# BUAN 6341.501: APPLIED MACHINE LEARNING

## 

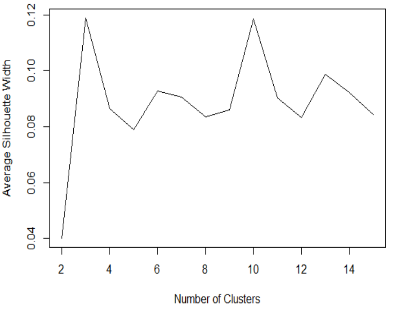
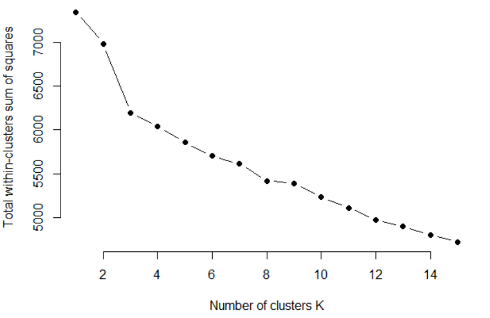
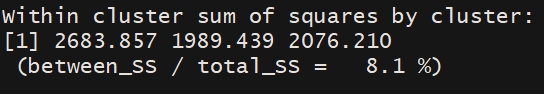
# Run Clustering Algorithms using the original set of Features:

## Data-Set 1: Car evaluation problem:

**K-MEANS:**

Elbow and Silhouette methods have been used to find the optimal number of clusters to avoid overfitting.

**Elbow Curve** **Silhouette Curve**

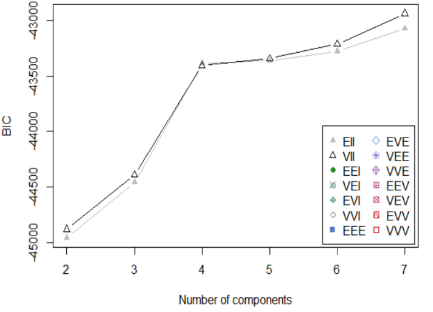
**Choosing K:** Optimum number of clusters using both the methods is **3**. Using Elbow Curve, we notice that the total WSS value doesn’t decrease drastically with K > 3. Also in Silhouette Curve, K=3 model achieves the highest score. We see that about **8%** of the variance in the Data has been explained by this model.

**The within cluster sum-of-squares is a measure of the variability of observations within each cluster. In general, the cluster that has a small sum-of squares is more compact than a cluster that has a large sum-of-squares.**

**CLUSTER 2 is more COMPACT**

**Expectation Maximization:**

As the distribution of this data is Uniform, MClust returned 1 Cluster as Optimum. So we ran the algorithm by passing K values explicitly ranging from 2 to 7. Based on the highest BIC value, we obtained 7 as the optimum number of clusters.

