Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)

Objective:

To implement the following clustering algorithms on two datasets: -
1. K-means
2. Expectation Maximization

In addition, the following feature dimensionality reduction algorithms has been implemented and the above clustering algorithms has been performed on both datasets: -
1. Any one feature selection algorithm (decision tree, forward selection, backward elimination, etc.)
2. PCA
3. ICA
4. Randomized Projections

In addition to this Neural Network has also been performed after the above dimensionality reduction.

Dataset Details:

Two datasets have been used to implement the above-mentioned algorithms: -
Dataset 1: Appliances Energy Prediction dataset
Dataset 2: Absenteeism at Work dataset

Dataset Description and Data Pre-processing:

Appliances Energy Prediction dataset (Energy Dataset):

• The dataset is downloaded from UCI Machine Learning repository. Here is the link to download the dataset: https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction
• The Dataset consists of 19735 observations on 29 variables.
• None of the column contains any missing value, so no missing value imputation is required.

The target column for the Appliances Energy Prediction dataset is "Appliances". The Appliances column here is a continuous variable containing values ranging from 10Wh to 1080Wh with e median value of 60Wh.

Since our objective is to implement the clustering algorithms, we are going to keep most of the variable as is in the dataset except for the below mentioned columns. We are going to use Euclidean distance as distance metrics for implementing the clustering algorithms. Since we are using distance metrics for clustering, the data will be scaled using StandardScaler() from sklearn.preprocessing.

The column "date", "rv1" and "rv2" has been removed from the input variables as this column is not relevant for our analysis.

Absenteeism at Work dataset (Adult Dataset):

• The dataset is downloaded from UCI Machine Learning repository. Here is the link to download the dataset: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
• The Dataset consists of 740 observations on 21 variables.
• None of the column contains any missing value, so no missing value imputation is required.

The target column for the Absenteeism at Work dataset is "Absenteeism time in hours". The Absenteeism time in hours column here is a continuous variable containing values ranging from 0 hours to 120 hours with a median value of 3 hours.

Since our objective is to implement the clustering algorithms, we are going to keep most of the variable as is in the dataset except for the below mentioned columns. We are going to use Euclidean distance as distance metrics for implementing the clustering algorithms. Since we are using distance metrics for clustering, the data will be scaled using StandardScaler() from sklearn.preprocessing. The column ID has been removed from the dataset since this column is not required for analysis.

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)

IMPLEMENTATION:

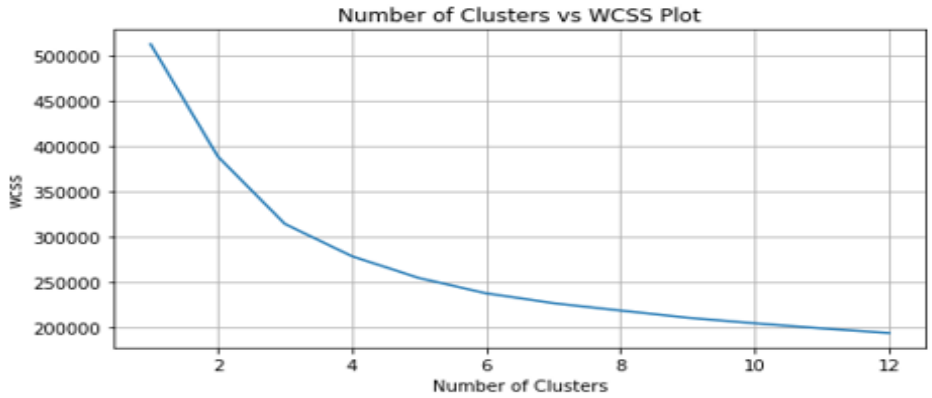
K Means Clustering:

The package used for implementing K Means Clustering is sklearn.cluster.Kmeans which was downloaded from SciKit learn. The k-means problem is solved using either Lloyd's or Elkan's algorithm. The average complexity is given by O(k n T), were n is the number of samples and T is the number of iteration. The worst case complexity is given by O(n^(k+2/p)) with n = n_samples, p = n_features. In practice, the k-means algorithm is very fast (one of the fastest clustering algorithms available), but it falls in local minima. That's why it can be useful to restart it several times.

