CS 7641- Machine Learning

Project 3 – Unsupervised Learning and Dimensionality Reduction

Madhu Mohan
MMOHAN8

Abstract

This report analyzes clustering and dimensionality reduction techniques. The 2 clustering algorithms analyzed were K-means and Expectation Maximization. The 4 dimensionality reduction algorithms analyzed were Principal Component Analysis (PCA), Independent Component Analysis (ICA), Random Projection (RP) and Recursive Feature Elimination (RFE). This report has 3 parts:

1. Clustering Algorithms Analysis
2. Dimensionality Reduction and Clustering Analysis
3. Neural Network Analysis with Dimensionality Reduction and Clustering algorithms

Datasets

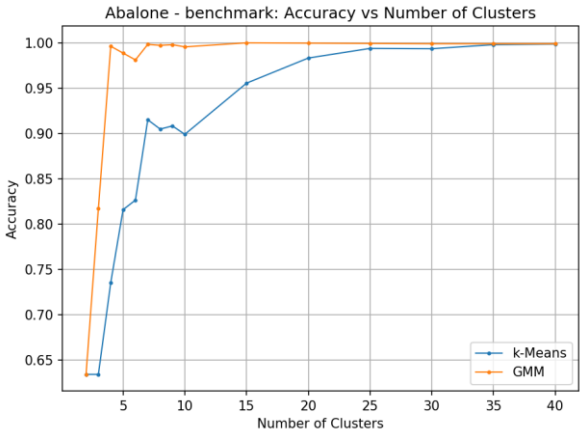
I have used Abalone and Wine Quality datasets in this analysis. Wine Quality dataset is from my assignment 1 and will be used to compare the results in Neural Network Algorithm with and without Dimensionality Reduction and Clustering techniques.

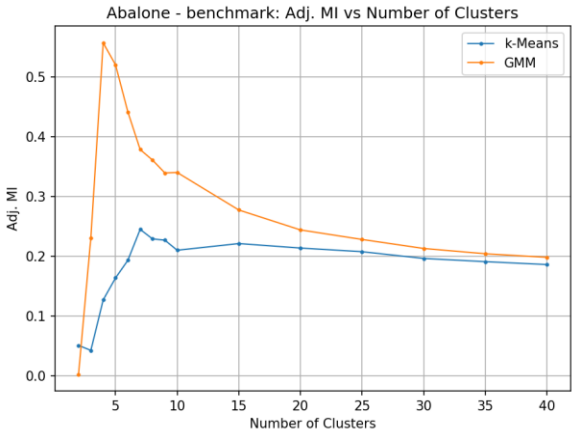
Part 1: Clustering Algorithms Analysis

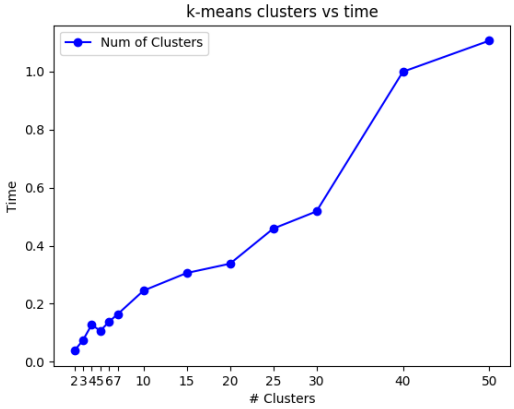
To analyze the K-means and Expectation Maximization algorithms, I measured 3 values: Accuracy, Adjusted Mutual Information Score and Time taken for increasing number of clusters for both algorithms.

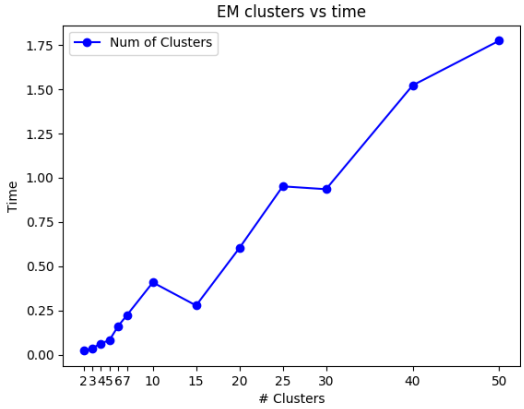
Following are the results that I got:

<u>Abalone:</u>



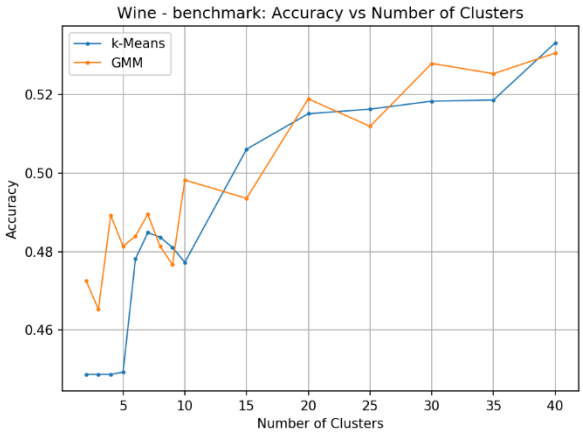


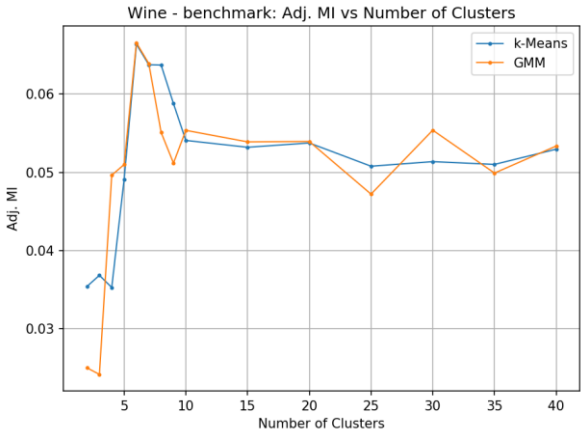


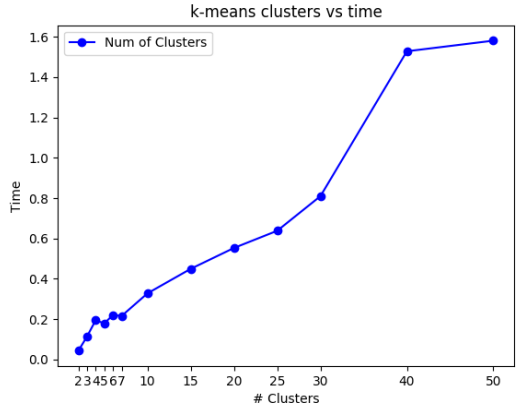


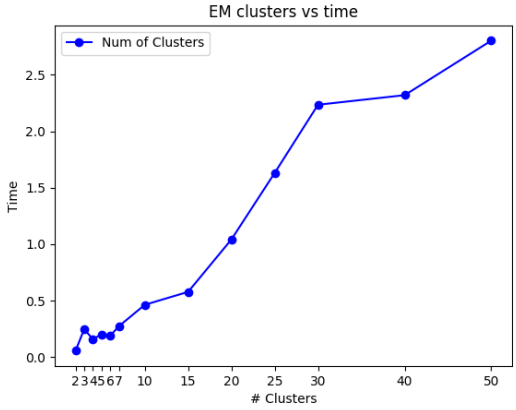
In the case of Abalone data, EM was doing significantly better. The accuracy score peaked at around 4 clusters in EM where as the score peaked at around 25 clusters for K-Means. The same observation can be seen with the AMI scores also. With respect to time taken, K-Means ran about 50% faster than EM. Despite this, I believe the performance of EM was significantly better for this dataset.

<u>Wine Quality:</u>









In the Wine data, I did not find any significant difference in accuracy between the 2 algorithms. K-means seems to be performing slightly better with a gradual increase in accuracy as the number of clusters increased. For this data set too, K-means was faster. Considering these, K-means seem to fit better for this data set with comparable accuracy and significant performance improvement.

Part 2: Dimensionality Reduction and Clustering Analysis

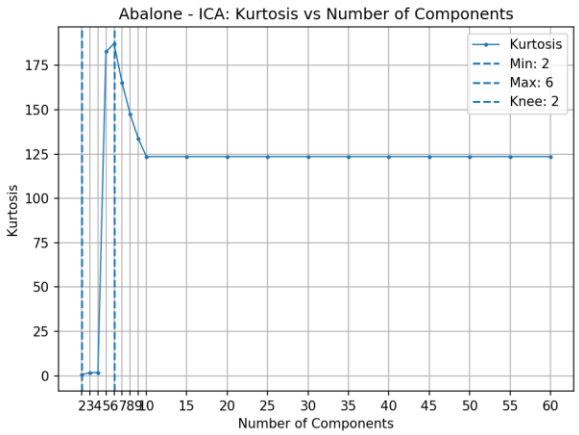
Dimensionality Reduction techniques help in transforming the input data to fewer dimensions. It helps in identifying features that are relevant/useful and can help in improving the accuracy of the learning algorithm downstream.