**BIC (Bayesian Information Criterion) Curve**

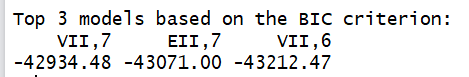
Cluster Statistics

Within cluster SS = 5724,

Average silhouette width = 0.06327032,

average between = 2.963151,

average within = 2.594822



**Note:** The clusters did not line-up as per the actual class labels. They were formed naturally.

## Dimensionality Selection

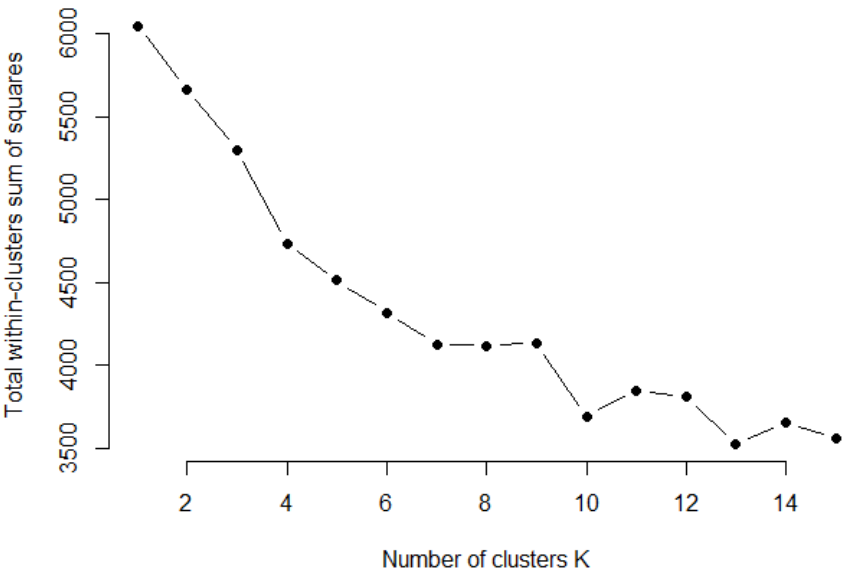
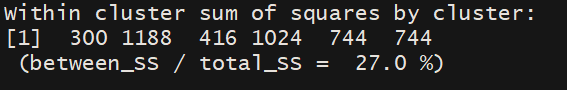
We used **Decision Tree for feature selection** algorithm. It helped us to find the significant variables that was enough to get the desired output.



**K-Means using Feature Selection**

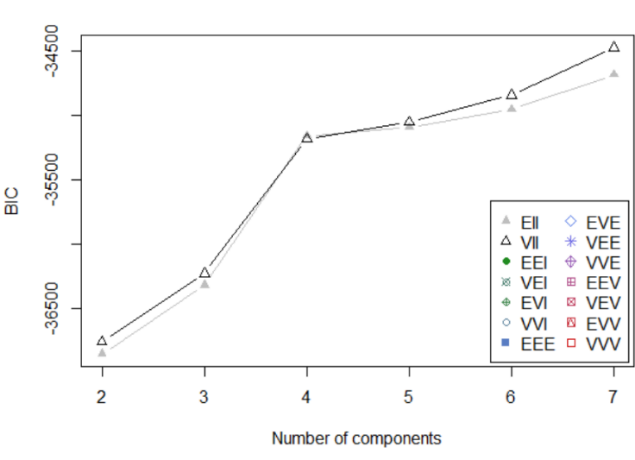
Next, we ran K-means using the significant variables that we got from the decision tree.

**Elbow Curve Cluster Variation**

We can see that the optimal cluster K = 6.

**EM Using Feature Selection**

**BIC Curve**

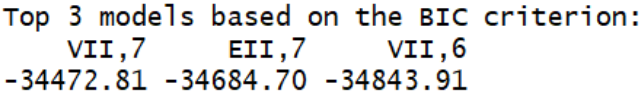
**Cluster Statistics:**

within cluster SS = 4428

Average silhouette width = 0.08577556

average between = 2.698067

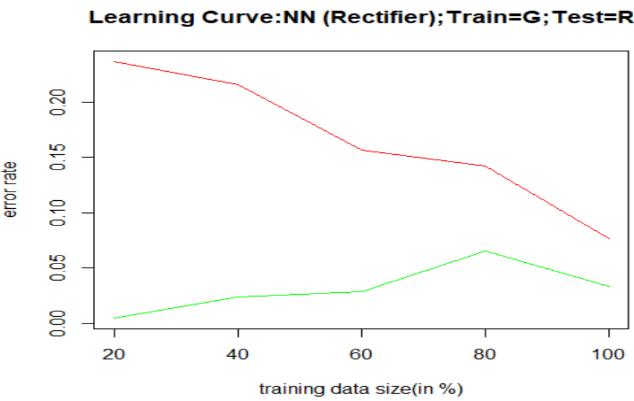
average within = 2.277841



We chose the model with 7 clusters. We see that the BIC here is maximum.