KMeans complete Dataset 1 (Energy Dataset):-

As mentioned above Kmeans from sklearn was used to implement KMeans clustering. The number of clusters to form as well as the number of centroids to generate was set to default value as 8. Method for initialization, defaults to 'k-means++'. Number of times the k-means algorithm will be run with different centroid seeds was set to default value of 10. The final results will be the best output of n_init consecutive runs in terms of inertia. Maximum number of iterations of the k-means algorithm for a single run is set to a default value of 10.

Inoder to determine the ideal number of clusters for Energy dataset, Kmeans clustering was performed for different number of clusters and the respective Sum of squared distances of samples to their closest cluster center (wcss) was calculated for different number of clusters and the below plot was made:-

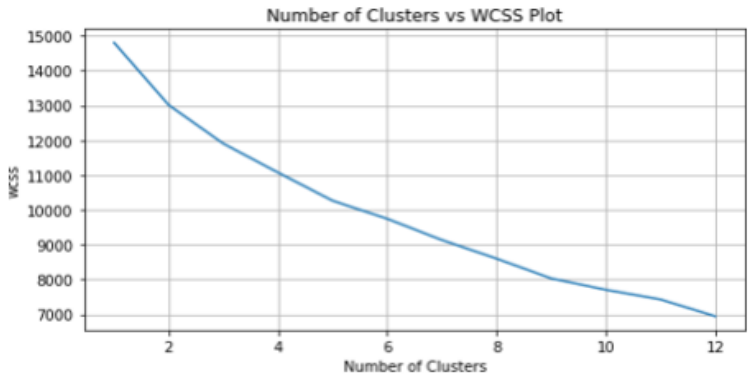


From the above plot it can be seen that the error keeps on decreasing with increasing number of clusters for the given energy dataset. The ideal number of clusters for the given dataset was determined using Elbow method and from the above plot we can see that 6 clusters looks ideal with minimum wcss for the Energy dataset

KMeans complete Dataset 2 (Adult Dataset):-

Now the same algorithm was run on the dataset 2, which is dataset to determine the wcss distances by iterating the number of clusters and the below plot was obtained for dataset 2. Here also it is noted that wcss decreasing with increase in the number of clusters. Applying the Elbow method, the ideal number of clusters that can be considered for the Adult dataset as per Kmeans clustering is 7.

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)



Expectation Maximization:

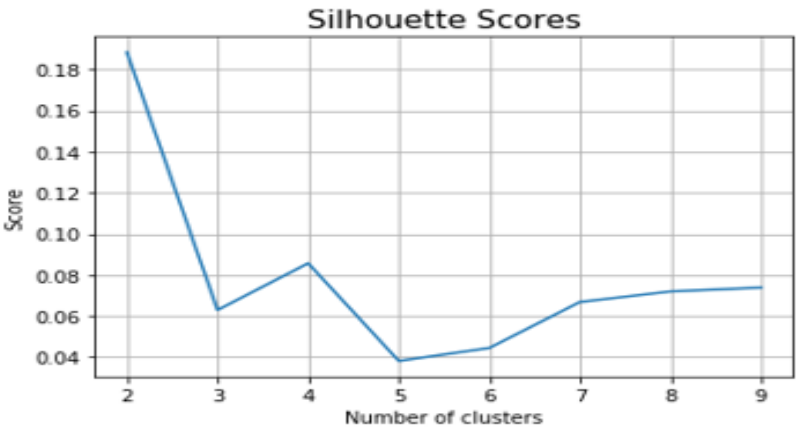
Expectation Maximization was implemented using sklearn.mixture.GMM. Gaussian Mixture Model is a representation of a Gaussian mixture model probability distribution. This class allows for easy evaluation of, sampling from, and maximum-likelihood estimation of the parameters of a GMM distribution. It initializes parameters such that every mixture component has zero mean and identity covariance.

