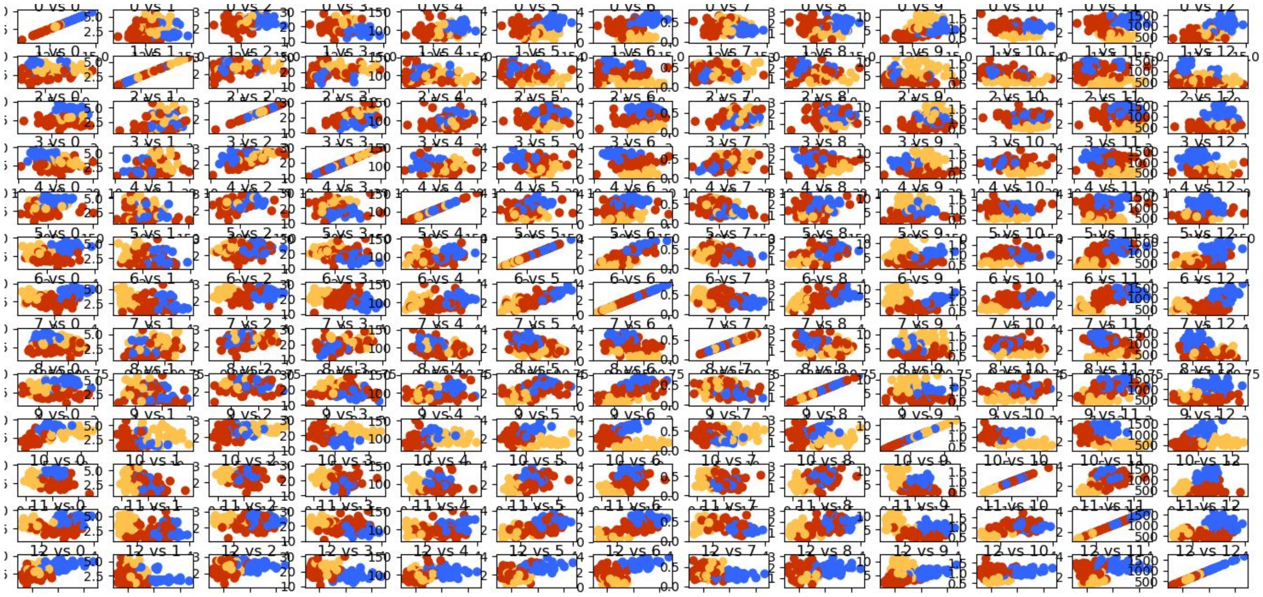
Andrei Alexandru : aa17956
Matei Culianu : mc16337

CW2: A Memory of Wine

Feature Selection

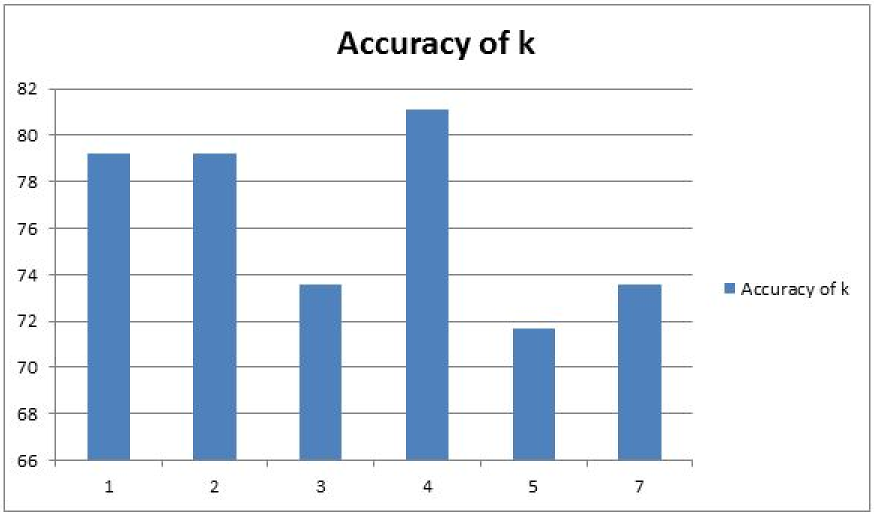


To select two features for our classifier, we did a scatter plot of all the pairwise combinations of the 13 features. After analyzing and discussing the results between us, we decided to choose features 10 and 12. Although there are multiple good pairs to choose from (e.g 9 vs 12 or 6 vs 9), we are confident these two features are able to separate and group the data in different classes in the best way possible.