**Neural network Using Feature Selection**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| Feature Selection | 0.033849 | 0.065089 | 93.49% |

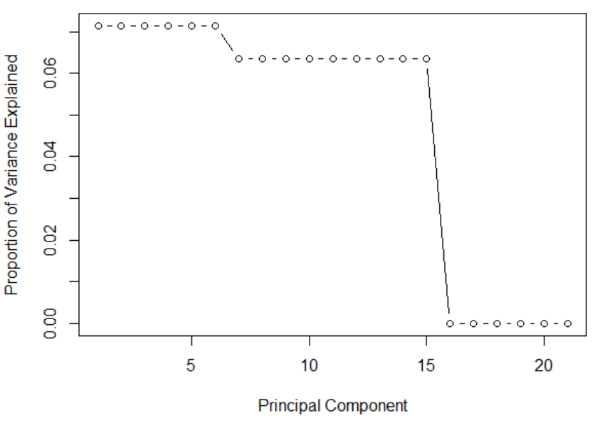


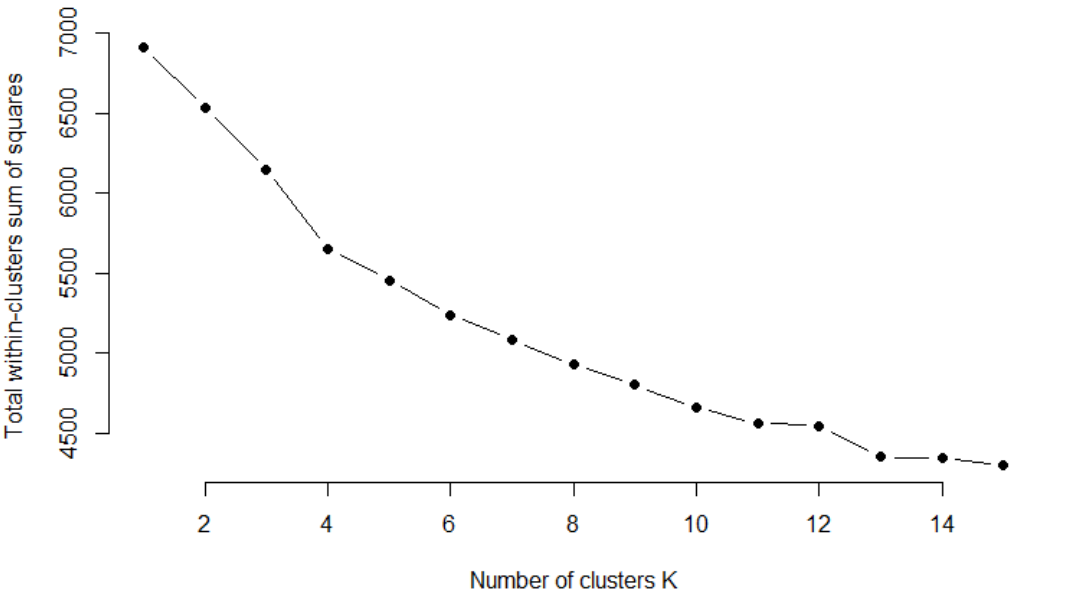
We can see the accuracy of the neural net is **93.49%** after using the significant variables obtained through Feature Selection.

**Note** that there is **a 3% dip** in the accuracy as compared to the original model run in Project 3.

## Dimensionality Transformation:

### PCA:

From the scree plot, the number of principal components chosen is **14**. These 14 components explain the maximum variance (in data) of about **90%  
K-Means using PCA**

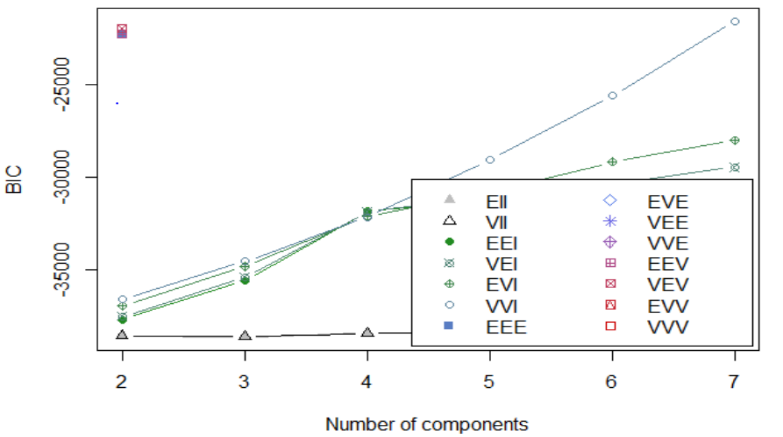
 Within cluster sum of squares by cluster:

