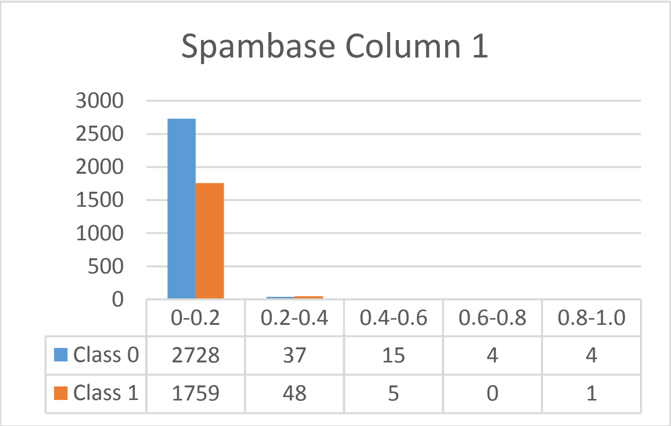
DMML Coursework 1.

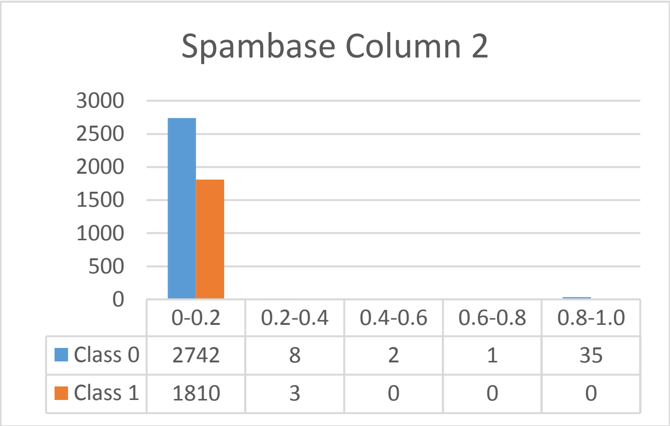
Tools

The tools that is used to create this report consist of Notepad++, Microsoft Excel and Microsoft Word. All data sets are prepared by simple “Awk” scripts. The scripts also do min-max normalization, calculate TPR-FPR and create data for Microsoft Excel to create all histograms. All “Awk” scripts are created using Notepad++.

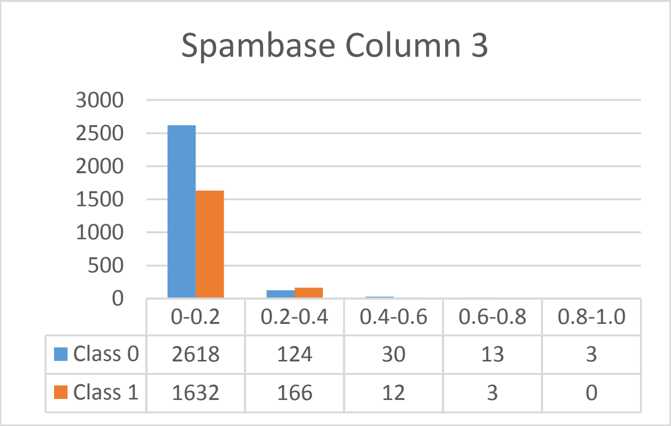
Histograms of Spambase Data Set.



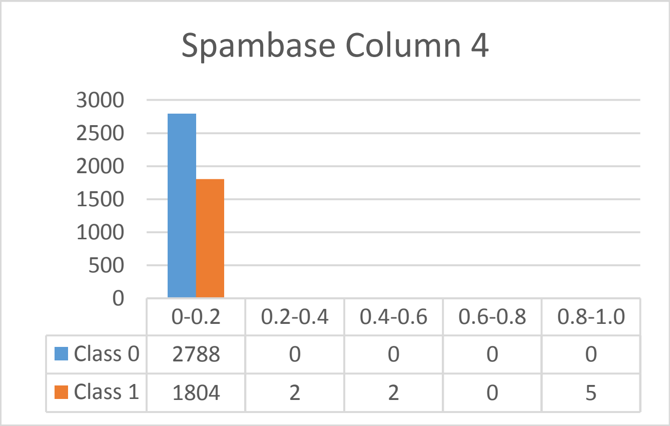
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Class 0	2728	37	15	4	4
Class 1	1759	48	5	0	1



