Assignment 4

Data Description

Dataset 1: Appliances Energy (https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction)

We have cross-sectional dataset with 19735 observations and 29 variables. The variables are Appliances and lights energy consumption in the house, 9 temperatures and 9 relative humidity in 9 different locations in the house, 6 weather variables such as outside temperature, outside humidity, Pressure, Windspeed, Visibility and Dewpoint. We also have date column, and two random variables that are being ignored in this model. Additional information about the dataset can be found out in the above-mentioned link.

Dependent variable is Appliances which is categorized as binary energy high variable(1 if energy is greater than median) and Date variable is dropped from the dataset for convenience

None of the variables consists of missing attributes so no missing value imputation is required

Data is partitioned randomly into train and test using 70:30 split percentage(13814 and 5921 observations respectively)

All the features are scaled to the range [0,1] by subtracting each feature from its mean and dividing by standard deviation in order to maintain evenness in the data

Dataset 2: German Credit Data (Attached)

We have cross-sectional dataset with 1000 observations and 21 variables(7 numerical and 14 categorical). Target variable is Default which is renamed as Creditability (binary, 1=Good, 0=bad).

None of the attributes have missing values so no missing value imputation is required

The numerical attributes are rescaled(min-max normalization) and for categorical attributes OneHotEncoding is used to create the dummy variables. The dataset now has 1000 observations and 62 attributes.

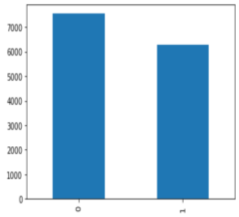
Data is partitioned randomly into train and test using 70:30 split percentage(700 train and 300 test observations respectively)

This dataset is chosen given the number of categorical variables it has for analyzing whether an applicant is creditable or not. The dataset provides a credit simulation close to the realistic credit simulation in the modern credit analyses with additional variables like the criminal records of applicants, their health information etc. This is similar to what is done by a bank before a credit is approved.

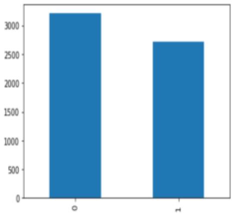
Modern credit analyses employ many additional variables like the criminal records of applicants, their health information, net balance between monthly income and expenses. A dataset with these variables could be acquired or complementary variables added to the dataset. This will make the credit simulations much realistic, similar to what is done by the banks before a credit is approved.

Number of responses for both the datasets for train and test data:

Appliances Energy Prediction Dataset

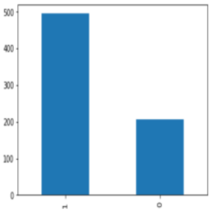


Train

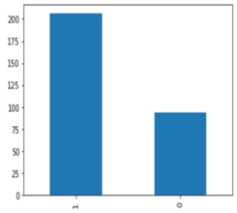


Test

German Credit Dataset



Train

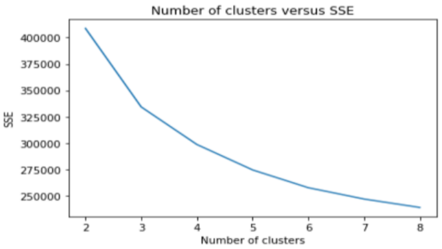


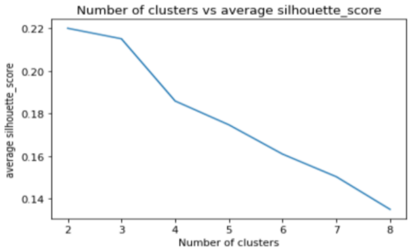
Test

Clustering

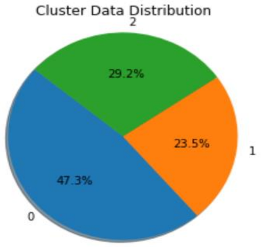
KMeans

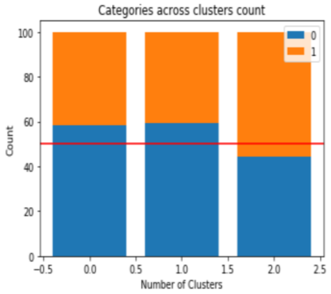
Appliances Energy





Elbow Graph(uses SSE) and Silhouette score graph(Silhouette Index) gives the number of optimal clusters for KMeans as 3

