cancer database, which provides different statistics about patients including: age, tumor size, and location of tumor.

Preparing the Data

When looking at the data, the first thing I noticed was that some categories (node caps, irradiation, and breast quadrant) had values listed as “?”. To fix this, these values were set as the most common value in the list of possible values. Similarly, there were duplicate rows that I was able to remove from the data. After fixing missing and duplicate values, the data had a size of 272 rows by 10 columns.

We can now manipulate the categorical data to make it easier to input into the machine learning models. I changed the categorical columns ("class", "menopause", "breast", "breast-quad") into a larger set of one-hot-encoded columns. However, we now have to deal with the numeric values that are given in a range, such as “age”. I took the average value of each range and used that as the data value. For example, if an age range was “30-34”, I instead assigned a value of 32 to that data point. Now we can analyze the data.

Data Insights

The first thing we can do to analyze the data is to find the data distributions for each non-categorical column. Age and tumor size were roughly symmetrical and had outliers at both ends of the data. Meanwhile, the other data sets were skewed, as shown in their box plots in the Jupyter notebook.

Training the Models

For this problem, we will be using a few different models to predict the recurrence of breast cancer. We must first split the data into a training set and test set at a ratio of 7:3, keeping roughly the same proportion of recurrence in each set. We can now fit different models from Scikit Learn to this data. For this data, we want to minimize the number of false negatives (improve recall), so that all patients receive the care they need to help with recurring breast cancer, even if it increases the number of false positives.

The models I used were the K-Nearest Neighbor Classifier (KNN), Random Forest Classifier (RFC), and Logistic Regression (LR). The resulting scores of these models can be found in the Jupyter notebook. For RFC and KNN, I searched for hyperparameters to improve accuracy. At this stage, I would recommend the RFC model, as it has a similar accuracy but higher recall score when compared to the other methods.

K-Nearest Neighbor Classifier Scores

Data	Precision	Recall	F1 Score	Accuracy
Train	0.73	0.14	0.24	0.73
Test	0.60	0.12	0.21	0.72

Random Forest Classifier Scores

Data	Precision	Recall	F1 Score	Accuracy
Train	0.55	0.54	0.55	0.73
Test	0.50	0.58	0.54	0.71

Logistic Regression Scores

Data	Precision	Recall	F1 Score	Accuracy
Train	0.65	0.23	0.34	0.73

Test	0.57	0.33	0.42	0.73

Improving Recall

The recall score for these methods are quite low, considering the goal of this model. We can make some changes to improve recall. For the RFC and KNN classifiers we can search for hyperparameters that optimize the recall score instead of the accuracy score. Similarly, we can modify the decision threshold. For all three methods, I plotted the precision and recall scores as a function of the decision threshold. The decision threshold is what gives us the classification from these models, as they all predict a value between 0 and 1 from each patient's data. Decreasing the threshold in all cases improved the recall of the methods. However, it had the least effect on the RFC model, as it already has 1.00 recall.

K-Nearest Neighbor Classifier Scores

Data	Precision	Recall	F1 Score	Accuracy
Train	0.49	1.00	0.66	0.69
Test	0.33	0.67	0.44	0.51

Random Forest Classifier Scores

Data	Precision	Recall	F1 Score	Accuracy
Train	0.30	1.00	0.47	0.32
Test	0.30	1.00	0.46	0.30

Logistic Regression Scores

Data	Precision	Recall	F1 Score	Accuracy
Train	0.32	0.98	0.48	0.36
Test	0.31	1.00	0.47	0.34

Based on these results, I would use the adjusted LR model because it has almost 1.00 recall while having slightly higher of every other score than RFC. This model will allow us to predict the recurrence of breast cancer with an extremely low chance to have false negatives, allowing proper care for all patients.