Final Project Report Team Python Charmer

During Phase I the team loaded, cleaned, summarized key statistics and metrics and provided graphics to allow visualization of breast cancer research data collected by the University of Wisconsin.

Through this process the team identified 16 missing values for Bare Nuclei Data. The average column value for Bare Nuclei was imputed for missing values to provide a method to perform consistent analysis. Summary statistics revealed that the Bare Nuclei Data had the highest standard deviation (3.60), followed by Normal Nuclei (3.054) and Uniformity of Cell Size (3.051). Clump thickness had the highest mean value (4.42), while Mitoses had the lowest mean value (1.59). Histograms revealed distributions that were skewed to the right for each attribute, with Clump Thickness having the most significant distributions across bins. Finally, there were roughly twice as many benign cells as malignant cells in the data set.

During Phase II the team implemented the KMeans algorithm unsupervised learning technique to conduct cluster pattern analysis to identify the optimal cluster size for the data. The Kmeans algorithm was employed to fit and predict data assignment to form clusters around the closest centroid. The team produced an inertia elbow chart to identify the optimal cluster size. In this data set, 2 was the optimal cluster size.

The team transformed the features of the data using Standard Scaler to derive better clustering. The team first standardized the data and then utilized KMeans algorithm to fit the data. The team produced a standard deviation bar plot and box plot for each attribute to visualize variation within the data.

At the end of this phase, the team produced a crosstab to visualize and authenticate clustering assignments. In this case, 226 of 238 (95%) malignant cells were assigned to cluster 0; 446 of 461 (97%) benign cells were assigned to cluster 1.

During Phase III the team once again implemented the KMeans algorithm with cluster size 2. In this phase, adjustments were made to the number of times initial centroids were chosen (10 to 20) and each run was increased from 300 to 500 iterations.

The following error rates and results were calculated:

**Benign error rate**:

Total number of points with predicted labels = 4 that correspond to CLASS = 2 / total number of predicted labels = 2.

The error rate for benign cells is 2.59%

**Malignant error rate**:

Total number of points with predicted labels = 2 that correspond to CLASS = 4 / total number of predicted labels = 4

The error rate for malignent cells is 7.66%

**Total error rate**:

Total number of datapoints with predicted labels (labels column) Not equal to actual class (CLASS column) / total number of datapoints.

The total error rate is 4.29%