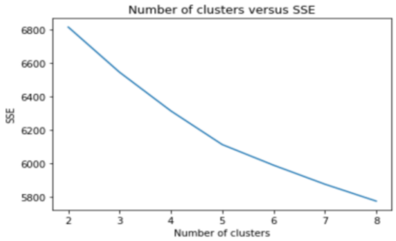


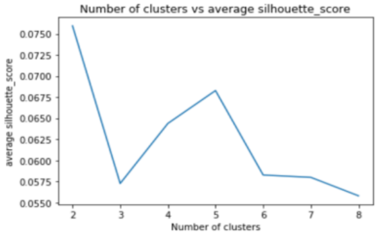


Percentage of observations in each cluster and the spread of observations in each cluster is varied

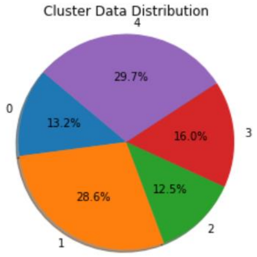
Percentage of two class labels in each cluster and the spread of observations in each cluster is approximately 50%

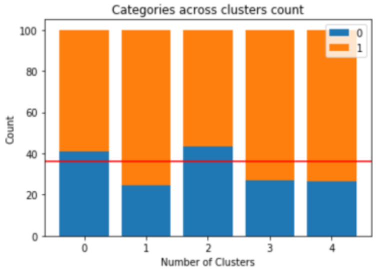
German Credit





Elbow Graph(uses SSE) and Silhouette score graph(Silhouette Index) gives the number of optimal clusters for KMeans as 5



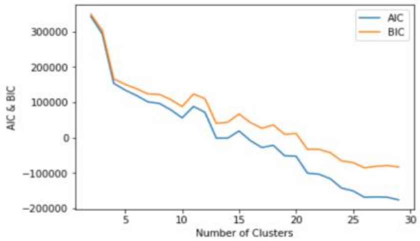


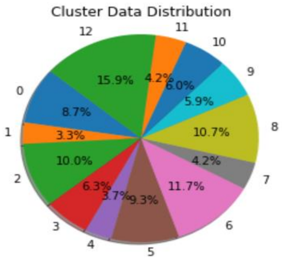
Percentage of observations in each cluster and the spread of observations in each cluster is varied

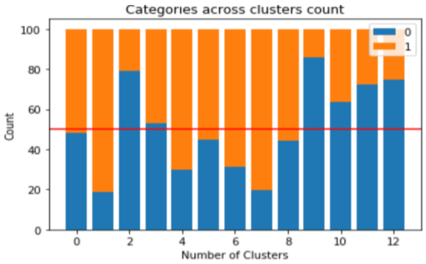
Percentage of two class labels in each cluster and the spread of observations in each cluster is approximately 50%

Expectation Maximization

Appliances Energy







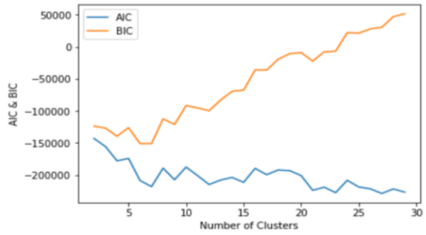
Based on BIC score the optimal number of clusters chosen is 13 because when compared to AIC, BIC has a penality term which leads to the better estimation

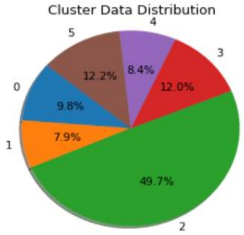
The left most graph suggests a marginal decrease in the BIC at k=13 which is why 13 has been chosen as the optimal number of clusters

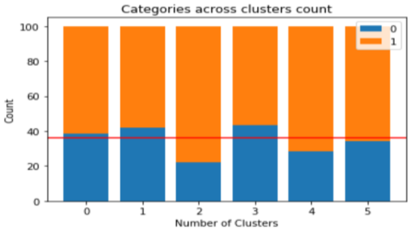
The pie chart in the middle shows the approximately equal spread of observations in each of the clusters

The right most graph shows class distribution in each of the clusters and the average distribution is 50%

