Executive Summary

A careful review was conducted of the Wisconsin Breast Cancer data to model the data in an attempt to identify a way to predict with a diagnosis would be benign or malignant. The data that was modeled consisted of 10 measurements taken from biopsy samples. Each sample was assessed on each of the 10 measurement variables, and recorded. The data set also included the true diagnosis of each of the patient samples that were measured.

The data was then modeled in two different ways, to assess the probability of successfully predicting the diagnosis. The two modeling methods used were Naïve Bayes Classification, and Logistic Regression Classification.

After completing the modeling process, the data was reassessed for accuracy and interpretability. The Logistic Regression method produced the most interpretable results. There were several variables identified as significant contributors to the accuracy of the prediction. Each data point was assigned a probability of being benign, or malignant, and the predictions very accurate, evaluated under a 50% probability of diagnosis.

The Naïve Bayes model was accurate as well, but was a bit more difficult to interpret, due to the way the data had to be broken down for modeling.

Overall, both models proved to be accurate within a 95% confidence interval. For future predictions, I recommend continuing with the Logistic Regression model.

Bayes Process

For the Bayes process, the data set was pruned down to only the mean values of the 10 variables measured. There was some data missing in the data set, and there were also some bad data fields entered as “999s” in several of the fields. For each instance, the mean value of the variables were used as a replacement value to correct for the bad data.

Each variable was broken into categories for the model to be effective in it’s evaluation. The variables were reviewed for their variance, and each variable was broken into 5 components that roughly equaled their quantile distribution.

The model evaluated the data fairly well. It identified a 92% Sensitivity and a 95% Specificity with a 6.6% misclassification rate. Overall this was very good, but it was a bit difficult to decifer the specific contributions of the variables to the overall model.

Additionally, there could be more work done in breaking the categories into more intelligent factors, based on better domain knowledge from clinicians.

Logistic Regression

For this process, we began by breaking down the variables into categories just like the Bayes process. After working with the categories, and running the model, the data shaping process was thrown out. It was decided to use the continuous variables without going through the categorization process.

The data was also corrected for the previously mentioned bad data, and then the variables were modeled without any further refinement. The variable labeled “Texture\_Mean” was the most significant variable identified by the evaluation (0.33 coefficient).

An analysis of the probability estimate that would maximize the Sensitivity (96%) & Specificity (95%) of the model identified a 54% cutoff as ideal, although the default 50% standard was very nearly as good. At 54% evaluation, there was a very low 4.6% misclassification rate as well.