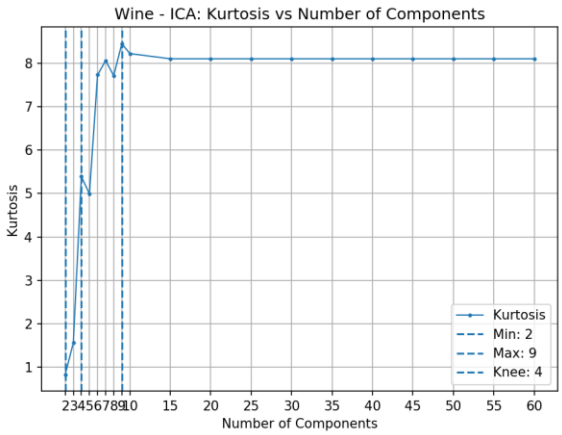
Independent Component Analysis:

ICA is a method to separate a multivariate signal into its subcomponents. It does this by maximizing independence of the transformed feature set.

Dimensionality Reduction:

To analyze ICA, I measured kurtosis for various components. Kurtosis measures the degree of non-gaussian. Based on the below observation, we should have a significant improvement in accuracy at component = 2 for Abalone data and at component = 10 for Wine data, as those are the points with maximum kurtosis.

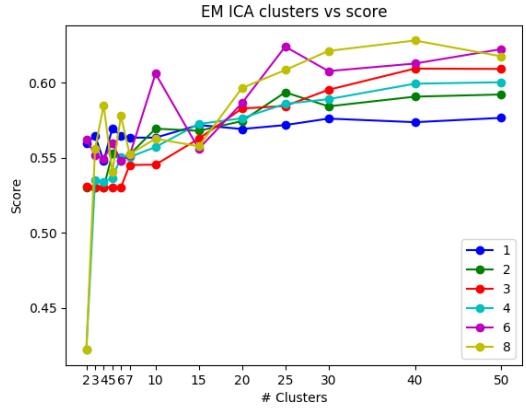


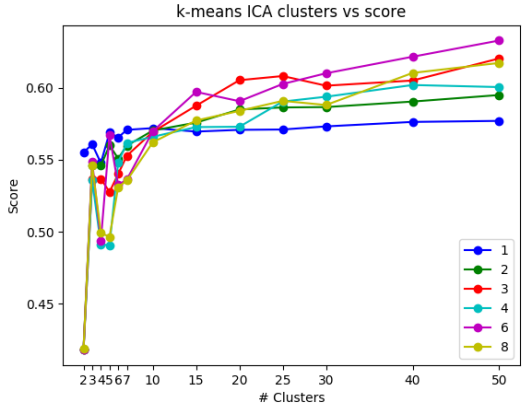


Clustering Analysis:

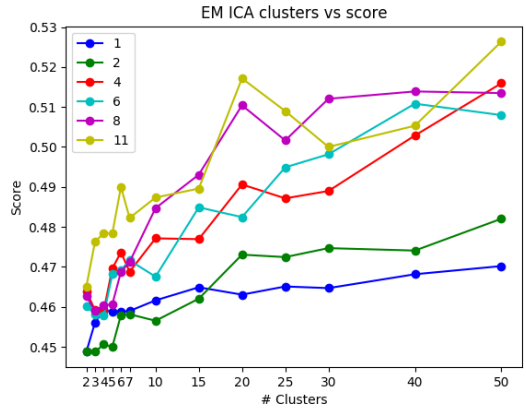
The clustering and the related accuracy measures below show this. They are as per the kurtosis measurement that we had seen earlier. For Abalone data, we have about 0.55% at around 6 components and then slower accuracy improvement after that. For Wine data, we see similar results 10 components.

