Objectives: To experiment Clustering (K-Means, EM), Dimensionality reduction(Feature Reduction, ICA, PCA, Randomized projections) algorithm's on two types of datasets and implement Neural Networks on the feature reduced data to understand the algorithms and report the findings.

Data description and reason for choosing:

1) The first dataset was provided by the instructor and it is used to apply all algorithms. The data is cleaned by removing unnecessary variables (the first two variables that had no role as variables. The variable shares are divided into two classes (high and low) based on the median value. Any other metric could also have been used but as the data is continuous, the median is used as the mean could be biased towards the high sharing values. The data is scaled wherever it is necessary and the class labels were used only where they are necessary.
2) The other dataset is picked from UCI ML repository where the challenge is to identify whether the user defaults his credit card in the next month based on the previous months data. The reason for choosing this dataset is this type of problems are famous in the industry and the weight of the data is like that of the one given by the instructor (except number of variables). Also, the data is mix of continuous and categorical which makes the things a bit challenging. Again, the data is scaled wherever it was necessary and the output was.

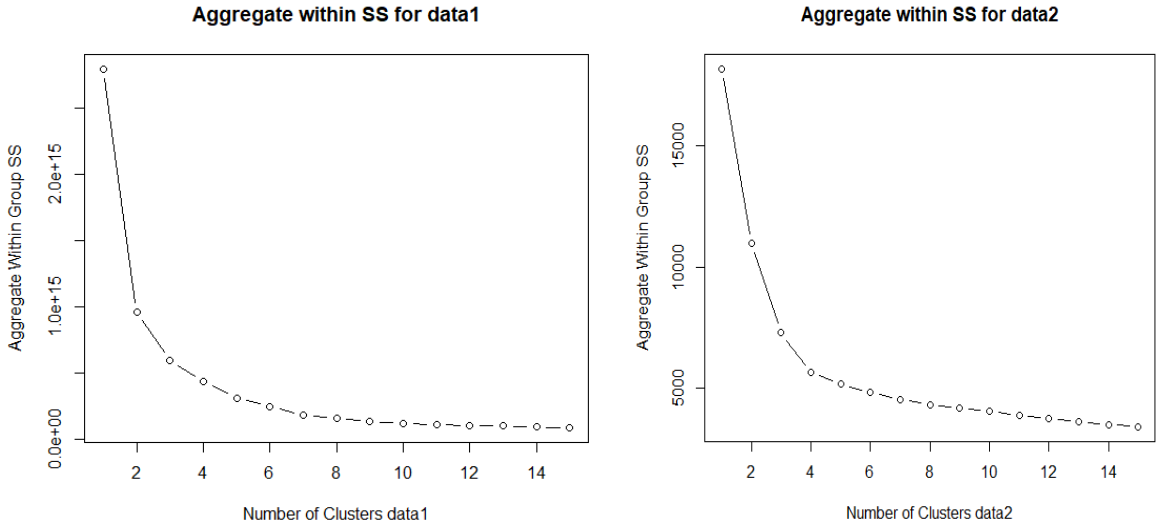
Experimentation 1: Clustering

For each dataset both the K-means clustering and the gaussian mixture model clustering algorithm's (Expectation Maximization) were performed.

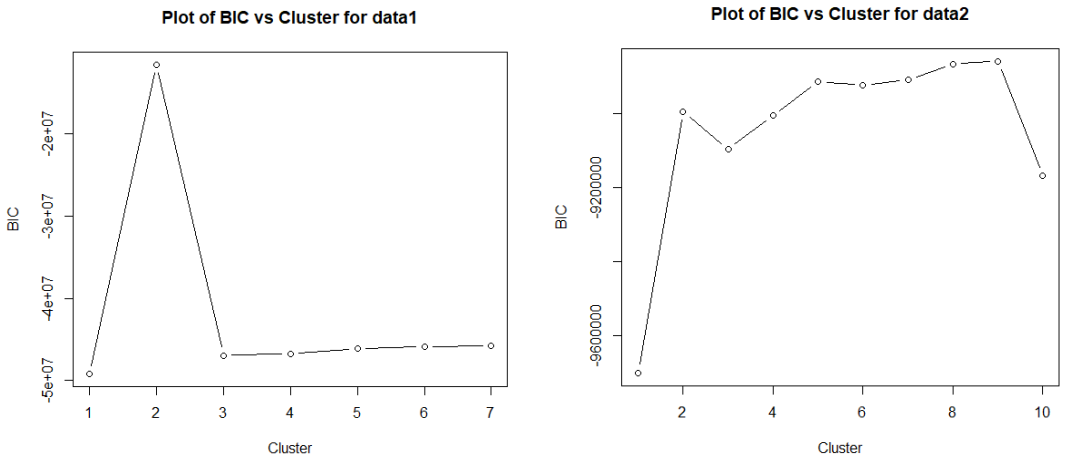
K-Means: For both the data sets a total of 15 centres were picked initially with an initial starting data nearest values of 10 were picked. Upon that both the dataset was looped iteratively to find the optimum number of K- Value. Below are the plots for both the datasets for K-means clustering.