German Credit







Based on BIC score the optimal number of clusters chosen is 6 because when compared to AIC, BIC has a penality term which leads to the better estimation

The left most graph suggests a marginal decrease in the BIC at k=6 which is why 6 has been chosen as the optimal number of clusters

The pie chart in the middle shows the varied spread of observations in each of the clusters

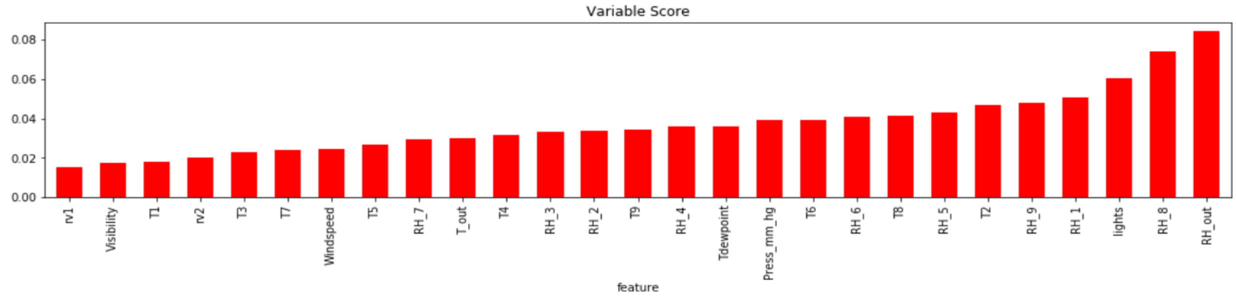
The right most graph shows class distribution in each of the clusters and the average distribution is 36%

<u>Dimensionality Reduction Algorithms</u>

Feature Selection

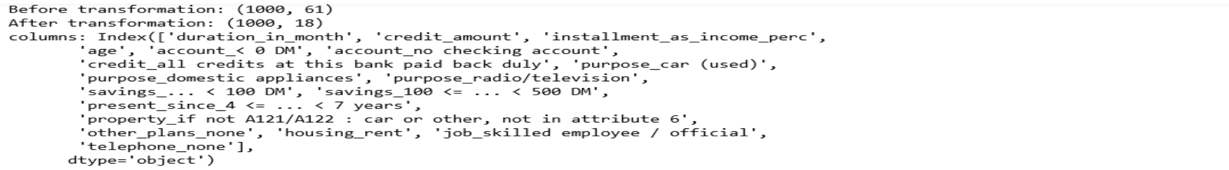
Appliances Energy

```
Before transformation: (19735, 27)
After transformation: (19735, 11)
columns: Index(['lights', 'RH_1', 'T2', 'RH_5', 'T6', 'RH_6', 'T8', 'RH_8', 'RH_9',
       'Press_mm_hg', 'RH_out'],
      dtype='object')
```

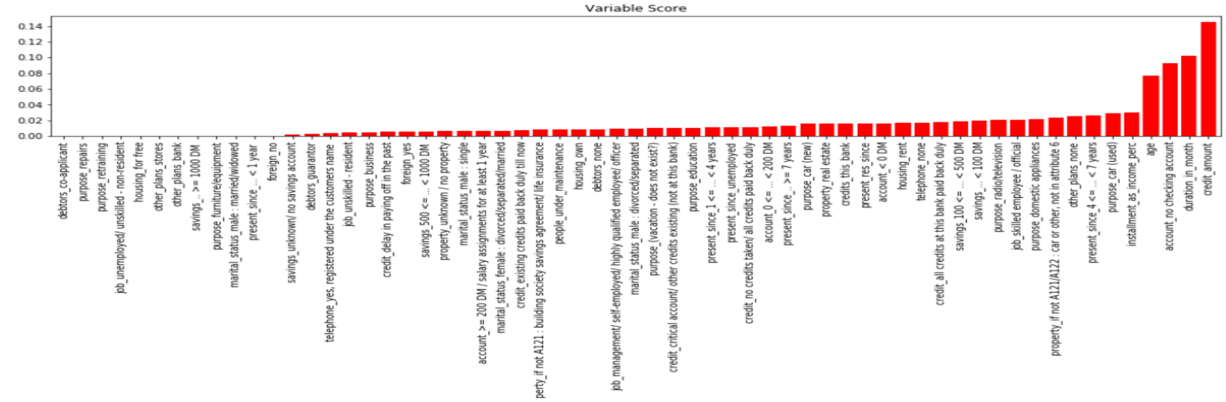


Decision Trees is used as the feature selection algorithm and 11 optimal features are selected among 27 original features. The graph represents the features from least important in the left to the most important in the right.

