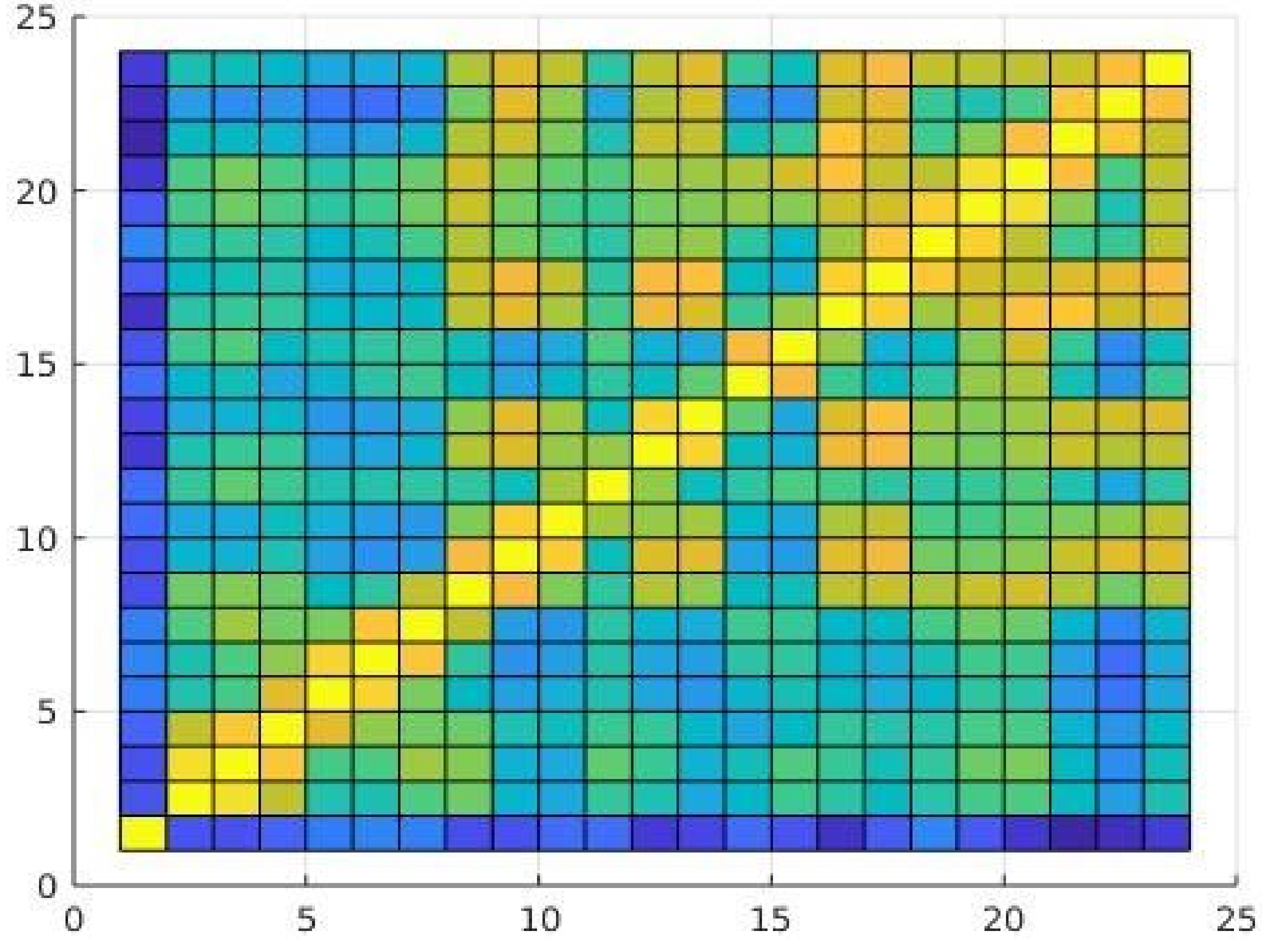
Task 1

Task1.2:

In this task I investigate the correlation matrix and describe the findings from that correlation matrix of my dataset for the coursework. Firstly, I will explain what weak correlation and strong correlation are. Weak correlation is when the value of correlation between 2 vectors is near 0. This is because correlation gives the value of the cosine between 2 vectors and if 2 vectors are perpendicular or nearly so, the value of the cosine is near 0. The opposite is when we have a strong correlation. We have 2 types of strong correlations : negative and positive correlations which depend on the direction of the vectors. From my correlation matrix, I found out that feature 1 is negatively correlated with all of the other features. To add to that, it is negative but closer to 0 than to 1 implying that feature 1 is weakly correlated to the other features. Contrary to that, the other features have positive and also strong correlation with most of the other features possibly implying that the other features are correlated and are dependent on each other. This would imply that a classification which utilises the naive Bayes would be inefficient since in naive Bayes we assume each of the features is independent of the others which seems to not be the case. Below is my graph showing the correlations between each pair of features.