Expectation Maximization: Both the datasets were experimented with gaussian mixture models GMM expectation maximization. In the process of generating the optimal centres both the datasets were normalized and the criterion of BIC was used and the distance between the points used was the Euclidean distance.

Upon running the algorithms 2 plots were generated for each type of dataset for each algorithm. The plot consists of the number of clusters Vs aggregate total sum of squares for K-Means clustering and number of clusters Vs Bayesian Information Criterion for expectation Maximization. Now the ideal number of clusters are chosen based at the point where the rate of change of total sum of squares and the maximum value of BIC for Expectation Maximization. The ideal number of clusters for data1 are found to be 6, 5 and the ideal number of clusters for data2 are found to be 4 ,9 for the algorithm's K-means and Expectation Maximization respectively. From both the experimentations and the data the number of clusters are inconsistent with the class label's that were assumed previously. It means that the data1 is not a 2-class problem and also, not the data2.. The clusters are inconsistent and are aligned naturally. When tried to understand the clusters the alignment was such that each class labels were aligned to each where the credit limit in high and credit limit is low. That is people who defaulted (with high and low credit card balance) and people who did not default (with high and low credit card balance). The understanding of clusters for data1 was not possible. The graphs for both the data and algorithms are shown below.



Graphs for data1 and data2 respectively after running Expectation Maximization.

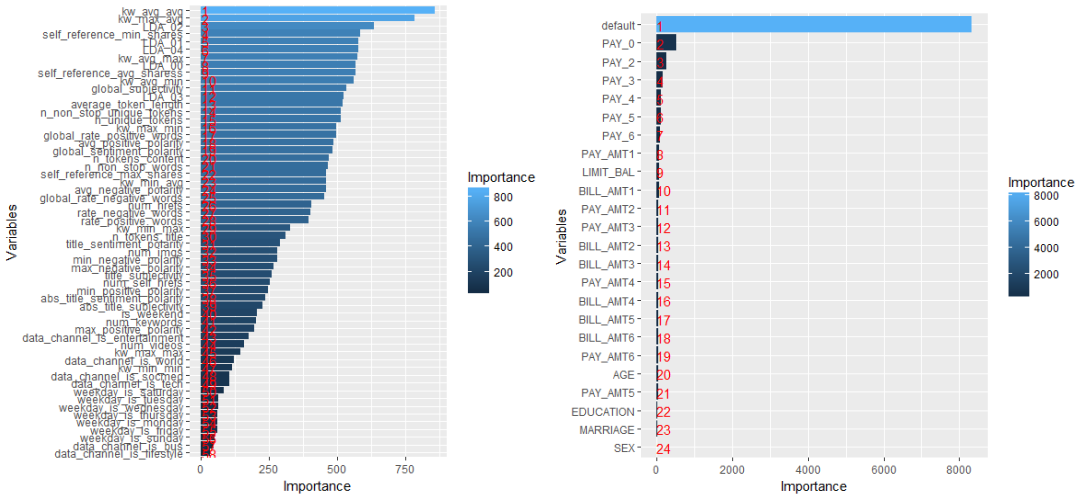


Experiment 2: Dimensionality Reduction:

Four dimensionality reduction algorithms are applied on two types of data all the necessary plots and interpretations are given below.