German Credit



```
Before transformation: (1000, 61)
After transformation: (1000, 18)
columns: Index(['duration_in_month', 'credit_amount', 'installment_as_income_perc',
       'age', 'account_< 0 DM', 'account_no checking account',
       'credit_all credits at this bank paid back duly', 'purpose_car (used)',
       'purpose_domestic appliances', 'purpose_radio/television',
       'savings_... < 100 DM', 'savings_100 <= ... < 500 DM',
       'present_since_4 <= ... < 7 years',
       'property_if not A121/A122 : car or other, not in attribute 6',
       'other_plans_none', 'housing_rent', 'job_skilled employee / official',
       'telephone_none'],
      dtype='object')
```

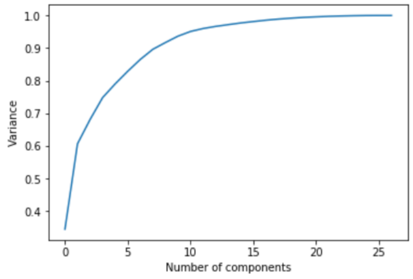


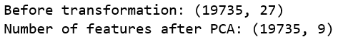
Decision Trees is used as the feature selection algorithm and 18 optimal features are selected among 61 original features. The graph represents the features from least important in the left to the most important in the right.

PCA

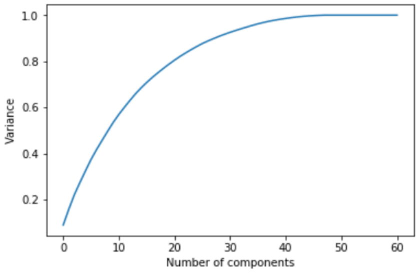
Appliances Energy

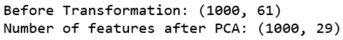
German Credit





```
Before transformation: (19735, 27)
Number of features after PCA: (19735, 9)
```





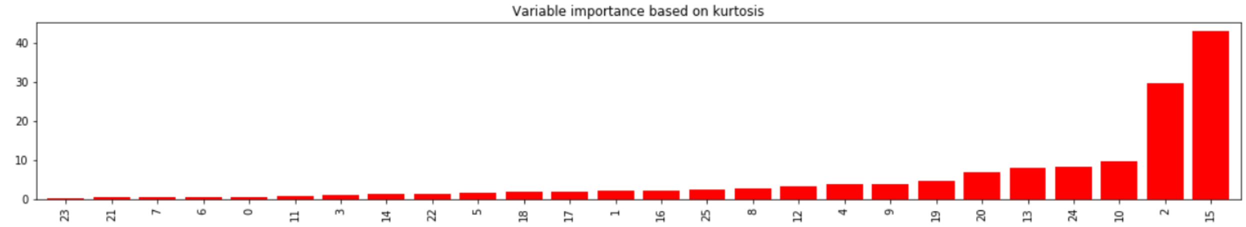
```
Before Transformation: (1000, 61)
Number of features after PCA: (1000, 29)
```

PCA generates orthogonal features. For the Appliances Energy 27 original features are transformed to 9 new features and for German Credit 61 of the original features are transformed to 29 new features

ICA

ICA retains all the components based on the kurtosis range of -1 to 1.