Since we cannot apply elbow method to determine the ideal number of clusters in Expectation Maximization, we are calculating silhouette scores based on which we are deciding the ideal number of clusters for a given dataset through Expectation Maximization.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is 2 <= n_labels <= n_samples - 1.

EM on complete Dataset 1 (Energy Dataset):-

The Silhouette scores for different number of clusters through expectation maximization has benn calculated and the below plot was obtained:-

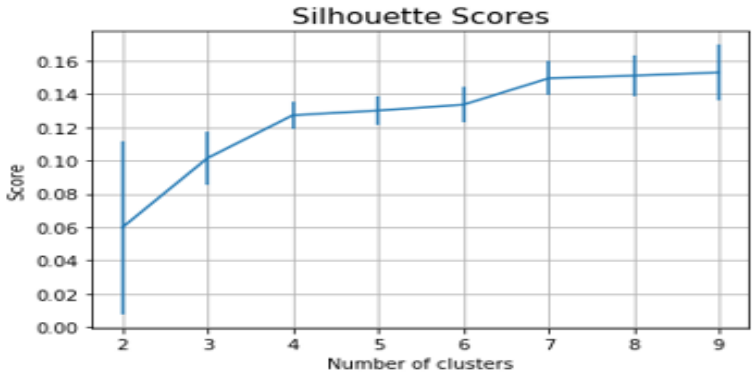


In the above plot it can be seen that there is a pike in the plot at the number of clusters equal to 4. Therefore, we are considering 4 to be the ideal number of clusters for the given energy dataset as the score is reasonable for this number of cluster.

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)

EM on complete Dataset 2 (Adult Dataset):-

Silhouette scores for different number of clusters was calculated using for the Adult dataset using GMM from sklearn and the following plot was obtained:-



As seen in the above plot, there is a pike in the graph at the number of clusters equal to 7. So as per the Expectation Maximization algorithm it is seen that the ideal number of clusters is same as that we obtained through Kmeans clustering for the Adult dataset.

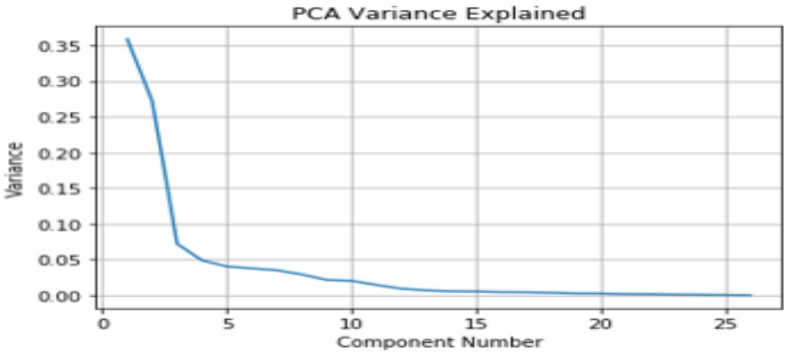
PCA Implementation:

PCA was implemented using PCA from sklearn.decomposition. Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. The input data is centered but not scaled for each feature before applying the SVD. It uses the LAPACK implementation of the full SVD or a randomized truncated SVD by the method of Halko et al. 2009, depending on the shape of the input data and the number of components to extract. It can also use the scipy.sparse.linalg ARPACK implementation of the truncated SVD.

Inoder to determine number of components that can be considered is based on the variance values that each component can describe the dataset. If a combination of components are able to explain atleast 70% variation in the dataset, then that number of components can be fixed and taken for further analysis.