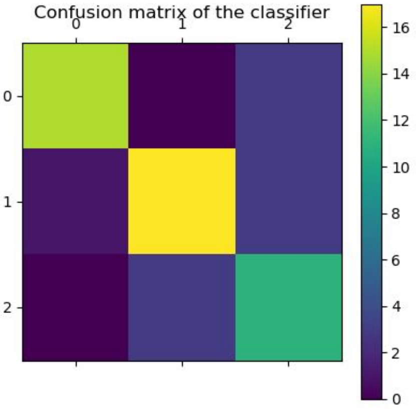
Alternative Classifier

As an alternative classifier, we chose and implemented the Naive Bayes classifier. We chose this classifier specifically because it was one that we both understood how it works and we were both confident we could implement it. Compared to our knn classifier, Naive Bayes delivered better results, with an accuracy of 84,90, the largest accuracy we recorded overall across our project. The main reason for this is the features we selected. Due to the fact our features don't separate the data well enough, it creates problems for the knn algorithm, which predicts classes by looking at a point's neighbours. Naive Bayes uses a Gaussian Probability Density Function to predict the class, which works better with our features. While separating the data in groups of classes helps both classifier work better, because the groups are too close to each other, the knn classifier has a great disadvantage.

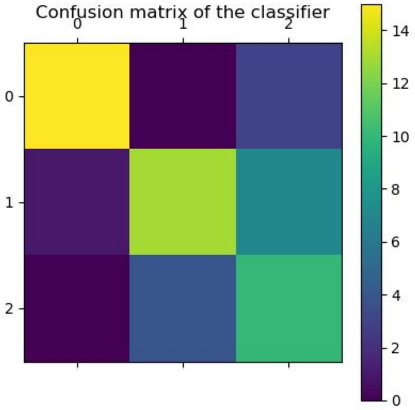
K-Nearest Neighbours



After implemented our knn classifier, we calculated the accuracy for all the given k values. The resulted are depicted in the figure above.According to k, the accuracy varies from 71% to 81%. The reason that a larger k doesn’t necessary means a better accuracy is because it makes the classification boundaries less distinct, which leads to the classifier confusion classes.



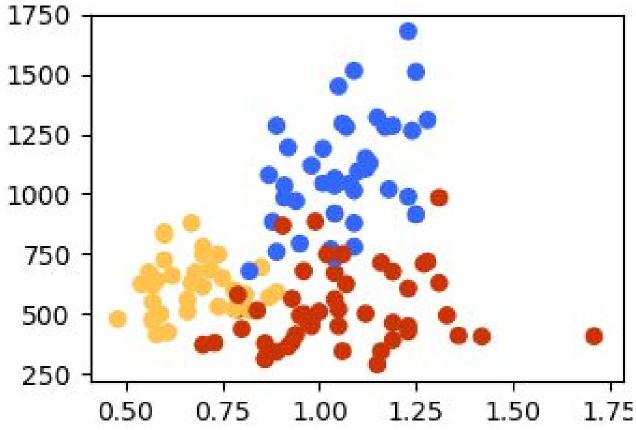
K = 4

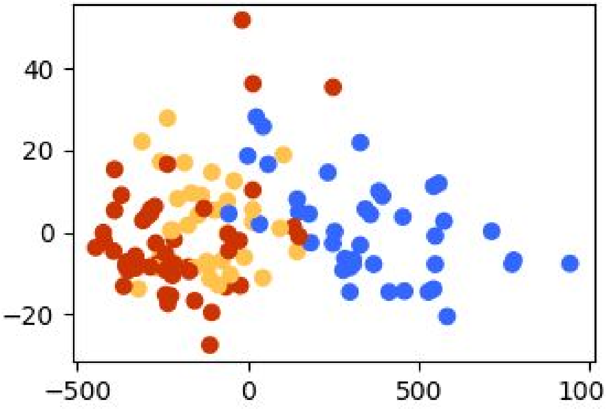


K = 5

We plotted the confusion matrices for k =4, which gave us the best accuracy, and k = 5, which gave us the worst accuracy. Judging from these figures, we can conclude that our knn classifier has problems with predicting the third class wrongly. This may be caused by the fact that, in our selected features scatter plots, the first and second classes are grouped in a more compact way, but are both closer to the third class group, which spans over a bigger surface. Because the groups are too close to each other, the classifier confuses the points between the classes.

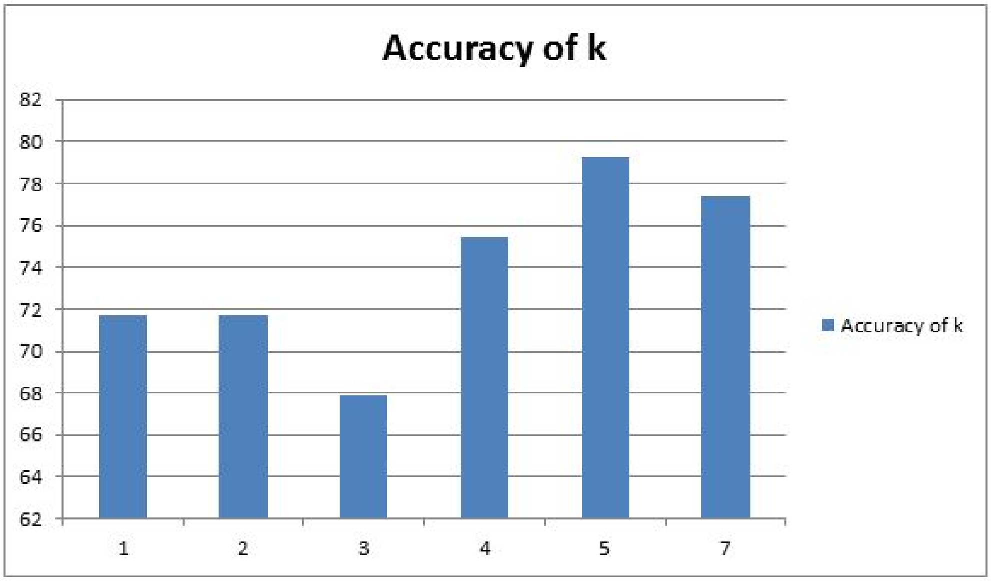
Principal Component Analysis (PCA)