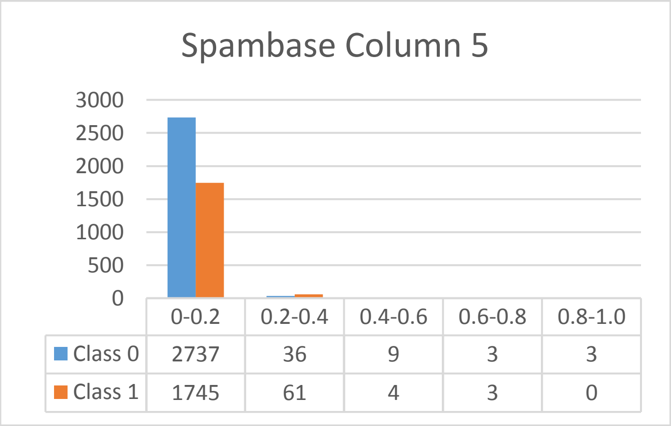
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Class 0	2742	8	2	1	35
Class 1	1810	3	0	0	0



	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Class 0	2618	124	30	13	3
Class 1	1632	166	12	3	0



	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Class 0	2788	0	0	0	0
Class 1	1804	2	2	0	5

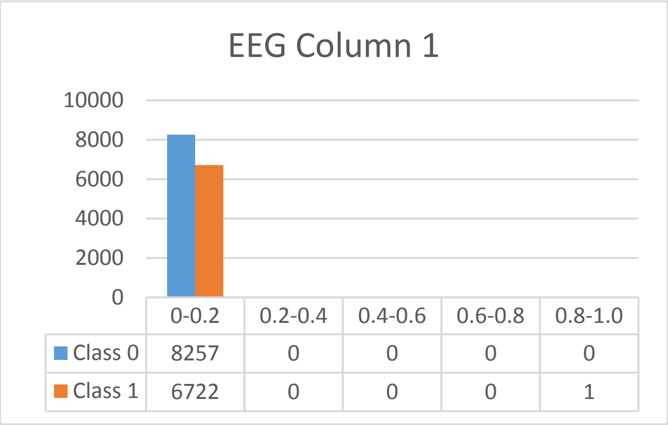


	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Class 0	2737	36	9	3	3
Class 1	1745	61	4	3	0

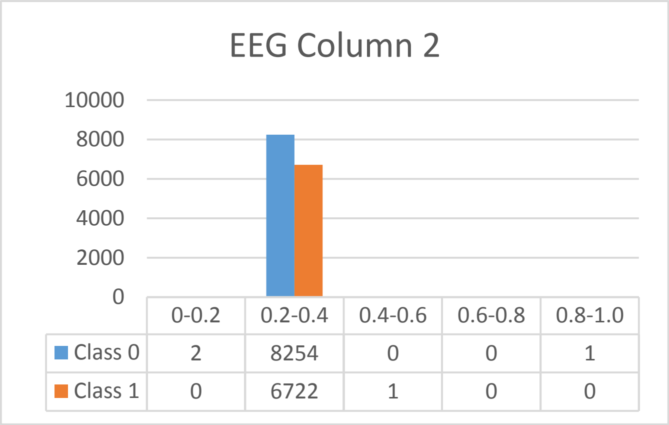
Spambase Data Set in detail.

A common bin for all histograms is in the first range from 0 to 0.2. In the first histogram, the rage of data that exceed a common bin is around 2.2% of 2788 total records of class 0 and 3% of 1813 total records of class 1. The second histogram, the rage of data that exceed a common bin is around 1.7% of class 0 and 0.2% of class 1. In the third histogram, the rage of data that exceed a common bin is around 6%% of class 0 and 10% of class 1. The forth histogram have smallest difference, class 0 at 0% and class 1 at 0.5%. In the fifth histogram, the percentages of email that exceed a common bin are 2% for class 0 and 4% for class 1. As you can see in histograms, the shape of are quite similar so it is hard to tell which column are better to use in classification. After calculate exceeding percentage for all histogram the result showed that column 3 is the word that is used the most in both two class of email and column 4 is smallest used for both type. After considered the gap between classes column 3 and column 5 may be the best two for classification.