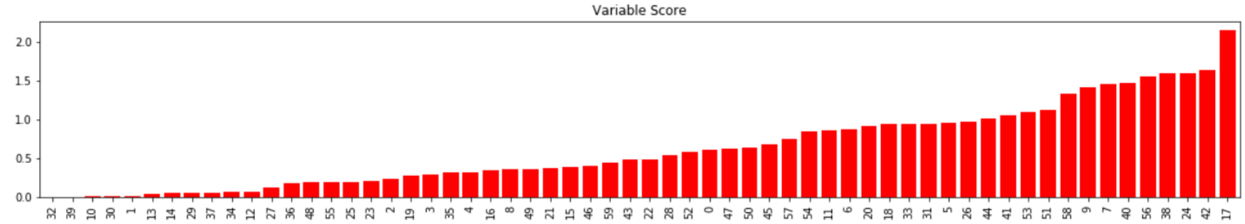
Appliances Energy



```
Number of features after reduction (19735, 6)
```

Based on the kurtosis 6 features starting from the left corner have been retained(23,21,7,6,0,11)

German Credit



```
Number of features after ICA (1000, 47)
```

Based on the kurtosis 47 features starting from the left corner out of the 61 original features

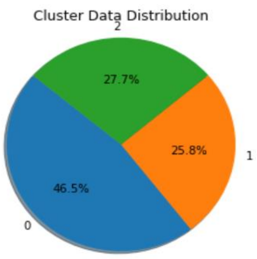
Randomized Projections

Random weights have been used to generate the features and approximately half of the features have been considered as optimal components beforehand. For Appliances Energy dataset 12 features have been retained out of the 27 original features and for the German Credit dataset 30 features have been retained out of the 61 original features

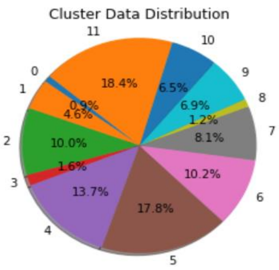
<u>Clustering after Dimensionality Reduction</u>