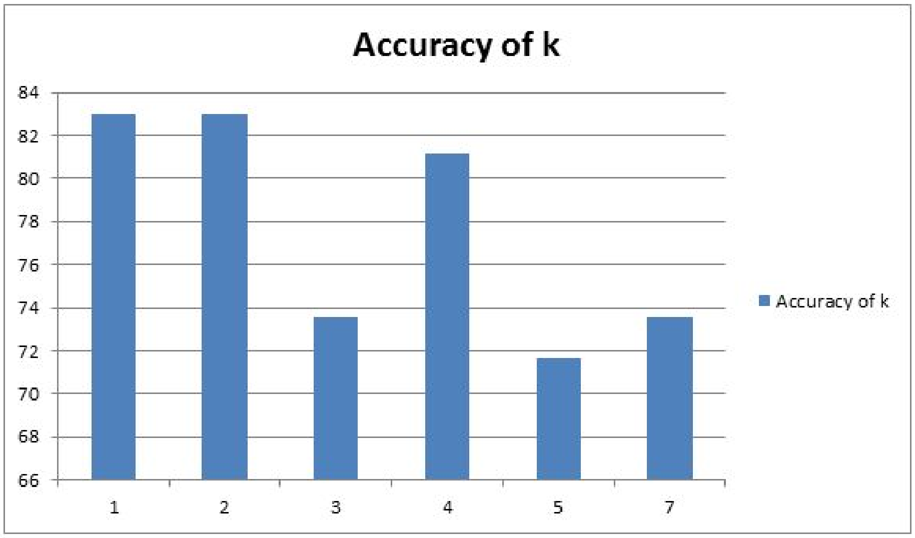
Selected features **PCA**

Comparing the 2 plots, we can determine that the PCA one did a worse job at separating and grouping the 3 classes. This could be caused by not normalizing the data. One of the attributes may be an order of magnitude higher than others, which causes PCA to assign it the highest amount of variance, leading to a bad result. However normalizing results in spreading the influence across many more principal components. In others words, more PCs are required to explain the same amount of variance in data. In the end, we decided not to normalise the data.

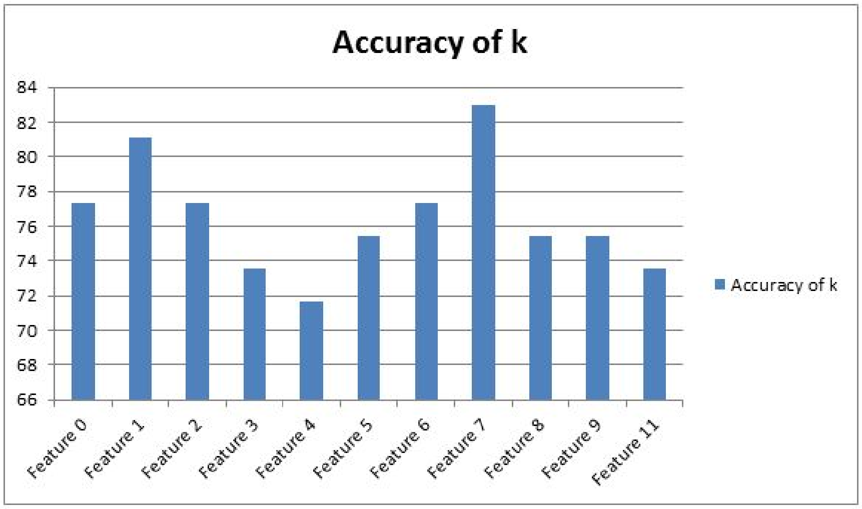


Compared with knn with our selected features, with PCA it brought a smaller accuracy overall. In addition, it had a larger accuracy for smaller values of k, but smaller accuracy for larger values of k, which could be caused what we previously discussed above.

Use Three Features



To manually select the third feature, we ran the program with each of the remaining 11 features and chose the one which gave us the largest accuracy. In the picture above, we represented the accuracies for each k with the features 10, 12 and 7. Compared to a classifier with two features, this classifier seemed to improve the accuracy for smaller k values, but it gave the same result for the larger values. Furthermore, adding an extra dimension for the classification class gave a larger accuracy, which is a great benefit for our classifier.



The image above represents the accuracy of the selected two values with each of the other 11 values, when k = 1.