PCA on dataset 1 (Energy dataset):

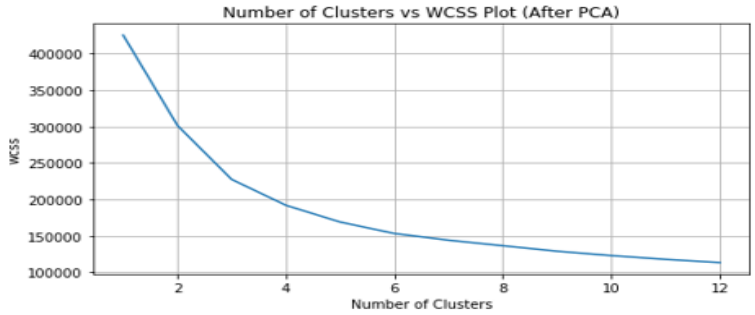
PCA was implemented and the following plot was obtained that shows the variance for each of components obtained after PCA for the given dataset



As in the above graph it can be inferred that 75% variance in the dataset can be described by first 6 components. There considering 6 components as the final set of features after feature transformation for Energy dataset and this will be used to determine the ideal clusters using Kmeans and EM.

Kmeans after PCA

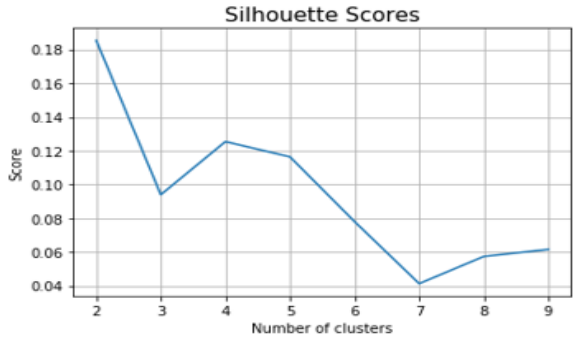
Considering 6 components after PCA, Kmeans was implemented and wcss was calculated and plotted against the number of clusters as shown below:



Now the number of clusters is 6 after applying Elbow method while it was the same earlier as well before PCA.

EM after PCA

With the output components of PCA , EM was applied and the below plot was obtained:



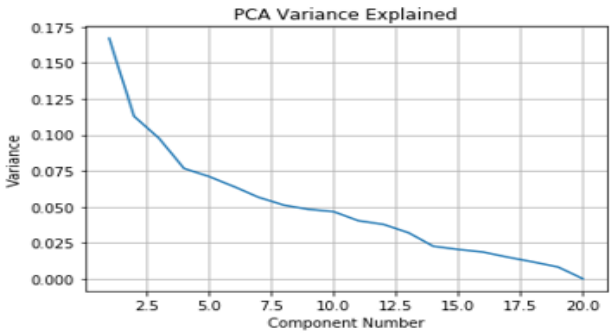
Here too, the ideal number of clusters is 4 after PCA while before PCA too it was 4 for the energy dataset by EM algorithm.

PCA on dataset 2 (Adult dataset):

PCA was implemented on the adult dataset and a plot was obtained for the number of components against the variance as shown in the below plot.

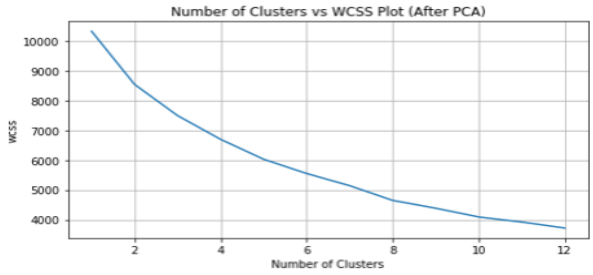
From the below plot it is seen that 8 components are able to explain 75% of the variance in the dataset. So we will consider 8 components now for determining the ideal number of clusters using KMeans and EM

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)



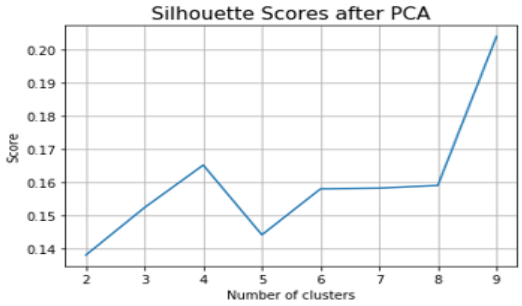
Kmeans after PCA

Using 8 components after PCA, it is seen that the ideal number of clusters can be taken as 7 which is almost similar as earlier obtained for the same dataset using Kmeans.