Appliances Energy

Feature Selection - Decision Tree

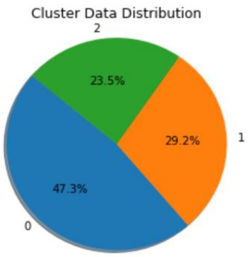


KMeans

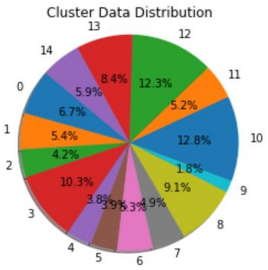


EM

PCA

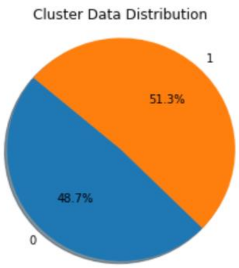


KMeans

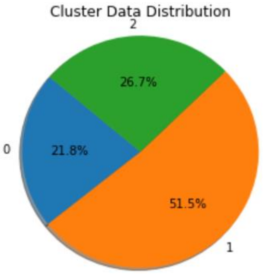


EM

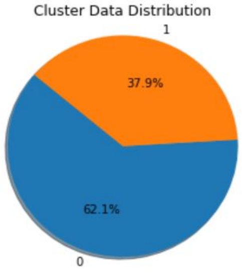
ICA RA



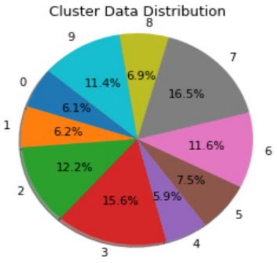
KMeans



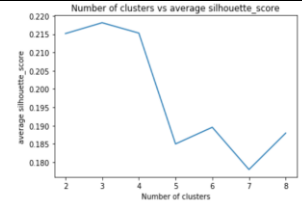
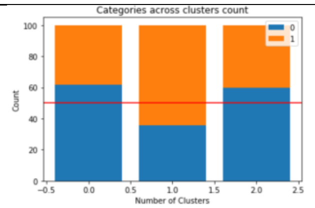
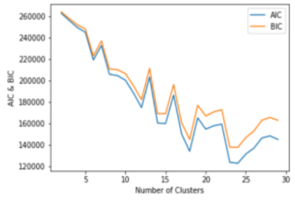
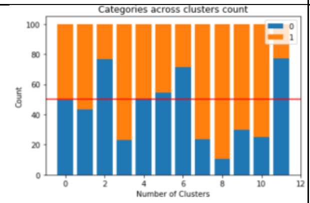
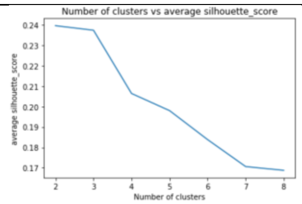
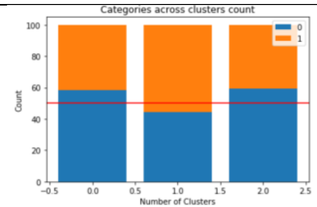
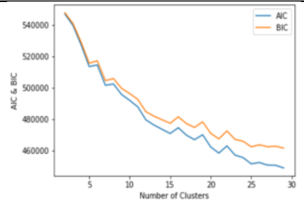
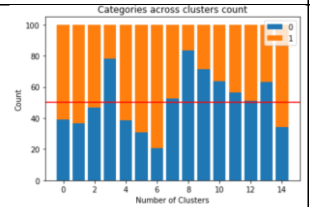
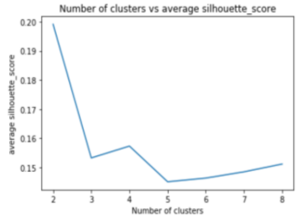
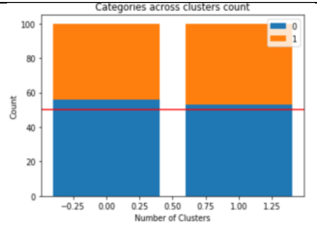
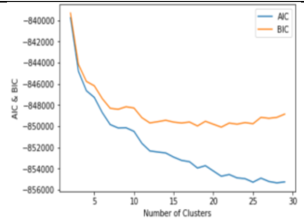
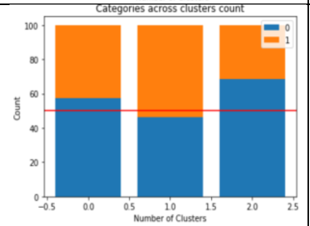
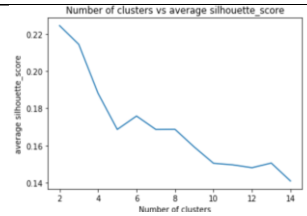
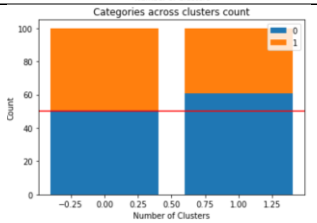
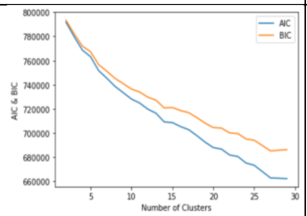
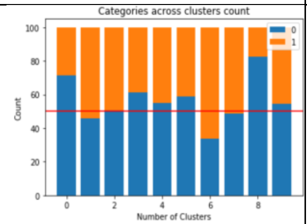
EM



KMeans



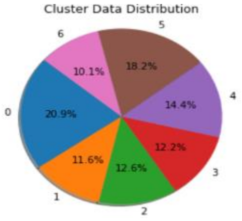
EM

Algo	KMeans Clusters	KMeans Categories	EM Clusters	EM Categories	Optimal Clusters
FS – Decision Tree					KMeans:3 EM:12
PCA					KMeans:3 EM:15
ICA					KMeans:2 EM:3
RA					KMeans:2 EM:10

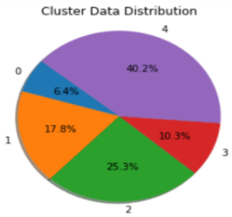
ICA has the least number of optimal clusters with respect to both KMeans and Expectation Maximization. Out of all the other dimension reduction algorithms, ICA retained least number of features(6). The class labels across all the clusters are approximately equally occupied by the observations.

