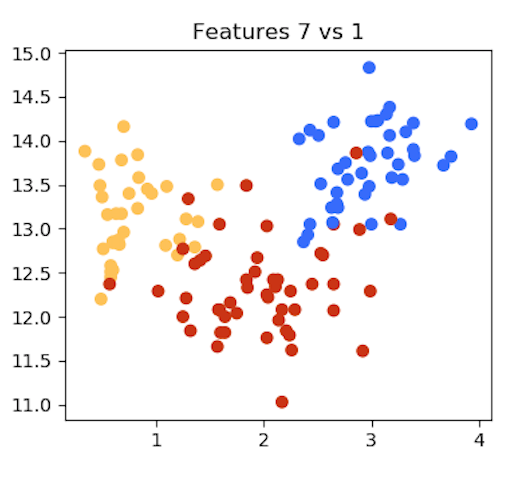
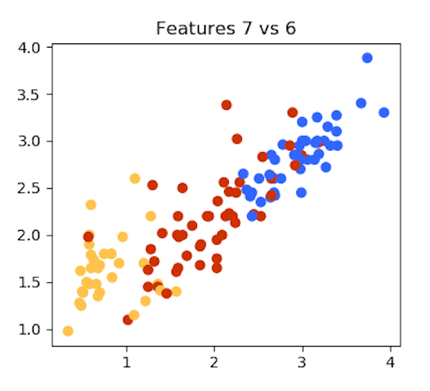
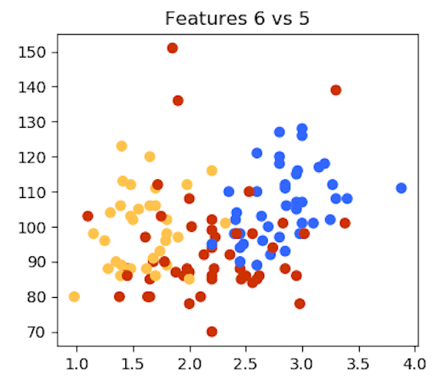
**SPS CW2 Report**

Feature Selection

When selecting the pair of features that we would use to train our classifier, there were a few features we were looking for.

Firstly we looked for feature pairs that gave distinct clusters of each group. Primarily this was focused on ensuring classes 1 (Blue) and 3 (Yellow) were linearly separable, as looking at the scatter plots it was quite apparent that class 2 (Red) had a decent amount of couldn’t have a linear separation from the other two classes regardless of which feature pairs were selected.

Secondly, as can be seen in the plot of our selected features, there is no strong correlation between them, while clusters are still present. This means that one feature cannot be simply predicted using the other, resulting in our classifier having more information to work with.

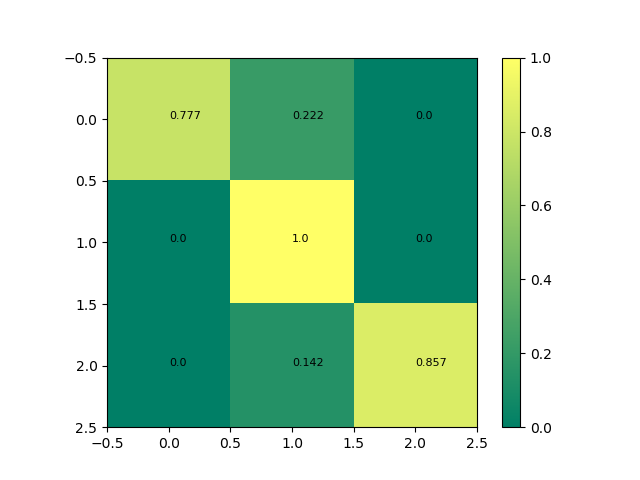
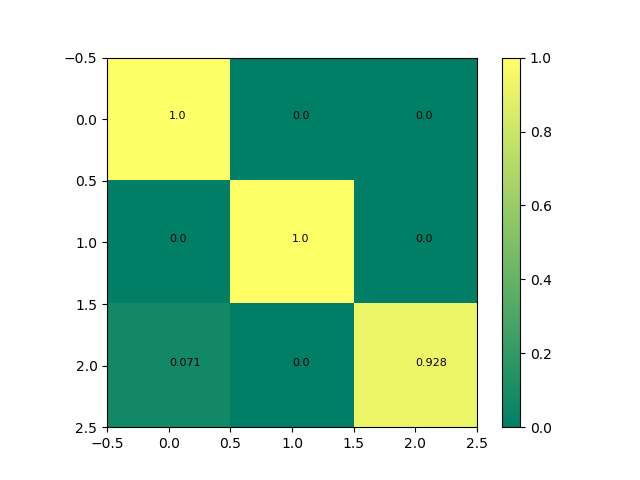


Shown to the side features 7 and 6 were passed over in favour of features 7 and 1 due to both pairs having a similar degree of clustering and separated groups, but with 7 and 1 being less clearly correlated with each other. On the other hand pairs such as features 6 and 5 weren’t considered as while they do also show a degree of clustering the lack of any clear separation between classes would result in any classifier we created being inaccurate for identifying classes 1 and 3, and almost useless for identifying class 2.