EM after PCA

After performing PCA in the Adult dataset and determining the ideal number of clusters using EM produced 4 clusters as ideal while it was 7 before doing PCA



ICA Implementation:

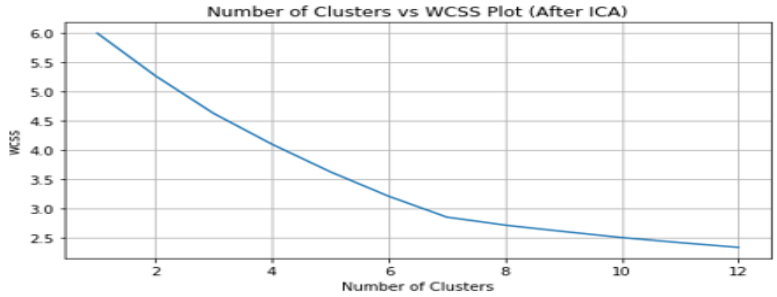
ICA was implemented using FastICA from sklearn.decomposition. FastICA is a fast algorithm for Independent Component Analysis. ICA was implemented as follows:

ICA on dataset 1 (Energy dataset):

Using FastICA it was determined to have 6 components which will be considered further to determine the ideal number of clusters.

KMeans after ICA:

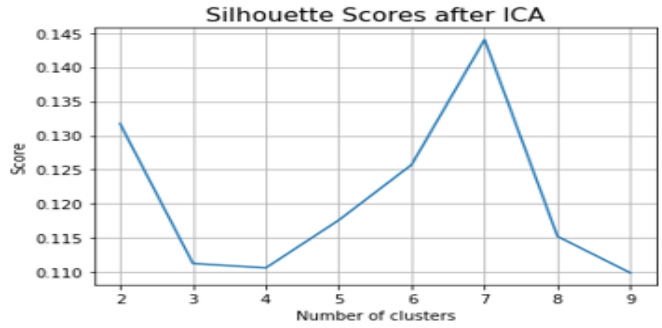
The below plot was obtained for wcss vs number of clusters after ICA:-



Using Elbow method, it is seen that the ideal number of clusters is 7 while the number of clusters for this dataset was 6 before ICA using Kmeans.

EM after ICA:

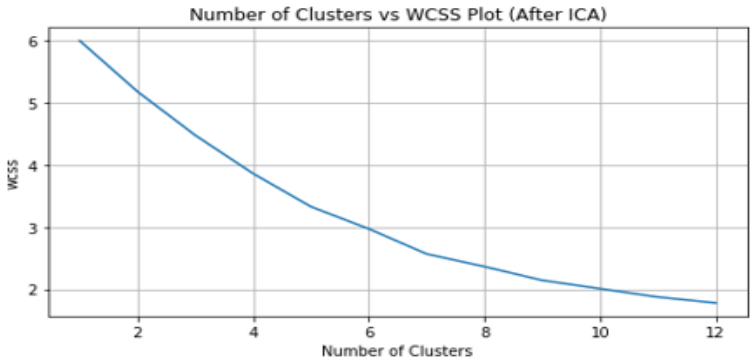
The below plot was obtained after performing ICA and then determining the ideal number of clusters using EM for Energy dataset:-



The ideal number of clusters that was obtained was seven as per the above chart after ICA while it was 4 for the Energy dataset before ICA using EM