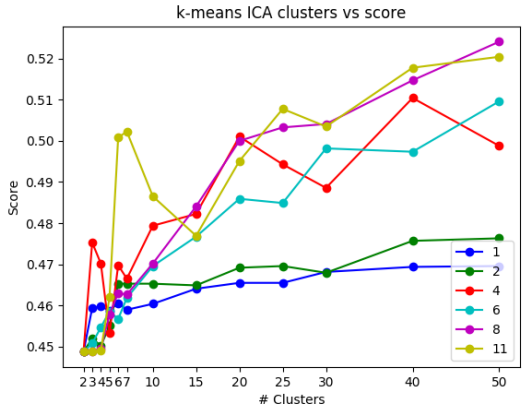
Abalone:





Wine Quality:



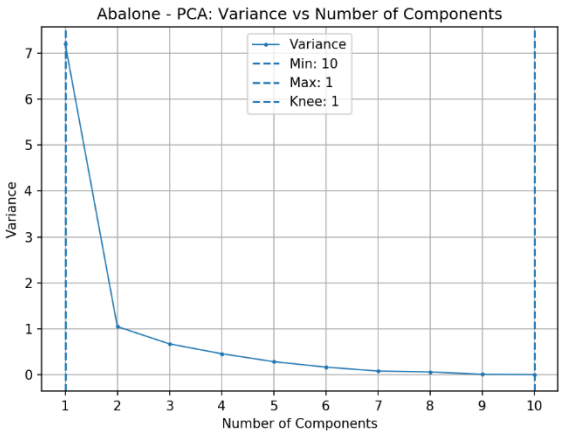


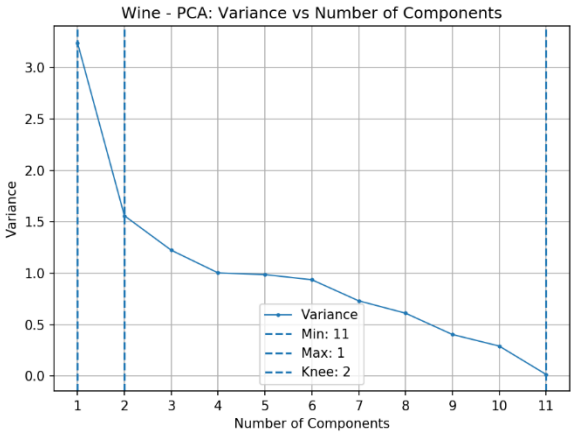
Principal Component Analysis:

PCA uses orthogonal transformation to convert a set of observations to a set of uncorrelated variables called principal components. It finds the orthogonal eigenvectors that best explain the maximum amount of variance.

Dimensionality Reduction:

To analyze dimensionality reduction, I measured variance vs the number of components for PCA.



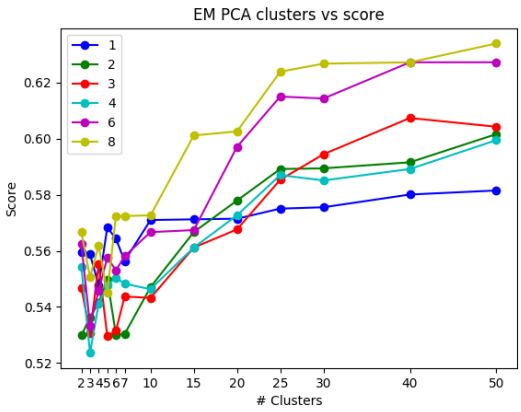


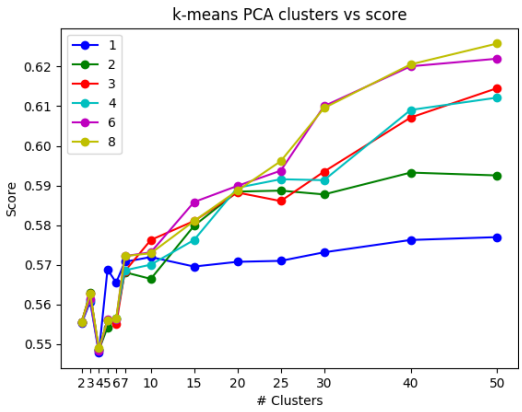
Variance was at the maximum at component 1 for both data sets. But for Abolone data, variance reduced faster than Wine Data. Based on this, I was expecting a less gradual increase in accuracy for Abalone and a more significant improvement in accuracy for Wine data as we increase the number of components.

Clustering Analysis:

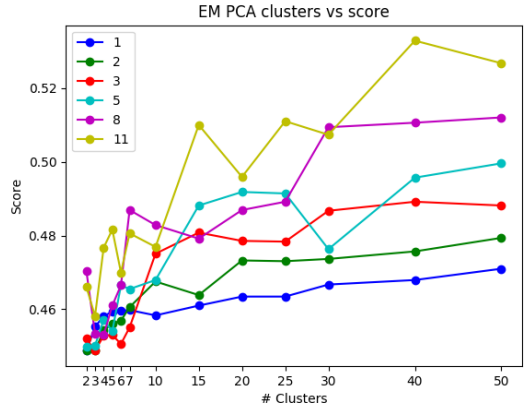
Below are the results that I got with PCA combined with EM and K-Means. As anticipated, Abalone data accuracy improved slower than Wine data. EM Vs K-Means also had an impact on the accuracy. K-Means accuracy was lower than EM's at each component number for Abalone and vice versa for Wine Data.