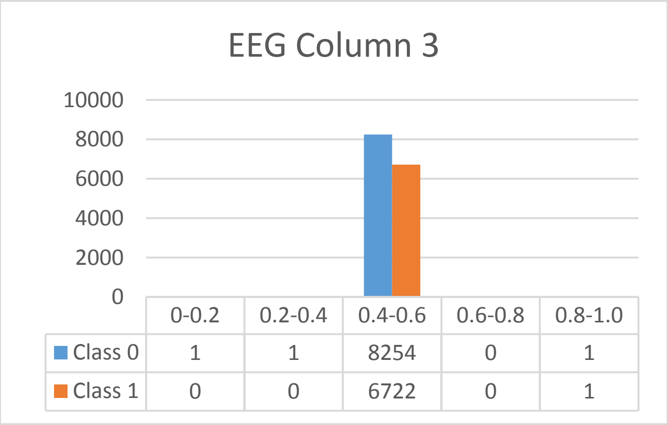
Histograms of EEG Eye State Data Set.



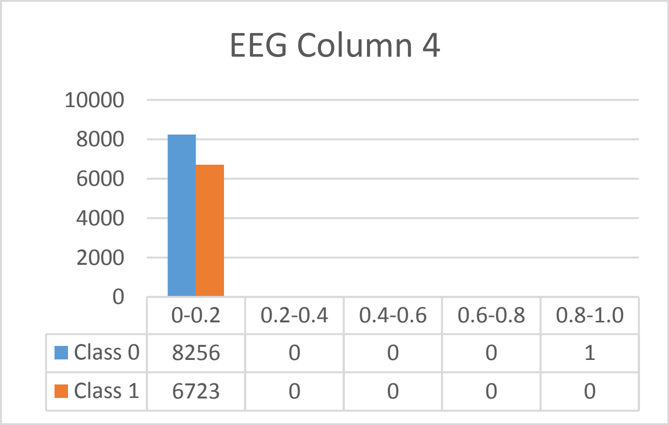
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	8257	0	0	0	0
🟧 Class 1	6722	0	0	0	1



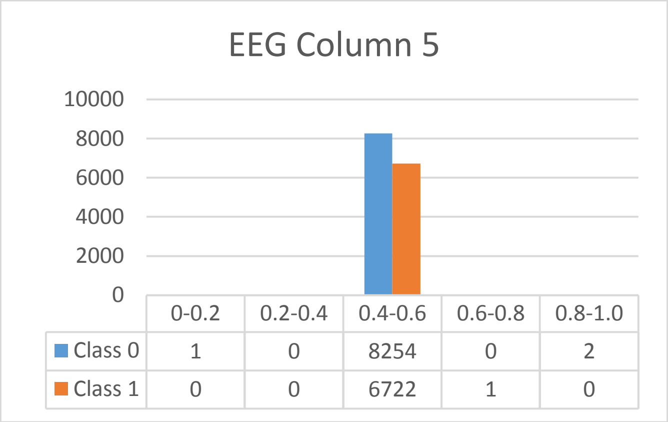
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	2	8254	0	0	1
🟧 Class 1	0	6722	1	0	0



	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	1	1	8254	0	1
🟧 Class 1	0	0	6722	0	1



	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	8256	0	0	0	1
🟧 Class 1	6723	0	0	0	0

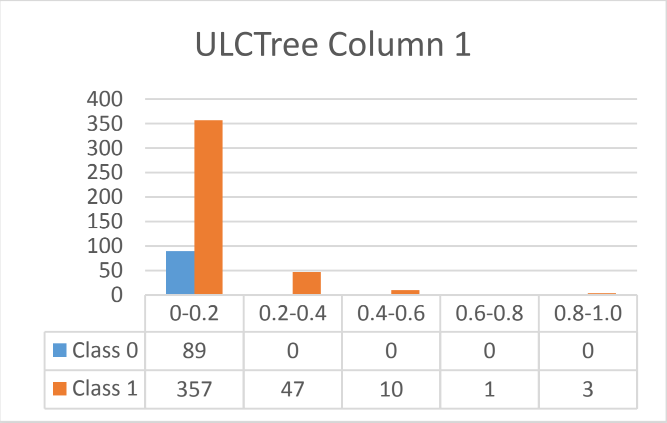


	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	1	0	8254	0	2
🟧 Class 1	0	0	6722	1	0