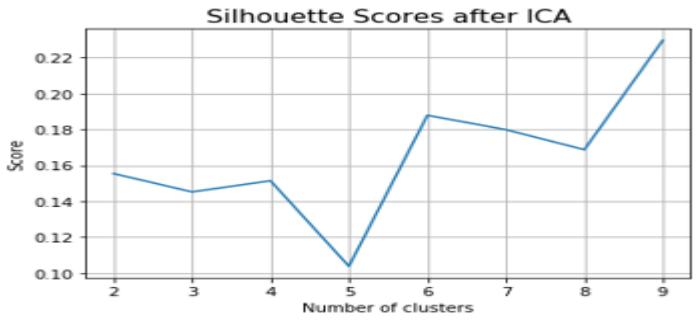
ICA on dataset 2 (Adult dataset):

Kmeans after ICA



From the above plot it is seen that the ideal number of clusters seems to be 6 after ICA for Adult dataset using K means while it was 7 before ICA

EM after ICA:



The above plot tells us that the ideal number of clusters after ICA using EM is 6 while it was 7 before ICA was done.

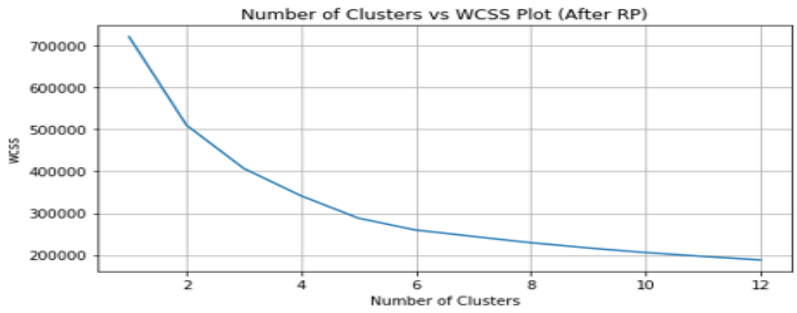
Randomized Projection Implementation:

Randomized Projection was implemented using SparseRandomProjection package imported from sklearn.random_projection. It Reduces dimensionality through sparse random projection

Sparse random matrix is an alternative to dense random projection matrix that guarantees similar embedding quality while being much more memory efficient and allowing faster computation of the projected data.

Randomized Projection on dataset 1 (Energy dataset):

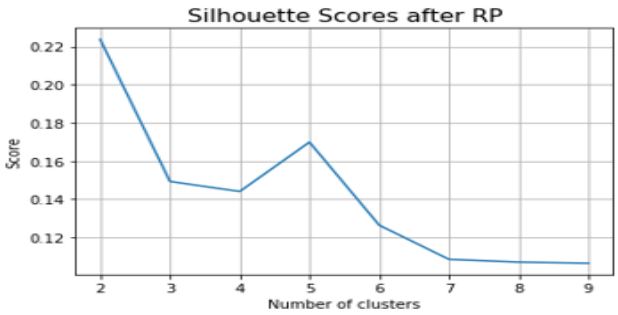
Kmeans after RP:



In Randomized Projection , we could see that the elbow is identified at cluster number 5 for ENERGY dataset . The WCSS is found to be 300,000 . This error value is less when compared to the model developed with original features .

EM after RP:

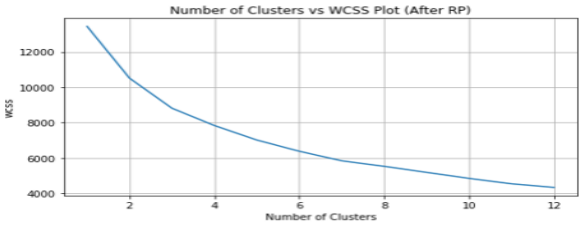
Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)



This plot shows the Silhouette scores for EM algorithm . Even though the cluster 2 has high score , we should select the cluster 5 configuration because the score is high when compared to 2 and 3 cluster configuration.