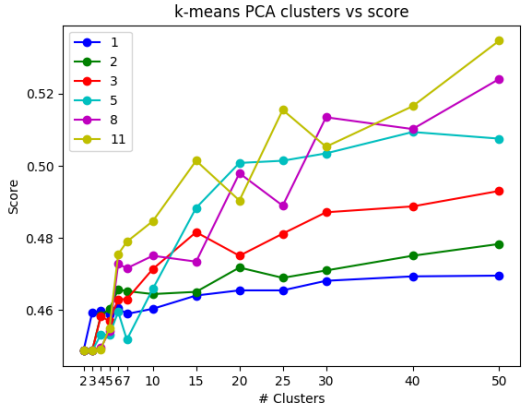
Abalone:





Wine Quality:



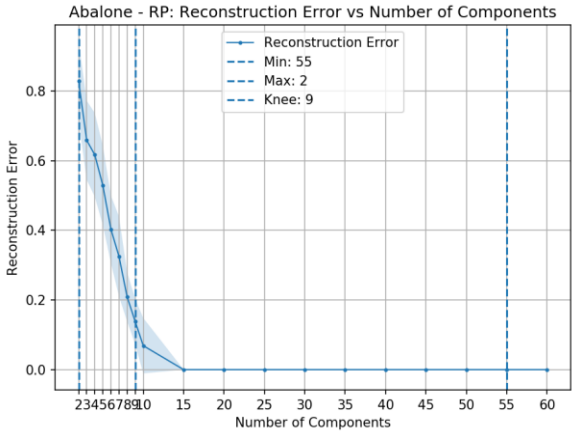


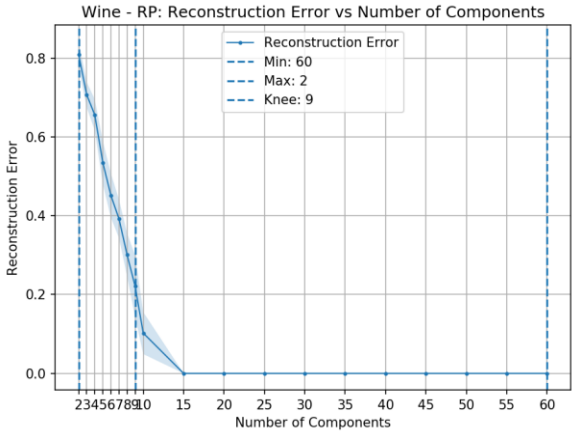
Randomized Projections:

Randomized Projections projects a point on vector space of a high dimension to a suitable lower dimension in a way that preserves the distance between points.

Dimensionality Reduction:

To analyze RP technique, I measured the Reconstruction error along with accuracy with other clustering algorithms. For both data sets, the knee was at 9 components.

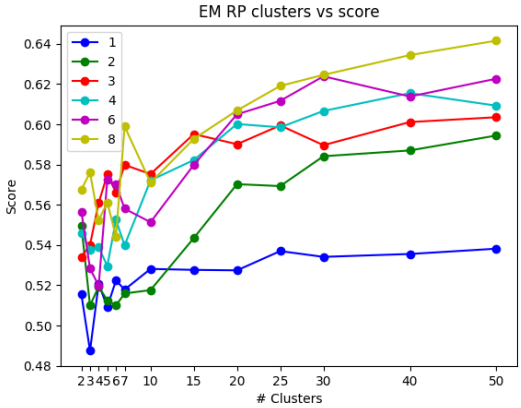


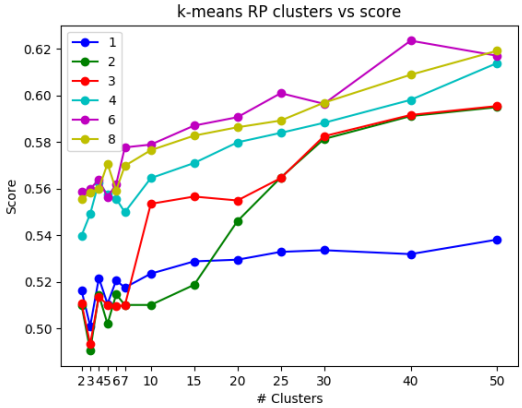


Clustering Analysis:

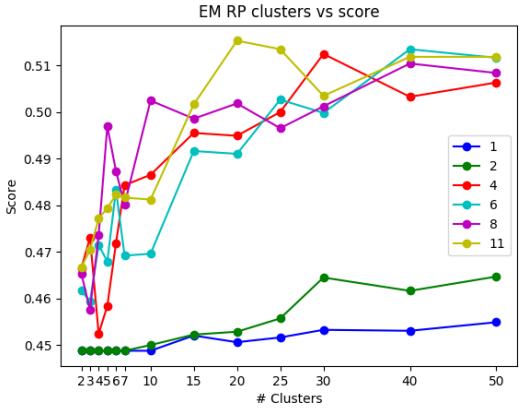
When I looked at the accuracy of RP with EM and K-Means, I found that K-Means accuracy increased significantly at number of components = 9. While for EM, it was closer to that and mostly at higher number of components. Similar performance was seen on both datasets.

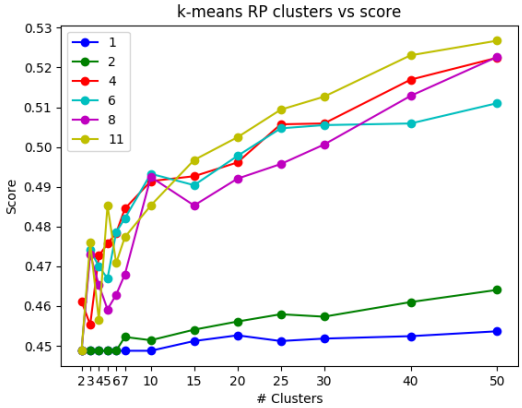
<u>Abalone:</u>





Wine Quality:



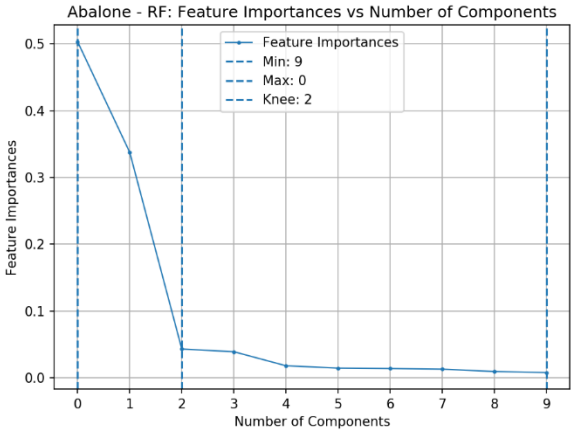


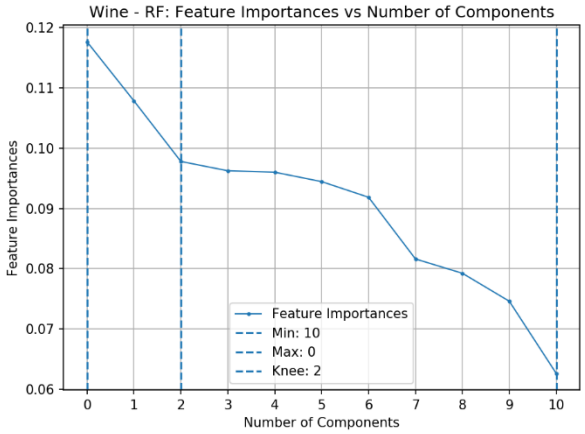
Recursive Feature Elimination:

Recursive feature elimination fits a model, and removes the weakest feature until the specified number of features are reached.

Dimensionality Reduction:

Below charts show the feature importance vs number of featured for both data sets. For Abalone, we see that at 2 features, the importance levels of. For Wine, while feature importance continuously reduces, it is not as steep as in Abalone.

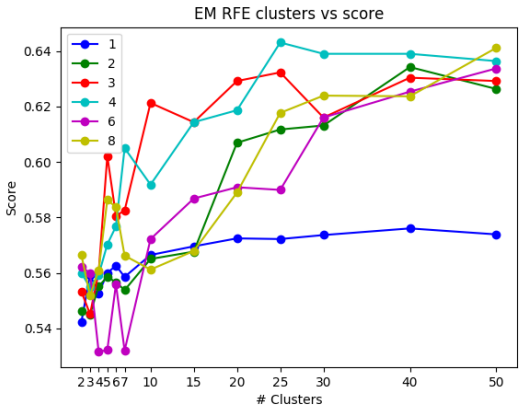


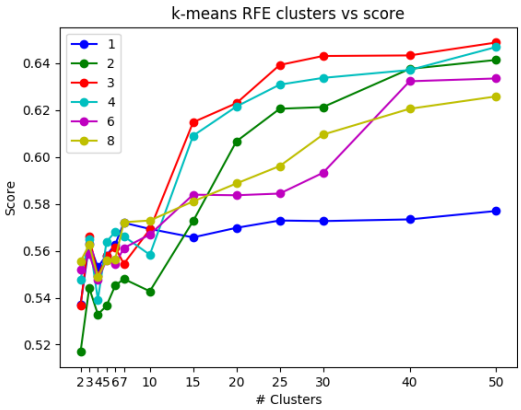


Clustering Analysis:

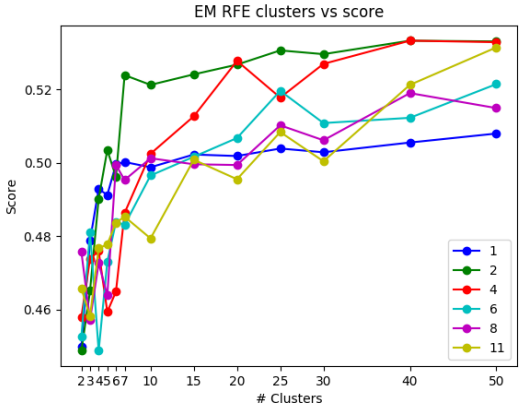
Below are the results obtained by running RFE in combination with EM and K-Means. We find that 2 features with around 20 clusters, gives a good result as we anticipated. On the Wine Data, we are seeing an increasing trend as the number of features are increased. A reasonably good accuracy is achieved at features=3 and clusters=20 and the accuracy levelling off after that.

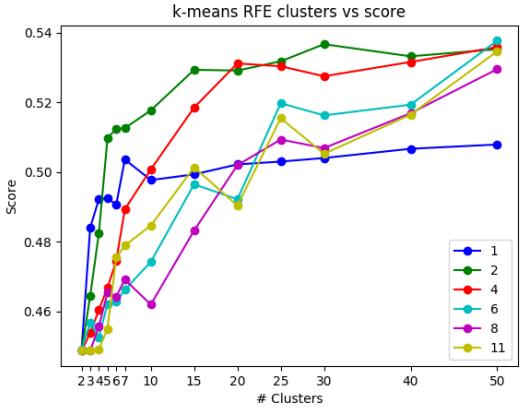
<u>Abalone:</u>





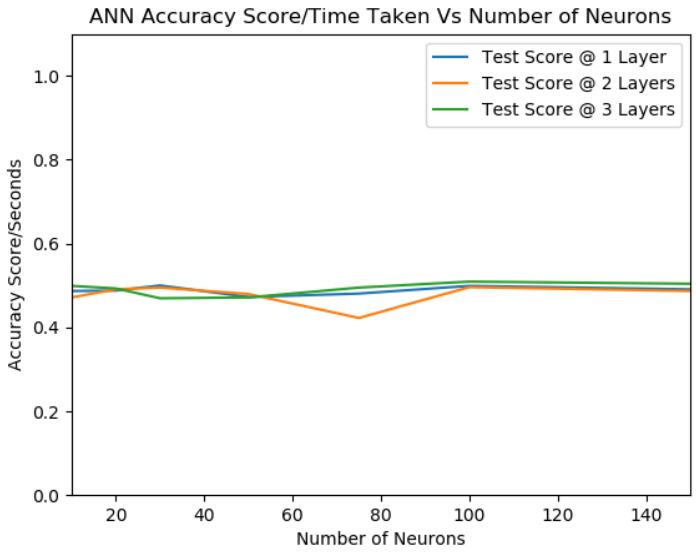
<u>Wine Quality:</u>





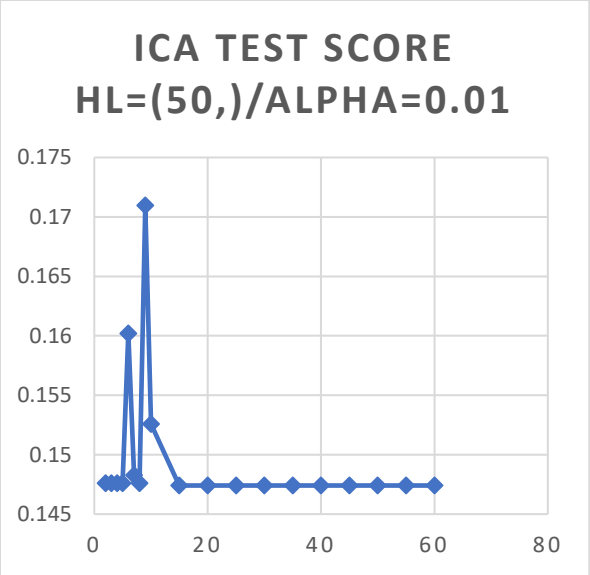
Part 3: Neural Network Analysis with Dimensionality Reduction and Clustering algorithms