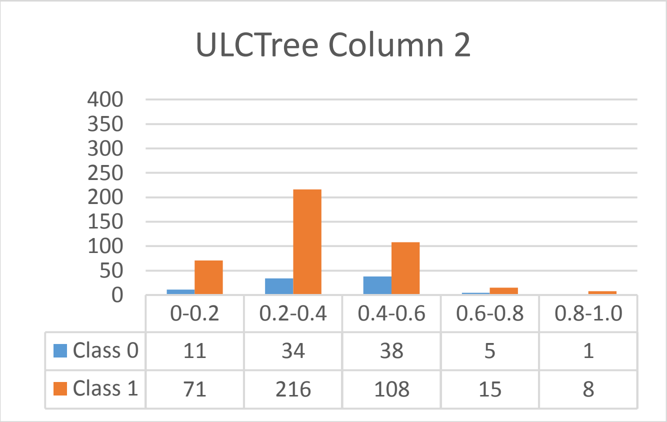
EEG Eye State Data Set in Detail.

In this data set, class 0 represent the state of an eye-open and class 1 eye-closed. An overall distribution of 5 histograms above, class 0 and class 1 tend to stay mostly in the same rage. The first histogram show that there is almost spread in distribution just one outlier which is mean all data for each classes are quite close to each other. The second histogram a common bin shifted from 0-0.2 range to 0.2-0.4 and most of the data of this column are within this range. The third histogram showed quite the same, the data shifted to bin 0.4-06 this time. The forth Histogram are almost the same as the first one as it filled the same bin and have 1 outlier in different class. The last histogram filled 0.4-0.6 bin as the third one but the distribution showed that it slightly better distribution. As you can see, this time class 0 and 1 have very similar distributions but each histograms have different outlier value that can use to discriminate the classes. Column 2 and column 5 seem to show more differences that other three. For class 0 there are some data in the first and the last bin but none in class 1. Also class 1 have data in the fourth bin but not in class 0. The other 3 columns have only 1 or 2 differences.

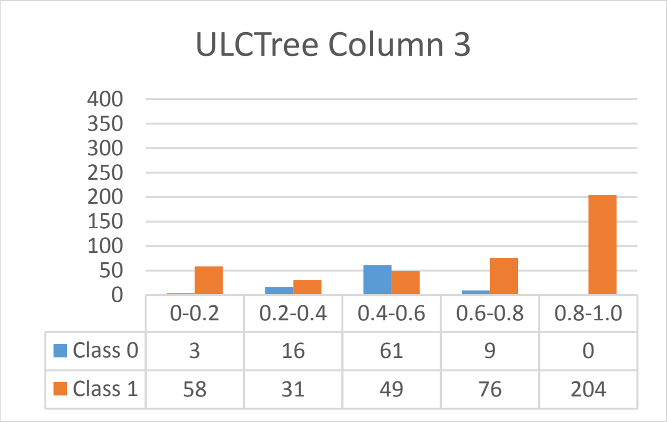
Histograms of Urban Land Cover (ULCTree) Data Set.


ULCTree Column 1

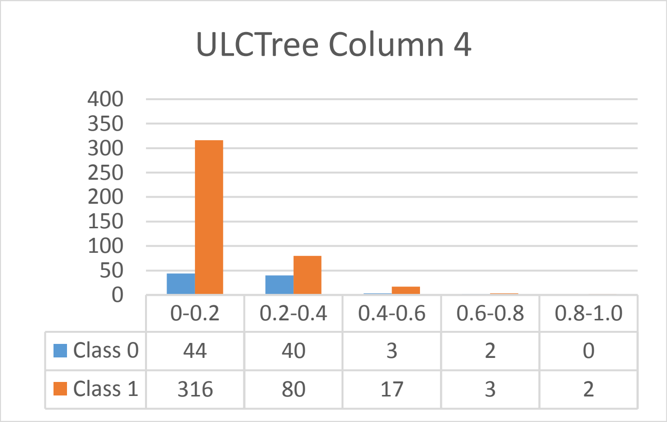
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	89	0	0	0	0
🟧 Class 1	357	47	10	1	3