KNN Classifier

Using the features we selected, our implementation of the KNN classifier has accuracies that range from 88% to 100% and appear to generally increase as K grows, with the exception of a small drop in accuracy when going from K = 5 to K = 6. This drop in accuracy could be for a variety of reasons, but is most likely due to an edge case from any of the classes that perhaps didn’t have a single modal neighbour class for K=5 having its extra nearest neighbour be from the incorrect class.

The trend across K’s 1 to 7 would seem to indicate that with the feature pair we have selected, that K’s closer to 7 provide the best performance, this is probably due to the fact that while we have looked for features with distinct clusters, we could not avoid each class having a small number of outlier. This would result in small K’s occasionally giving the wrong results for test values in cases where the data point happens to be closer to an outlier from a different class. Larger K’s would avoid this as it would have a wider range of points to query, with the outlier being ignored in favour of the majority from the clusters class.

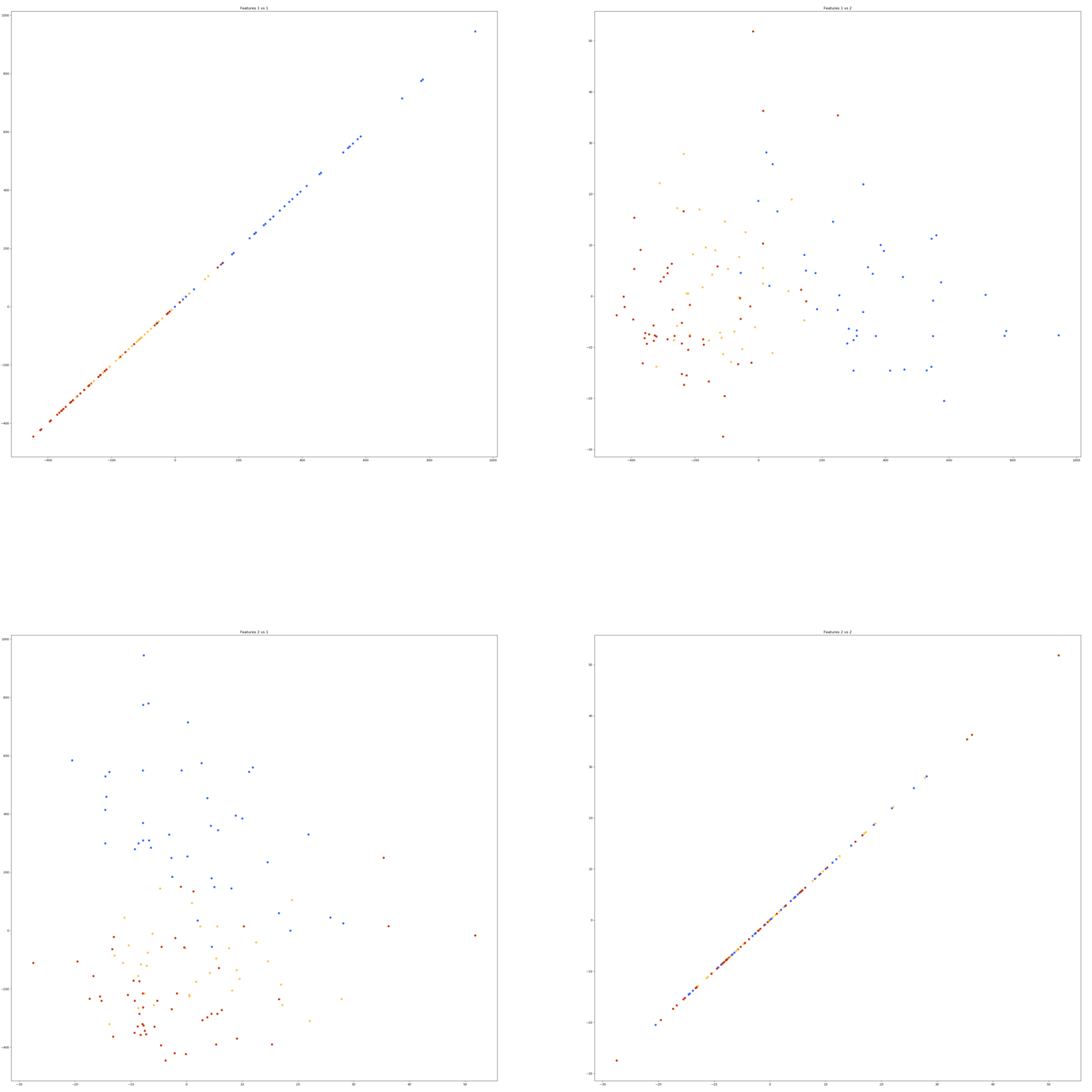


Another point of interest can be seen in how the confusion matrix develops as K grows.

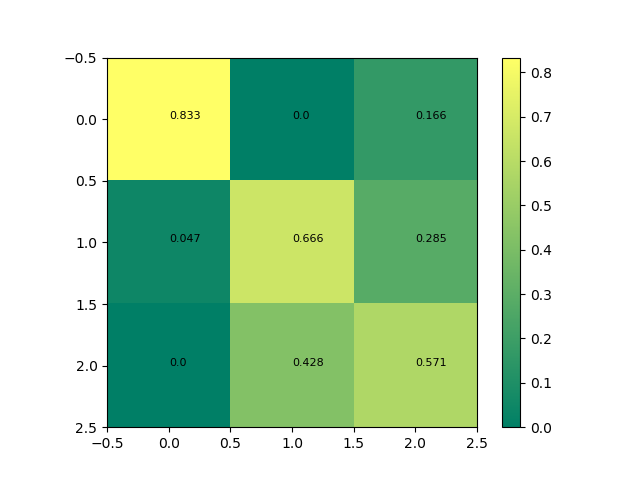
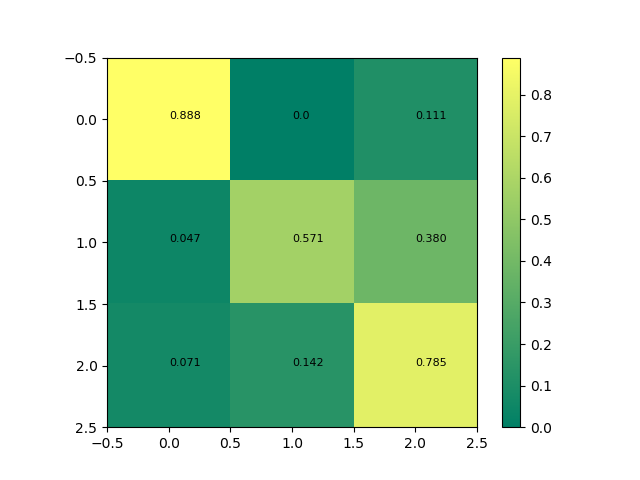
For K = 1, the only errors to occur are between Class 2 and the other classes, which can be attributed to the feature pair we selected causing class 2 to have some spread into the class 1 and class 3 clusters. This changes when K = 6 as the wider range of nearest neighbours being checked allows for some test values to be mistake for class 1 when they belong to class 3. This also agrees with what was stated above, as the possible reason for why the accuracy of the classifier decreases from K = 5 to K = 6.

K = 6

K = 1

KNN PCA Classifier

Shown to the side is a scatter plot of the features from our training data after being orthogonally projected onto a 2-dimensional space. As can be seen in this plot, Classes 2 (Red), and 3 (Yellow) have main clusters, with a few outliers, while Class 1 (Blue) doesn’t really have a main cluster of data at all. This is in contrast to our manually selected features which showed strong clustering for all classes. Another major difference between the PCA transformed data and the data obtained from our feature selection is the PCA data’s complete lack of separation between the data clusters for class 2 and 3.



The effect of the data clusters for classes 2 and 3 having almost no separation from each other has had a large effect on the general accuracy of the KNN classifier when provided the transformed training data and test data, having accuracies between 69 and 82% for all K’s 1 to 7 with no clear trend showing an increase or decrease of accuracy as K grows. However, by looking at the confusion matrices of our results, the reason behind this loss in accuracy is shown to be primarily due to the classifiers difficulty differentiating between data that belongs to class 2 and class 3. The scatter plot shows this is because of the clusters for classes 2 and 3 not being separated, which means any data from these classes could be surrounded by data from the other.

K = 7

K = 3