German Credit

Feature Selection – Decision Tree

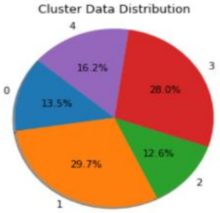


KMeans

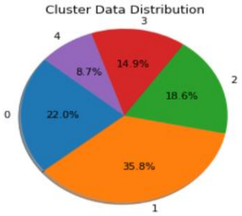


EM

PCA

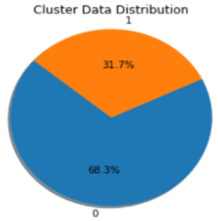


KMeans

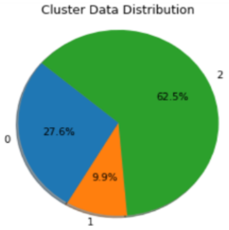


EM

ICA

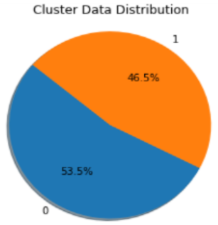


KMeans

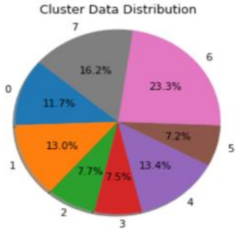


EM

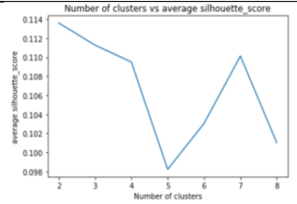
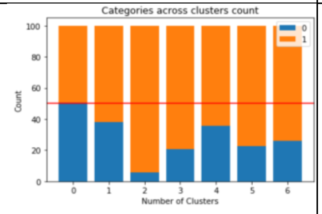
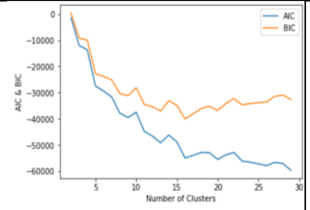
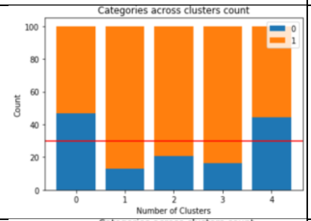
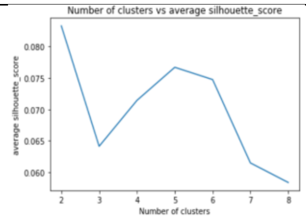
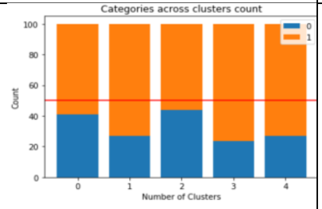
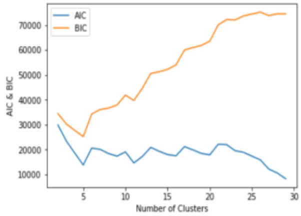
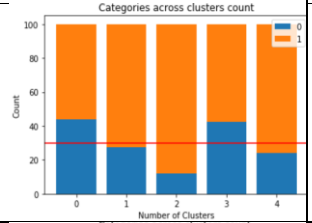
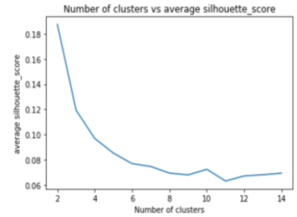
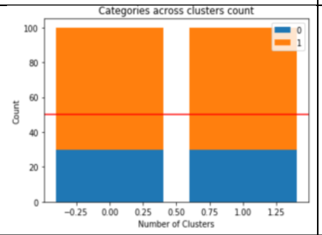
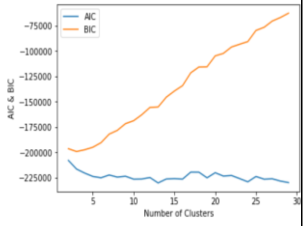
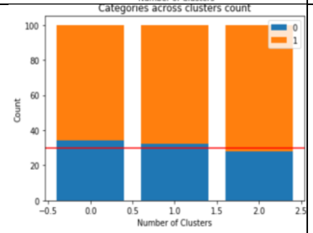
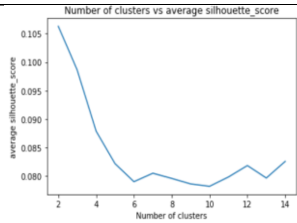
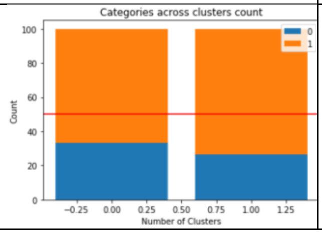
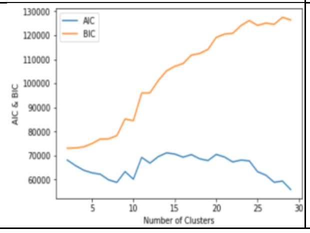
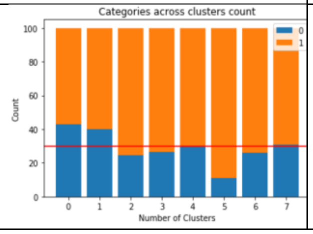
RA