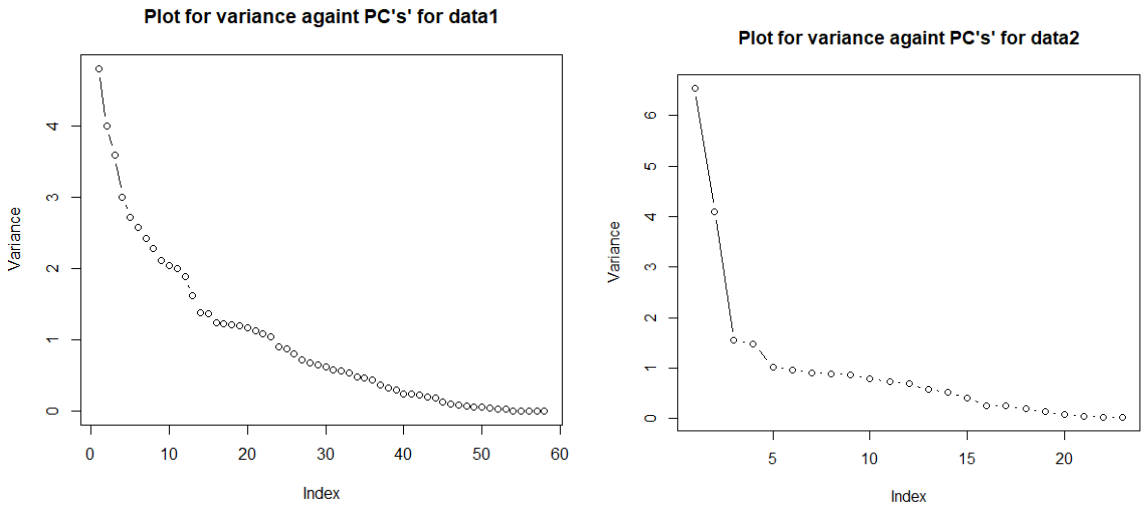
1) Feature Selection: The importance of features can be selected either using forward selection , backward elimination or decision tress(GINI or Information Gain). Instead, the feature selection criteria were used from the randomForest algorithm where the importance of the variables is calculated based on the GINI index which is exactly similar to the decision tress in CART. Based on that nearly 25% of the features were selected from both the datasets. Below are the plots for both the datasets where the variables are ranked along with the importance of the features. Though there is no hard and fast rule to identify the best and number of best features the following idea is used to find them.

Ranked features for data1 Ranked features for data2



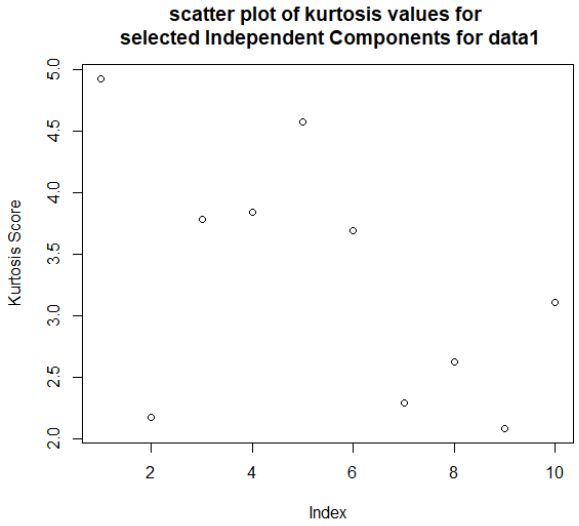
2) **Principal Component Analysis:**

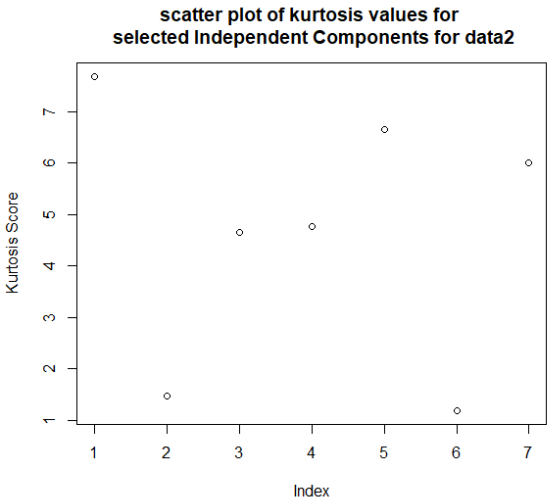
For both the datasets the principal components were extracted and are plotted against the number of principal components. For both the datasets the dependent variable is removed and the principal components are extracted and the data is scaled as well. Manually the variances for each of the PC's are calculated and the number of components are defined based on approximately the number of components that could explain the variance by around 75% to 80% when combined. Also the same number of components are evident from the graphs given below. To sum up in brief the number of principal components for data1 and data2 are 16 and 12 respectively. Post extraction, all the principal components are normalized to variance 1 and are stored for further exploration.



