**Computer Science 361 - Machine Learning**

Assignment 1 – Worth 5%

Due: 15th of March, 2020.

**Title: Signal or No Signal?**

**A). Which dataset you think is the real file**

I think the ObscureA is the real file and the ObscureB is randomised file of ObscureA.

**B). 2 paragraphs, one on each of the 2 different ways you found to tell which dataset had a signal in it.**

Result table and Confusion Matrix

ObscureA file showed more precision and TP, FP rate to determine the real given signals to show as much as data possible. However, ObscureB file showed only one TP, FP and Precision out of the whole data which shows that the file is randomised and cannot give the real information as the randomised file cannot produce signals. Also, the ObscureB file showed that the algorithm of the data was randomised and biased as it only showed numbers on one column.

Visualizing Data

The visual data of both file have shown similarity between the class distributions. However, when randomized both data, file A showed a significant difference in class distribution than B which determines that B may already have been randomized, therefore A showed more signification different to its’ original real file distribution.

**C). A paragraph discussing whether these techniques continue to work as the dataset gets smaller, please indicate the reasons for this behavior.**

The result table and confusion matrix showed the same behavior to the original file with larger data sets. As the result and confusion matrix is very statistical attribute which shows the possibilities.

Also, visualizing data showed similar behavior to the original data, but the distributions were not a significant as before. This reason is due to reduction in dataset and cluster of the data.

**D). A final paragraph discussing which method you feel is more reliable and why.**

I feel more reliable using the results and confusion matrix to determine the difference between real data and randomised data. This is because results of FT and NP shows how the data is analyzed to determine the data set and is more accurate with a confusion matrix. However, randomised data method relies on the possibilities more than the accurate data of give file.