KMeans



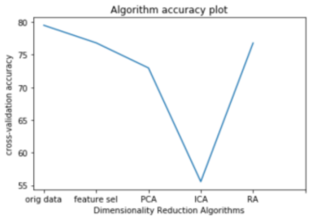
EM

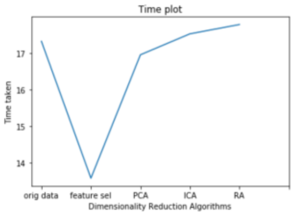
Algo	KMeans Clusters	KMeans Categories	EM Clusters	EM Categories	Optimal Clusters
FS – Decision Tree					KMeans:7 EM:5
PCA					KMeans:5 EM:5
ICA					KMeans:2 EM:3
RA					KMeans:2 EM:6

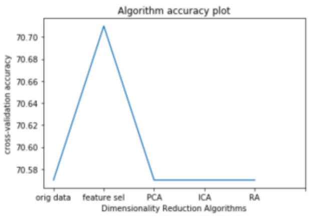
ICA and RA have the least number of optimal clusters with respect to both KMeans and ICA has the least number of optimal clusters with respect to Expectation Maximization. Out of all the other dimension reduction algorithms, Feature Selection – Decision Trees Classifier retained least number of features(18).

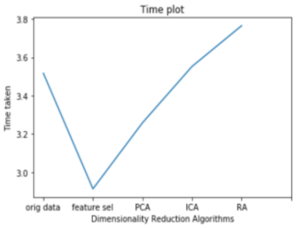
Neural Networks

Appliances Energy *German Credit*









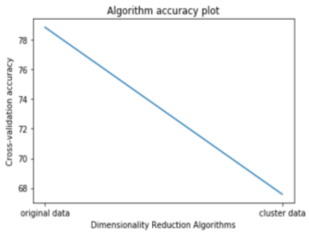
Neural Networks algorithm is applied using the optimal parameters obtained in the previous problem where hyperparamaters are tuned. Based on the cross validation accuracy graphs are plotted for the experimentation done using orginal data, feature selection, pca,ica and randomized projection algorithms.

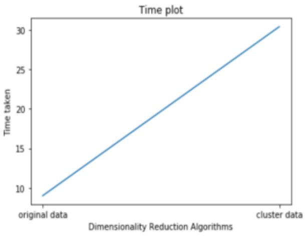
For the Appliances Energy Dataset, Original dataset performs better than the rest of the algorithms when cross validation accuracy is considered and randomized projection algorithm took the highest amount of time out of all the other algorithms. This in turn helps us prove that there is no problem of curse of dimensionality in the neural nets since the best cross validation accuracy score if for the original data set which has the highest number of features. Randomized projection algorithm performed second best.

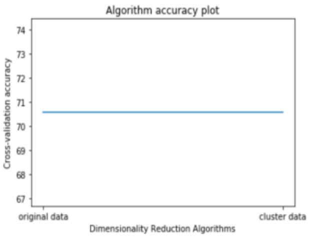
For German Credit Dataset, feature selection is the only algorithm that performs the best and the rest of the algorithms have same accuracies. Randomized projection algorithm took the highest amount of time when compared to other algorithms.

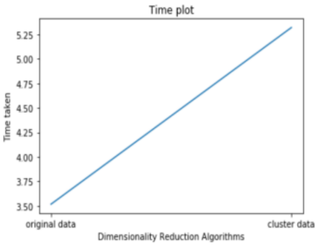
Clustering with Neural Networks

Appliances Energy *German Credit*









For Appliances Energy, optimal number of clusters with k-means is 3 and with EM is 13 (best cross validation accuracy when worked with the original dataset). 14 columns are created leaving one for k-means since it is the hard cluster the rest for the probabilty values. For German Credit, optimal number of clusters with k-means is 7 and with EM is 5 (best cross validation accuracy for the feature selection). 6 columns are created leaving one for k-means and the rest for the probability values.