3) Independent Component Analysis:

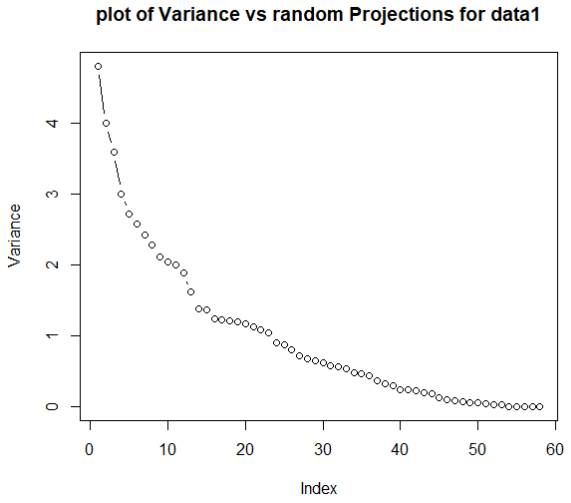
For both the datasets the independent components were extracted using the library fastICA. Post the extraction, to find the best components that can be used as reduced dimensions the independent components Kurtosis values were calculated to find the spread of the components. Finally, the values that are around the perfect kurtosis score (3) are extracted and are reported. The scatter plot of the independent component's Kurtosis scores is shown in the graph below.

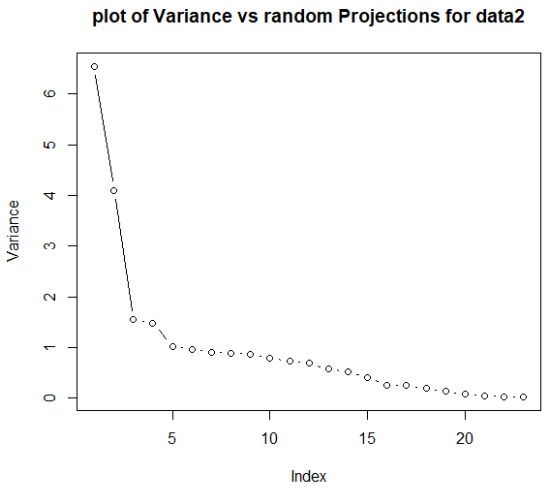




4) Randomized Projections:

The individual principal components for the data are calculated using rpca library. In this all the default values for the variables were taken and the data is scaled and centred. Also, the variance for each principal component is calculated and the components were taken which would approximately explain about 70-80% of the variance in the data. The number of principal components for the data1 and data2 are 16 and 6 approximately.





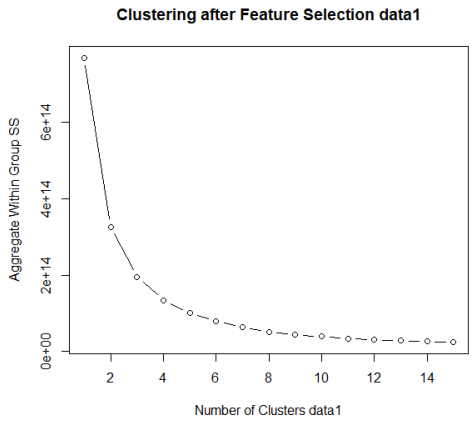
Interpretations:

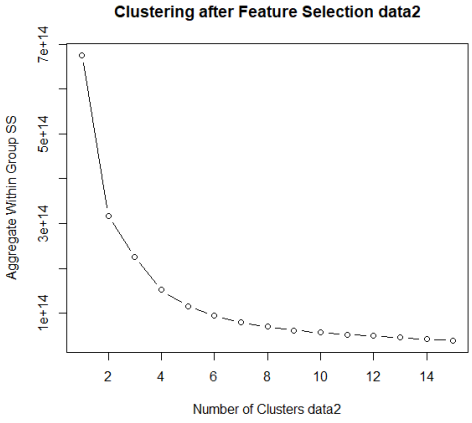
It can be interpreted from the above graph that as the number of principal components increase the variance associated with each principal component decreases. There is a steep point that can be calculated for data2 but for data1 the variance reduces drastically only after 9 principal components.

