**Data Set- Heart disease**

Using the heart disease dataset (**heart\_modified.csv** - found in the blackboard assignment page), show your understanding of classification techniques. The dataset has been modified to include additional noise in the data. Features have been added that have either been filled with random values or sampled from other features. Please choose two classifiers of your choice that have been covered in the module. The target feature for classification is called “class”, with a particular interest in classifying people that *do develop heart disease (class = 1).* Remember to use some method of training on a sub-sample of your data, while a sub sample is used for testing.

What will be assessed in this question is *not* what classifier achieves the best accuracy, but rather the justification of your choices and the interpretation of the results. As long as you have shown an understanding of what your classifiers are doing, achieving the highest accuracy is not the point of interest.

**Q1. For each classifier, please answer the following:**

1. Did you undertake any prepossessing? If so, why?
2. Run the classifier with default parameters. Print the output from Weka.
   1. How accurately can the classifier predict those that develop heart disease? What is in the output that signifies this?
   2. How many people are misclassified as developing heart disease? Where is this answer found in the output?
3. Plot and submit the ROC curve for the class that develops heart disease. What is another measure of accuracy commonly used? Please provide this.

**Q2. Now choose one classifier to further optimize.**

1. Why did you choose this classifier over the other?
2. Briefly explain how this classifier works from a theoretical point of view.
3. Try to optimize the classifier to achieve a higher accuracy (no matter how small) than first found. Remember that we have a particular focus on predicting those that develop heart disease.
   1. Were there any features that could be removed? How you could you determine this using Weka? Please print the output that helped you make this decision.
   2. Did changing the way data is sampled during training/testing affect the accuracy?
   3. What about some of the internal parameters specific to the classifier? Please explain how one of these parameters can affect accuracy.
4. In general, a classifier is only as good as the data it is trained on. Please comment on what is needed from training data to train a good classifier. How can utilizing classifiers help feed back into healthcare settings with regards to data collection?