[1] 1897 1820 1965

(between\_SS / total\_SS = 19.7 %)

Optimal K=3. There is no difference in the number of clusters when run with original features and when run on the reduced dimensions after PCA.

**Expectation Maximization on PCA dimensions**

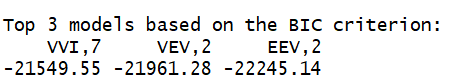
 **Mclust - BIC curve**

**Cluster Statistics**

Within cluster SS = 5294.206,

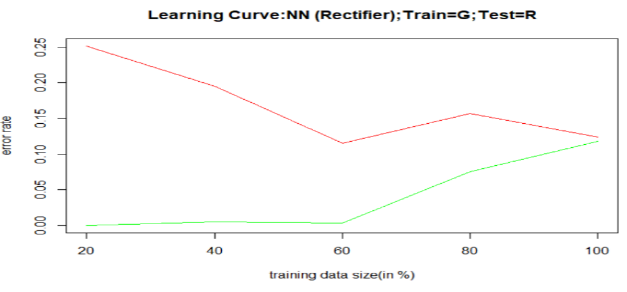
Average silhouette width = 0.06834474

Average between = 2.876747

Average within = 2.49422 We chose the model with 7 clusters. We see that the BIC here is maximum.

**Neural network Using PCA**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| PCA | 0.058027 | 0.097633 | 90.24% |



We can see the accuracy of the neural net is **90.24%** after using the first 14 Principal components from PCA.

**Note** that there is **a 6% dip** in the accuracy as compared to the original model run in Project 3.

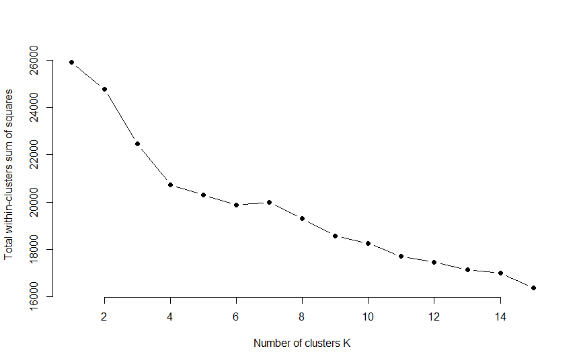
### ICA:

We used the package **fastICA** to compute the Independent components from the entire set of original features.

And then ran decision tree over the generated ICs to find the significant subset of features.

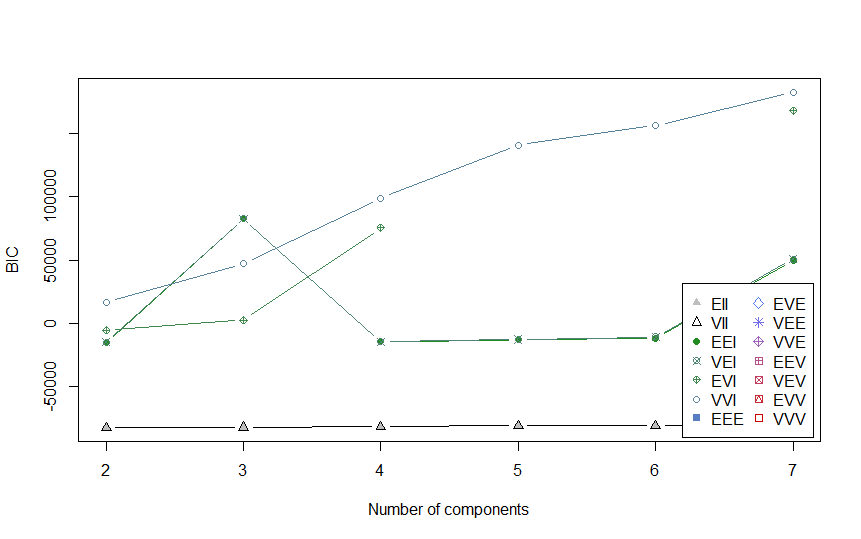
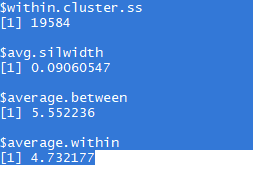
Chose **18 components** after using feature selection.