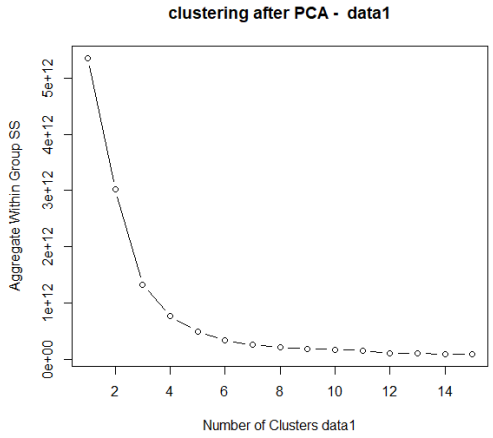
From the graphs the number of principal components for OnlineNewsPopularity dataset is approximately 9 and for Credit data the number of optimal principal components is 6.

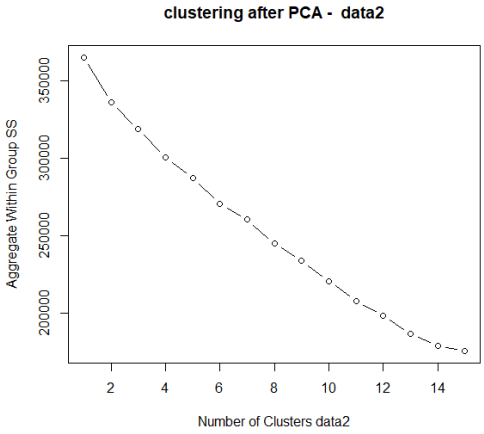
Experimentation 3:

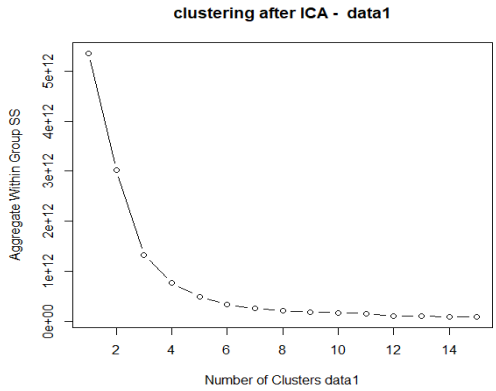
Cluster Plots for data1 and data2 after each dimension reduction modelling

