Arvind Vinod
Stubbs
COE 379L

Project 1 Report

The goal of this project was to predict the recurrence of breast cancer in patients using a dataset with 9 independent features and 1 target variable (class). The project involved multiple exploratory data analysis methods, data preprocessing, and the application of classification models to predict whether a patient's cancer will recur or not. This paper explains the details regarding the treatment of the data before the models are run, the choice in models, and the resultant and its accuracy.

Firstly, the raw data was presented within a csv file. A Pandas DataFrame was used in order to turn the data into working material for the upcoming use cases. By utilizing the shape function, I was able to deduce the dataset size to be 385 rows. Duplicate rows were dropped as they do not improve the accuracy of our models. Next, by taking a look at the data types of the columns, they were mostly cast as objects. By treating this with type conversions, we are left with class, irradiat, and deg-malig being an int. Age, tumor-size, inv-nodes were ranges so a midpoint was used in order to convert these columns into floats. Node-caps were first turned into floats, a median was used in order to fill empty spaces, and then converted into ints. Menopause, breast, and breast-quad were turned into categories as they had a few unique values that they could be. Other null values were also sorted using the median. Utilizing seaborn, a few histograms, box plots, and bar graphs were drawn for certain columns. In order to help our model, we also put our categorical columns through one-hot encoding which produced extra columns but with boolean values instead.

From finding the shape of the dataset, we see that it is a sizable but functionally small dataset for machine learning purposes. I do not expect an extremely high accuracy from the models we will run. However, the data had very few duplicate rows which means that we are getting information from a wide range of patients. It also had very few null values which further creates variety in the data. Through univariate analysis producing a number of plots, some overarching conclusions can be drawn. These include deducing that most of the women were between their mid 40's to

60's, most tumors seem to fall in the 20s - 30s size, and most of the tumors seem to be present on the left breast.

To train the models, I utilized different procedures based on the nature of each classification algorithm. Firstly, for the K-Nearest Neighbors (KNN) classifier, I initialized the model with a specified number of neighbors (in our case 3) and trained it using the dataset by computing distances between data points and classifying among the nearest neighbors. In an attempt to refine and optimize this KNN model, I used GridSearchCV, a cross-validation technique that systematically evaluates different values of n_neighbors to find the k-value that betters the model. This attempts to realize better generalization and reduces risk of overfitting. This did not go as planned as explained below. For linear classification, I used Logistic Regression, a statistical model that applies the sigmoid function to estimate class probabilities. After training, all models were evaluated on the test set using metrics such as precision, recall, f1-score, support and accuracy to quantitatively evaluate their performance in distinguishing between classes effectively.

Now we evaluate the effectiveness of each model in predicting the class target. The first value will be for no recurrence and the second value will be for recurrence. For the regular KNN classifier, we produced the results: precision - 0.69, 0.40, recall - 0.94, 0.08, f1-score - 0.79, 0.14, support - 51, 24, and accuracy - 0.67. After using GridSearchCV, the results are: precision - 0.68, 0.31, recall - 0.82, 0.17, f1-score - 0.74, 0.22, support - 51, 24, and accuracy - 0.61. Finally for linear classification, we have: precision - 0.71, 0.80, recall - 0.98, 0.17, f1-score - 0.83, 0.28, support - 51, 24, and accuracy - 0.72. By all metrics, the linear classification model did the best for predicting recurrence and non-recurrence. As for GridSearchCV, one would obviously expect it to do better than the regular KNN classification. I am not completely sure why the results are worse but it is something I need to address in the future. I think the most important metric here is recall as cancer involves a high risk when producing false-negatives and low risk when producing false positives. So I would recommend the model to prioritize recall to reduce as many false negatives. In terms of confidence in the model, I think it can be much better. 0.72 in terms of accuracy is not great, especially when dealing with public health and so I would like to do more refinements.