**K-Means using ICA**

From the above elbow curve, we observe that the optimal no. of clusters is **4** as opposed to 3(in original K-Means model)

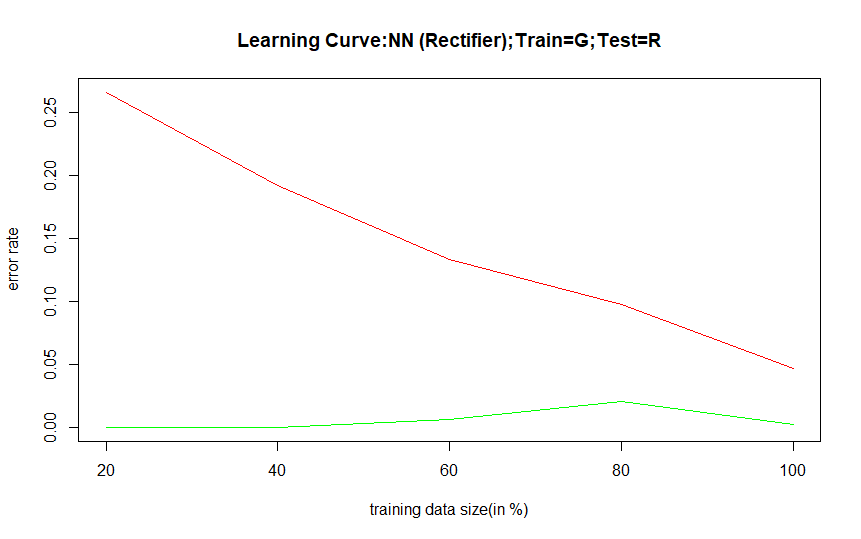
**Expectation Maximization on ICA dimensions**

We chose the model with 7 clusters. We see that the BIC here is maximum.

**Neural network Using ICA**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| PCA | 0.009671 | 0.038462 | 96.15% |



We can see the accuracy of the neural net is **96.15%** after using the 18 Principal components from ICA.

**Note** that this is the exact accuracy as compared to the original model run in Project 3. No performance improvements.

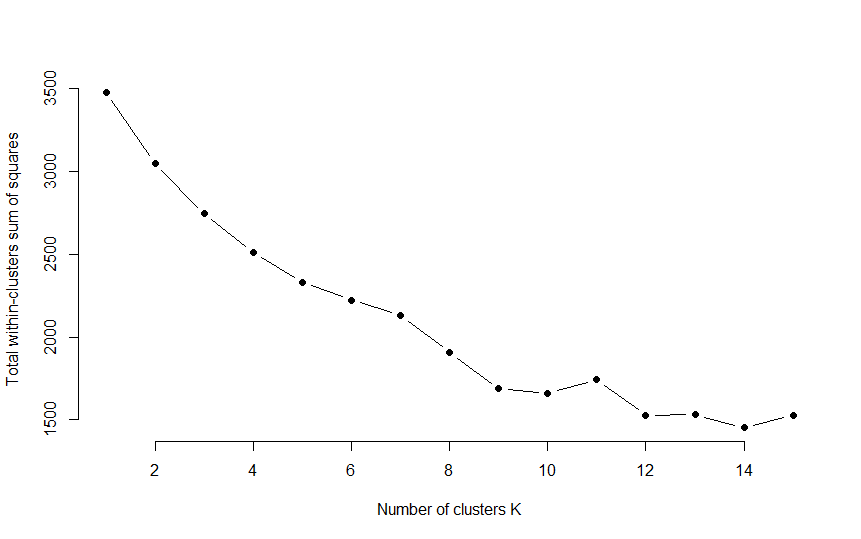
### RCA:

We implemented the code for RCA and generated same number of Random Components from the original features set.