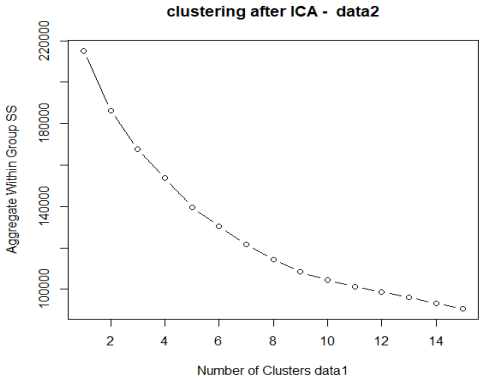


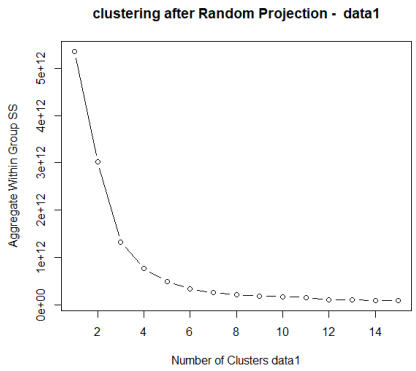


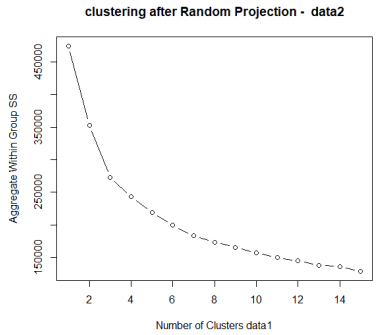




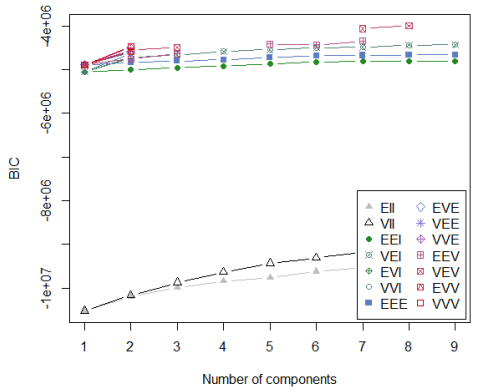


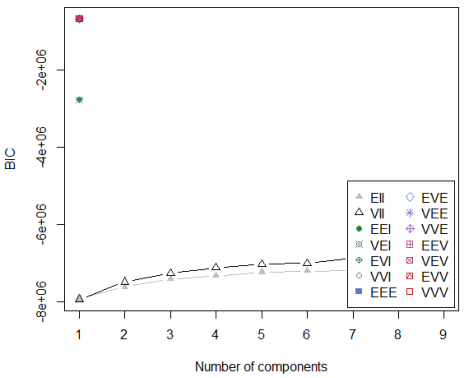


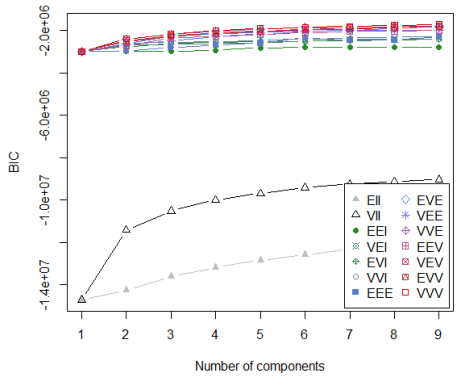


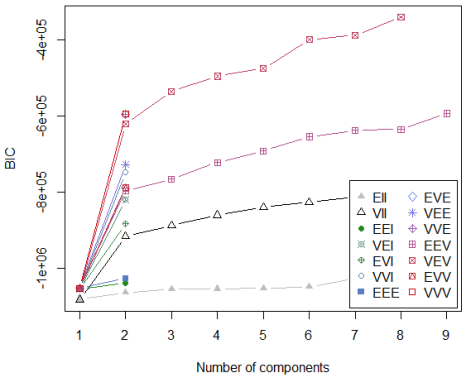


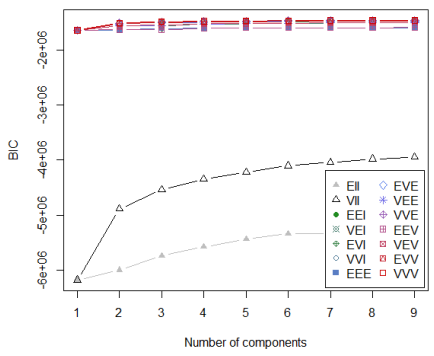
Expectation Maximization results after each dimension reduction modelling