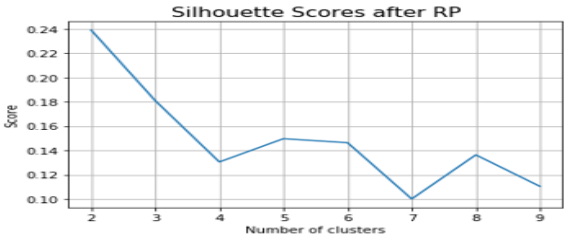
Randomized Projection on dataset 2 (Adult dataset):

Kmeans after RP:



In this RP model by K – Means , the optimal number of clusters is 5 . Because after this point , the WCSS doesn't seem to decrease significantly. So , we have to choose cluster 5 configuration .

EM after RP:



In EM for Adult dataset , the Silhouette Scores spiked at 5 even though the 6 also has the same Score . So , if there is a clash , the lower cluster number should be chosen . So , we will choose the cluster 5 configuration .

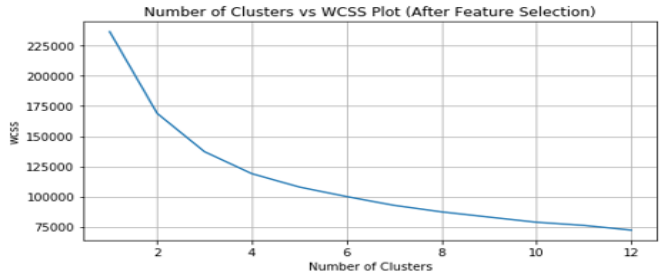
Feature Selection (backward elimination) Implementation:

Here, backward elimination algorithm was used to implement feature selection for dimensionality reduction. Backward elimination was implemented using RFECV package imported from sklearn.feature_selection. This is just feature ranking with recursive feature elimination and cross-validated selection of the best number of features.

Backward Elimination on dataset 1 (Energy dataset):

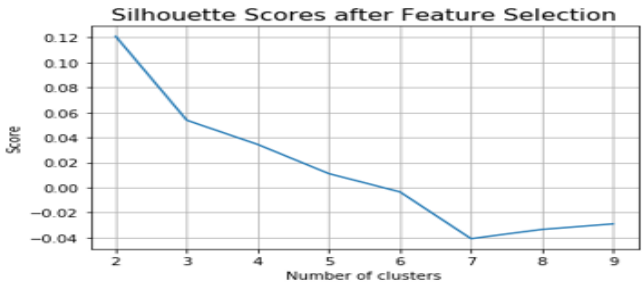
Kmeans after backward elimination:

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)



By comparing the WCSS error values , we should choose cluster 4 . Because after cluster 4 , the WCSS doesn't drop significantly after that . So , we should choose the cluster 4 than any other cluster configuration .

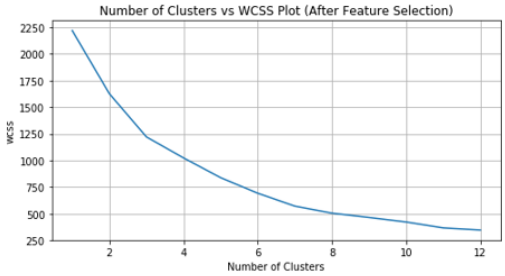
EM after backward elimination:



In EM , we should choose cluster 4 configuration . Because for soft clustering , the number of clusters will always be more . So we shouldn't choose higher cluster values . So , we have chosen the cluster value to be 4.