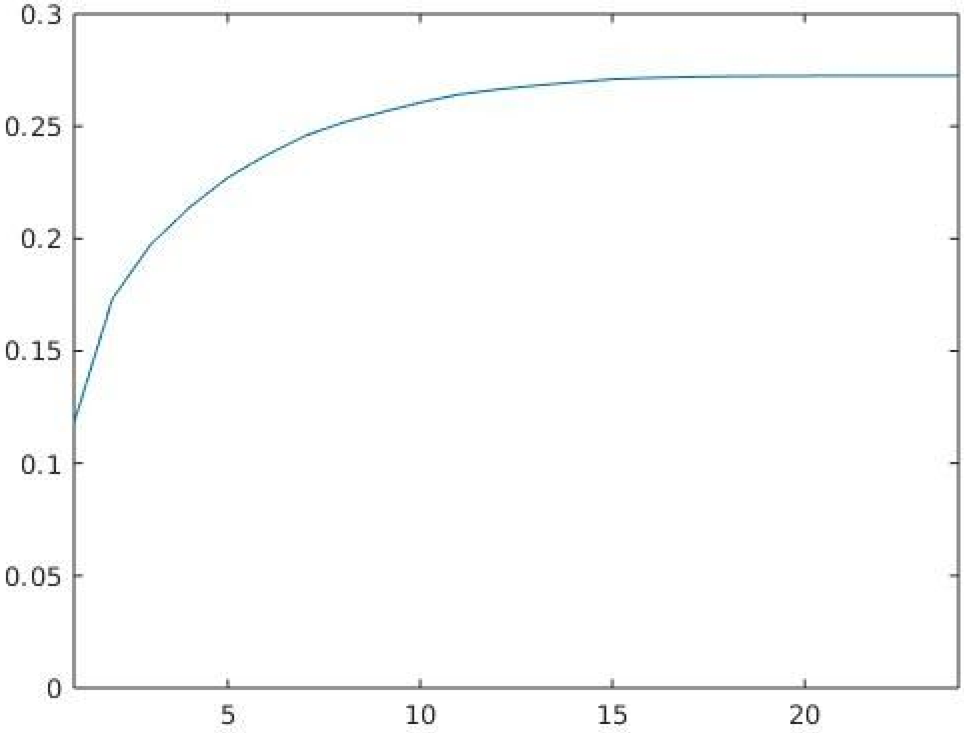


The graph shows how the feature in x-axis is correlated to the feature in y-axis. Here the blue color shows the negative correlation. The light blue shows a correlation near 0 and as the colors go from blue to green to yellow the correlation gets bigger and closer to 1 as it is shown in the diagonal of the correlation matrix in which all of the elements are equal to 1.From here we can see that most of the elements are strongly correlated and this further supports my claim of the features being dependent on each other.

Task1.3:

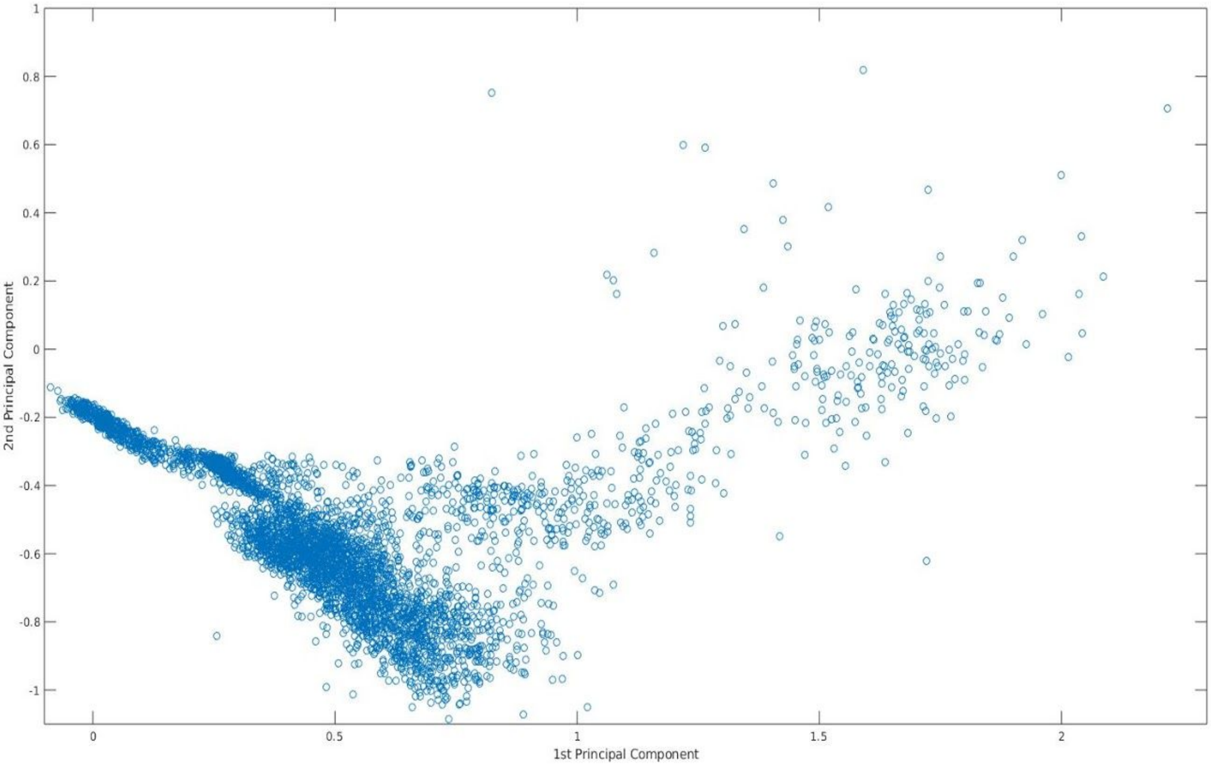
Part b:

This is the graph of cumulative variances



Part c:

This is the plot of the PCA 2D taking only the first two principal components

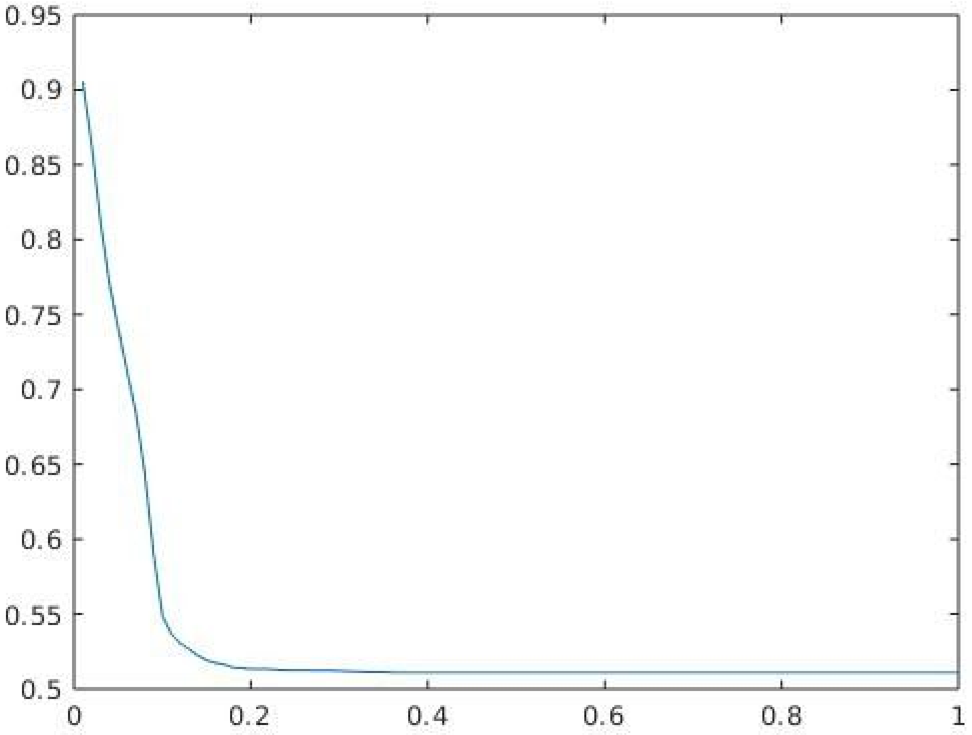


Task 1.4:

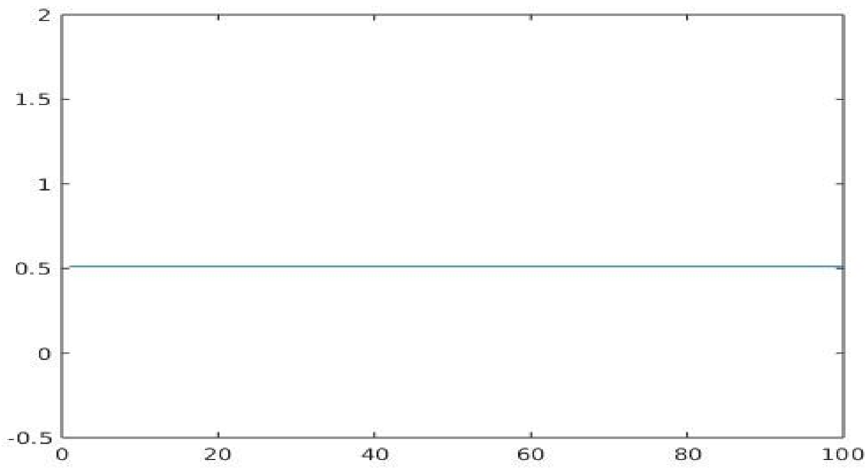
In task 1.4 I run the classification with epsilon 0.01 and kFolds=5.
The accuracies for each type of covariance matrix are listed below:
Covkind =1 : accuracy = 90.52%
Covkind=2 : accuracy = 9.22% I suspect this is low because the features are dependent
Covkind=3 : accuracy = 88.01%

Task1.5:

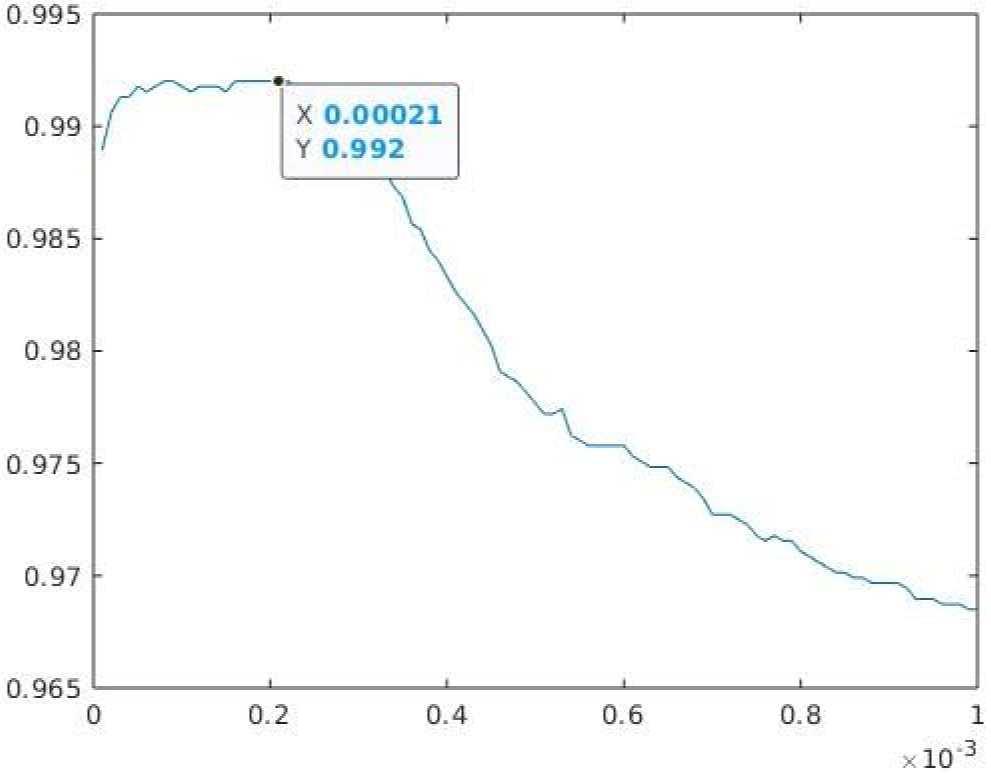
I recorded the accuracies for values of epsilon from 0.01 to 1 using a full covariance matrix and kFolds=5. I noticed that the results for the accuracies were dropping dramatically from 0.01 to approximately 0.2 where they were getting to flatten to an accuracy of 51.09%. Meanwhile for 0.01 I got an accuracy of 90.52% so this is a huge drop in accuracy. The graph for that is shown below:



The x-axis are the epsilon values and the y-axis are the accuracies recorded with that particular epsilon. I also tried values bigger than 1 and they resulted in a straight line as shown below:



Lastly, I tried values of epsilon between 10^-5 and 10^-3. Surprisingly, I got a maximum between 10^-3 and 10^-5 where the accuracy reaches 99.2% and then it drops down. The graph is shown below:



0.995
0.99
0.985
0.98
0.975
0.97
0.965
0
0.2
0.4
0.6
0.8
1
×10^-3
X 0.00021
Y 0.992