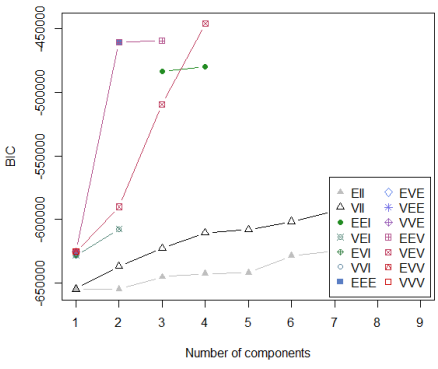


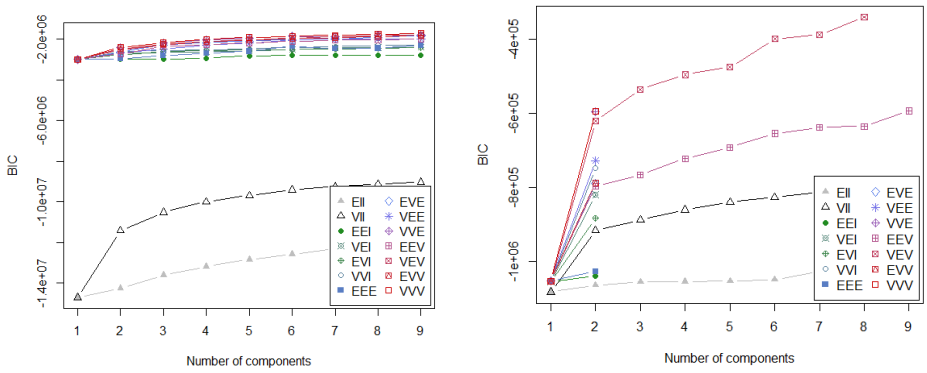












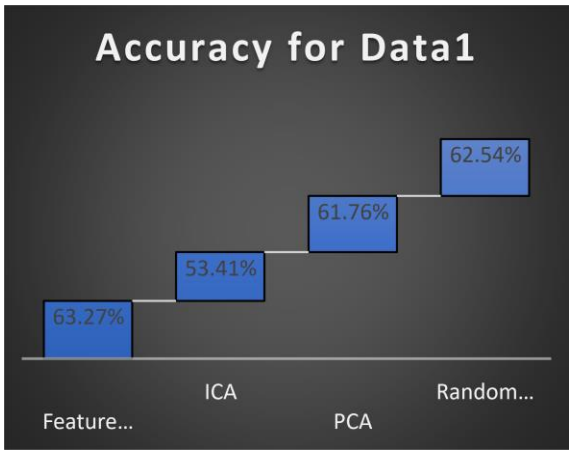
Observations of Experimentation 3:

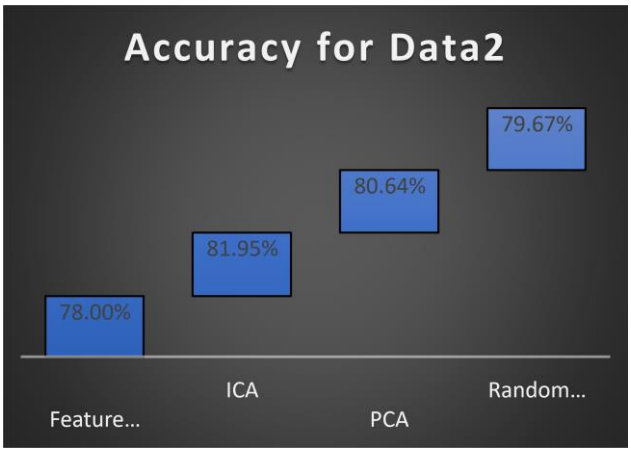
After running the clustering algorithms on the reduced featured data, the number of clusters for the second data set were perfectly 2 for all the feature reduced algorithms but the K- means did not give the proper clusters for the data2.

Where as in for data1 the K-means algorithm as well as the Expecation Maximization yielded almost the same results. Where the number of clusters are around 5. So from the above experimentation and for the future consideration it's suggested to use Expectation Maximization.

Experimentation 4: Neural Networks on PCA data

Post completing the PCA using 4 algorithms for 2 data sets the data is passed through the neural network skeleton. In the previous assignment for each individual data the optimal neural net is trained and the best model is picked out from it post tuning for epochs, batch size, activation function and validation split. The ideal combination for neural net is for 3 layers where the size of neurons were 26,16 and 1 respectively for layers. Also, the activation function sigmoid was found to be the optimal one with the epochs 25 and batch size of 5. With that the data for each principal component is split into train and test dataset by 75-25% and trained over the above given model and is tested for the test data set. Below is the graph for the performance metrics(accuracy) for data1 and data2 for the four principal component algorithms.





Observations:

1) Post comparing the performance as accuracy as the metric, the performance has not improved much. This might be because the most of the variance in the data of the PCA imight not be explaining the overall problem.
2) The speed with which the model increased drastically. The reason is a) The data is reduced by a lot of number of features and b) the data is standard normalized with variance is set to 1.
3) Feature reduction has the best performance from the the data1 which is approximately matched to the ANN from the previous assignment and Independent component analysis stood as topper for the second data set and is approximately equal to the ANN from the previous assignment.

Experimentation 5: Neural Networks after Clustering and Expectation Maximization:

From the first experimentation after the clustering and Expectation Maximization, the clusters were separated from both the models and are assigned with the class label's as they were done previously. The clustered data is again converted to dummy variables as the continuous data leads to wrong interpretation. The clustered data and the EM data is then passed to the previously built neural net and ran with the sigmoid activation function. In total four experimentations were done with in the fifth experimentation and the accuracy for each experiment are noted.

Below is the graphical representation of the accuracy for the data1 and data2 after running neural net on clustered experimentation and the Expectation Maximization experimentation.

Online News Popularity		Credit Data	
Algorithm	Accuracy	Algorithm	Accuracy
K-Means	53.62%	K-Means	67.21%
EM	61.71%	EM	68.27%

Conclusions:

The best results for classification were produced when the data were run through Feature selection using random forest algorithm for first data and then Neural network for data1 in dimension reduction. Where as for data2 the best results were produced by ICA and Neural Networks.

And when applying algorithms for clustering the best results were produced by Expectation Maximization for both the datasets after applying Neural Networks.