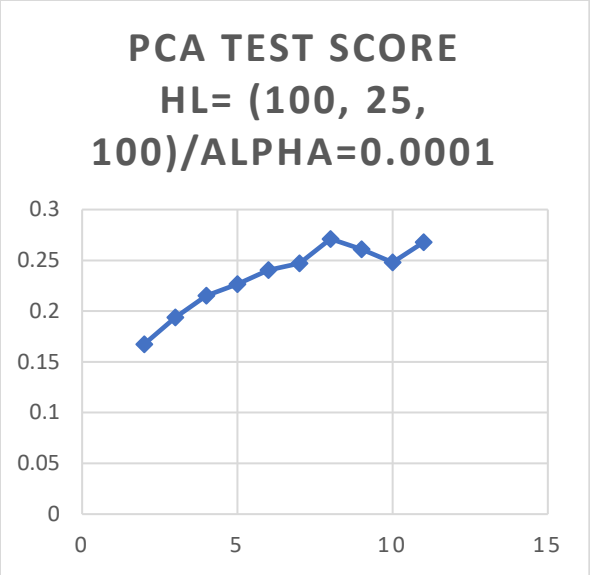
Finally I ran experiments to measure the accuracy of ANN that has dimensionality reduction applied on that. The following is the result that I obtained for Wine Data as part of Assignment 1. As you can see, the accuracy was around 50% for my ANN.

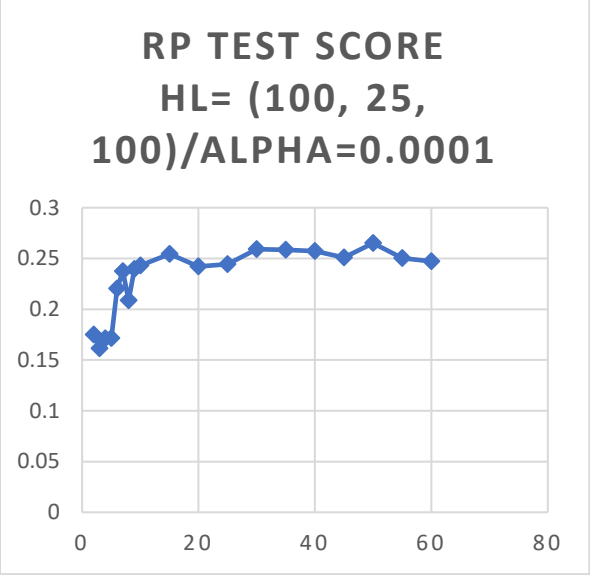


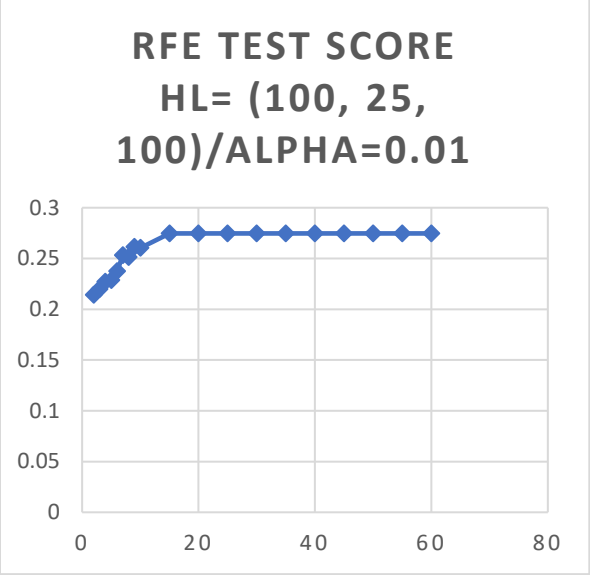
When I ran ANN after dimensionality reduction, I was not able to get the same accuracy. The best was close to 27% that I got with PCA, RP and RFE.

Following charts show the results that I got:









The worst performing of the lot was ICA. I got only about 17% accuracy with ICA. The best seem to PCA where an accuracy of 27% was achieved with about 8 components. RP and RFE also got about 27% but with more components.