**Spambase Summary and Conclusions**

Milestone 1:

By comparing the performance measures of Complete Feature Set, Feature Selection using Forward and Backward Search, and PCA, Feature Selection using Forward Search with kNN (k=5) is found to be the best classifier. Feature Selection using Forward Search has greater performance and more consistent values compared to other processes. For Forward Search, as k is increased, the performance values becomes greater. But when k=7, the values decreased. MLP, Decision Trees and Bayesian Network did not give greater results compared to kNN.

SVM Conclusion

Clustering

3. Explain the reasons for choosing such a number of clusters. You can apply the objective function (mentioned in the slides) to assess the quality of each cluster. You can mention the results of your previous experiments that may help justify for the choice of k clusters.

Starting with a number of 1 to 10 for the number of clusters did not show appropriate results. It only displayed 1 major cluster with a similar color. Choosing the k as 20 displayed several clusters that had different colors from each other. Also, even though there are different clusters, only 1 cluster was different from the others based from color shades. This solution gave a result that could be used for clustering two types of samples easily. The solution may become better if k is greater than 20.

4. Summary and Conclusion of the experiments

The k-means clustering is a process that groups samples together with similar characteristics from other clusters with different qualities. Based on the data recorded for Milestone 1, the top 2 features for PCA was referred to for the x and y axis for the Cluster Plot. The charts above have been generated by doing this process. Based on the data gathered, the clusters are more similar than dissimilar. Because of this issue, the class is difficult to distinguish.

SOM

Based on all the maps, it is difficult to distinguish what kind of samples are assigned to the parts of the map especially with no labels on it. Just like the clustering process, the maps are unorganized though the map with 1000 training rounds is more “clean” compared to the others. Based on the results, Using SOM with different training rounds such as 1000, 10000, and 15000 on the dataset did not organize the map. However, when compared to all the outputs, SOM with 1000 training rounds look more organized than the other training rounds. Even with unsupervised learning, the dataset was still unorganized due to similarities of the data in the features. The data can still be used for analysis, but it will not be as accurate for predictions.

**Diabetes Summary and Conclusions**

Milestone 1:

Clustering

There are only 2 possible values for the class, which are 1 and 0. These are boolean values to indicate of a patient has diabetes or not. This results to making k=2. Since the clusters are distinguishable from one another, this is a good reason to choose k as 2.

The k-means clustering is a process that groups samples together with similar characteristics from other clusters with different qualities. Based on the results, k-means clustering successfully grouped the samples into two clusters. However, the centroids of these 2 clusters are similar that at some point some samples cannot be differentiated from each other.

SOM

Based on the results, taking the SOM map of dataset 2 with 1000, 10000, and 15000 training rounds produced a varying number of nodes. With 1000 training rounds, the results displayed 17 distinct nodes. At 10000, 23 nodes were produced. And finally at 15000 training rounds, 28 nodes were retrieved. All of the information here were taken at jitter set to a minimum producing a 10x10 SOM map with clear, specific, and distinct nodes that allows the classification of each sample more precise and accurate. Although the actual relationship between each node were not discussed, this still shows a limited display of the SOM of diabetes.