ULCTree Column 2

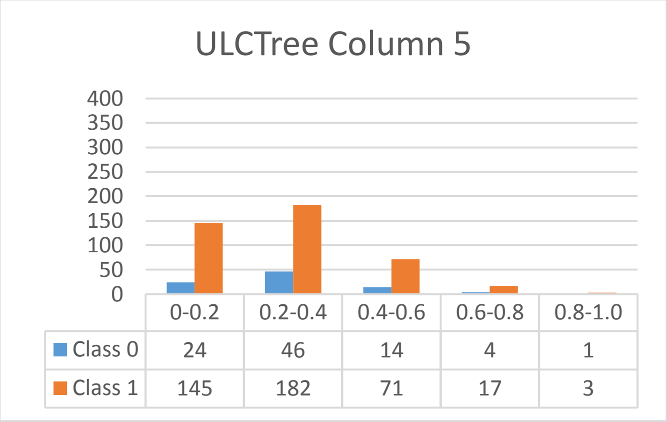
	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	11	34	38	5	1
🟧 Class 1	71	216	108	15	8


ULCTree Column 3

	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	3	16	61	9	0
🟧 Class 1	58	31	49	76	204


ULCTree Column 4

	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	44	40	3	2	0
🟧 Class 1	316	80	17	3	2


ULCTree Column 5

	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
🟦 Class 0	24	46	14	4	1
🟧 Class 1	145	182	71	17	3

Urban Land Cover (ULCTree) Data Set in Detail

The distributions in the third data set are clearly provide more useful information to discriminate the class than the first two data set. Class 0 is mean tree on the satellite images and class 1 mean other things such as building, car or asphalt. The distribution of class 0 in the first histogram are only in the first bin (0-0.2), class 1 also stack high up in the first bin but it also skew to right. The second histogram are quite well distributed across all bin for both classes but the difference of ratio between tree and other surface are quite high so it is more likely to be class 1 in every bins. The third histogram clearly showed that this column is good to discriminate the classes. If the data is in range 0.4-0.6, it is more likely to be class 0 and if data is in range 0.8-1.0 it is more likely to be other surface. Most of class 1 in the fourth histogram are in the first bin. The second bin showed high number of class 0 is in this bin (less than class 1 but it have a good ratio compare to other column). The fifth histogram is similar to the second one, it is more likely to always be class 1. With all of information the best two are column 3 and Column 4.

1NN Table of accuracy

Data set	Accuracy
ULCTree	89.9408%
EEG	83.3244%
Spambase	91.2193%
Reduce ULCTree	83.8264%
Reduced EEG	55.7677%9
Reduced Spambase	42.6646%

A result of the original tree data set clearly show better accuracy than the reduce data set. Especially in Reduced EEG and Reduced Spambase data set. This may be the result from the rage of the data in those two data set from each column are very close to each other. And 1NN predict the class by using sum of the square difference between each record. Reducing a number of column in this kind of data may reduce a range of data that might help discriminate the classes significantly. The accuracy of 1NN of reduced ULC data set seem to have only little reduce of accuracy because the data range of ULC are clearly show better distribution between the classes than the other 2 data set.

Part2: TPR and FPR

ULC1

Classes	TPR	FPR
Class 0	0.764045	0.0717703
Class 1	0.626506	0.0589623
Class 2	0.4	0.026694
Class 3	0.827957	0.0748792
Class 4	0.733333	0.017316
Class 5	0.618557	0.0536585
Class 6	0.809524	0.0102881
Class 7	0.714286	0.0020284
Class 8	0.777778	0.025974

ULCtree2

Classes	TPR	FPR
Class 0	0.764045	0.0717703
Class 1	0.92823	0.235955

ULCcar2

Classes	TPR	FPR
Class 0	0.809524	0.0102881
Class 1	0.989712	0.190476

Discussion

Compare to the original ULC data, the TPR of ULC tend to be higher. It is quite obvious that by grouping most of the data into one class covered more range of data. When predicting a class value using 1NN, it is more likely to be correct in the class that cover more range. But with a high number of positive it also increase more chance that 1NN will produce incorrect class. As you can see FPR of 2 big classes in ULCTree and ULCCar are very high (24% and 19%). All of the FPR in the original data set are very low the highest one is 7.4%. As the classes in the original ULC are separate around range (Landscape), TPR are clearly lower than other 2. The lower TPR of the original is acceptable (Still more than 70% accuracy in most classes) but the FPR of the other 2 may not be a good thing because 24% of the predicted positive class are wrong.