Backward Elimination on dataset 2 (Adult dataset):

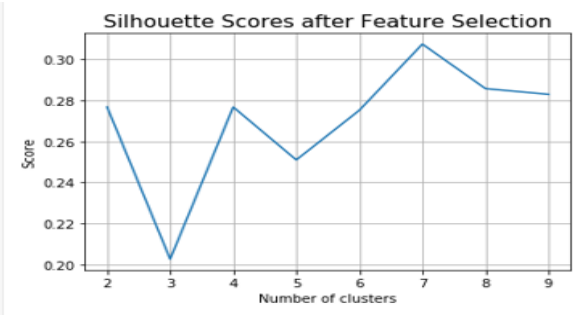
Kmeans after backward elimination:



In K-Means , we should choose the cluster 5 configuration. Because any cluster above this point , the WCSS error doesn't seem to decrease significantly . So , we should choose the cluster 5 configuration. Any point above this , doesn't make sense .

EM after backward elimination:

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)



From the above plot , we should choose cluster 4 because it has good Silhouette Scores when compared to other cluster configuration . So , we should choose Cluster 4 configuration .

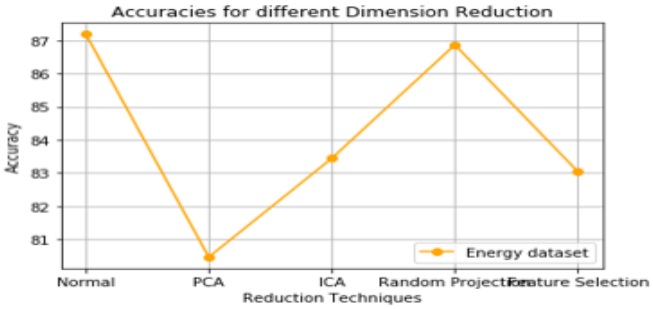
Artificial Neural Network Implementation:

For implementing Artificial Neural Networks "MLPClassifier" package was downloaded from Sci-kit Learn. Artificial Neural Networks using MLPClassifier was implemented for the normal dataset and then the dimensionally reduced datasets using the implemented algorithms: PCA, ICA, feature selection and Randomized projection.

ANN was implemented in this section using the dimensionally reduced algorithms on dataset 1 and dataset 2 using the above implemented algorithms.

Accuracy Scores of ANN implemented after dimensionality reduction For Energydataset:

The accuracy scores of each of the dimensionality reduction algorithms was calculated using the implemented algorithm and the was plotted to compare the accuracies of different energy datasets.

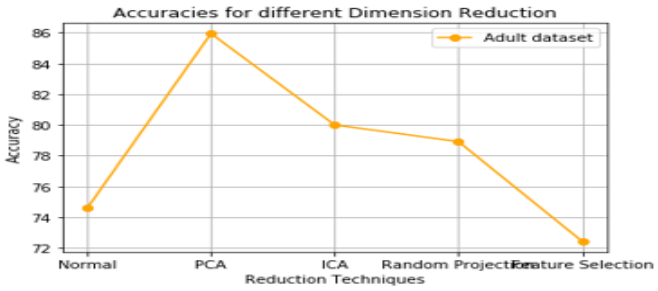


From the above plot it is seen that the Accuracy score of the normal dataset is the highest when compared with other datasets with an accuracy over 87%. Accuracy score of RP dataset is close to it nearing 87% The least is the PCA transformed dataset with an accuracy below 81%.

Accuracy Scores of ANN implemented after dimensionality reduction For Energydataset:

Below is the accuracy scores plot of the Adult dataset:-

Bharath Kumar Natarajan (BXN180005)
6341.003 – Applied Machine Learning (Assignment 4)



The above plot tells us that PCA transformed Adult dataset has the highest accuracy of over 86% Next to that is the ICA transformed dataset with an accuracy score of 80%. The least is the feature selection (backward elimination) with an accuracy score of 72%.

Conclusion:

From the implementation and experiments done in this assignment, we can infer that the number of clusters varies, given the data, when we apply various dimensionality reduction algorithms. We cannot say a particular dimensionality reduction algorithm is always the best as it depends on the dataset as studied in this assignment.

Maybe the optimization of these algorithms can further be improved by looking into factors such as checking and treating the extreme values or outliers in your variables. This could be one reason why your predicted estimate values might vary as they are getting skewed by the outlier values.