And then ran decision tree over the generated RCs to find the significant subset of features.

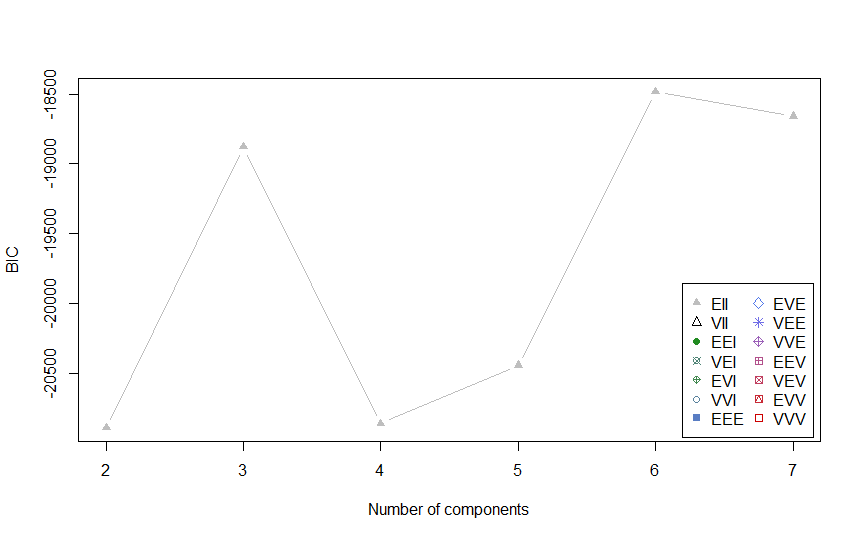
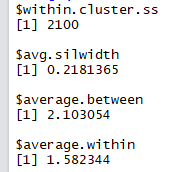
Chose **10 components** after using feature selection.

**K-Means using RCA**

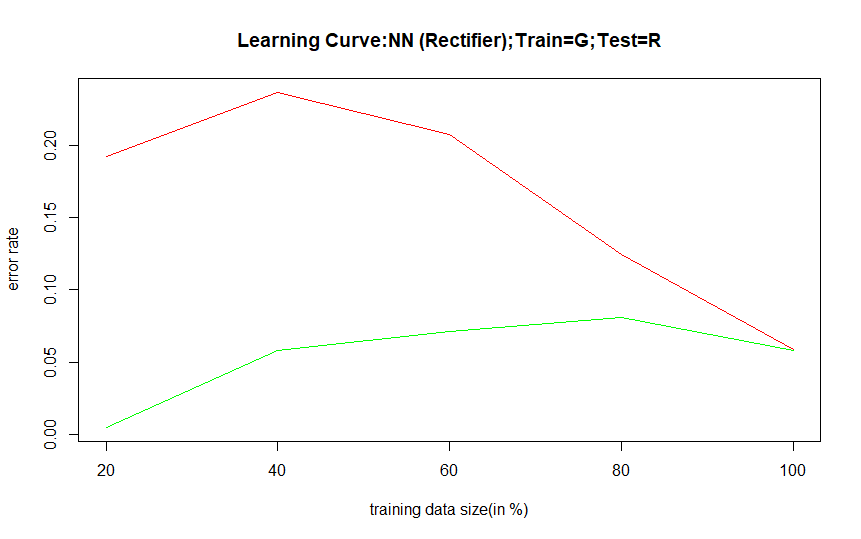
From the above elbow curve, we observe that the optimal no. of clusters is **7** as opposed to 3(in original K-Means model)

**Expectation Maximization on RCA dimensions**

    
We chose the model with **6** clusters. We see that the BIC here is maximum.

**Neural network Using RCA**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| PCA | 0.047389 | 0.050296 | 94.97% |

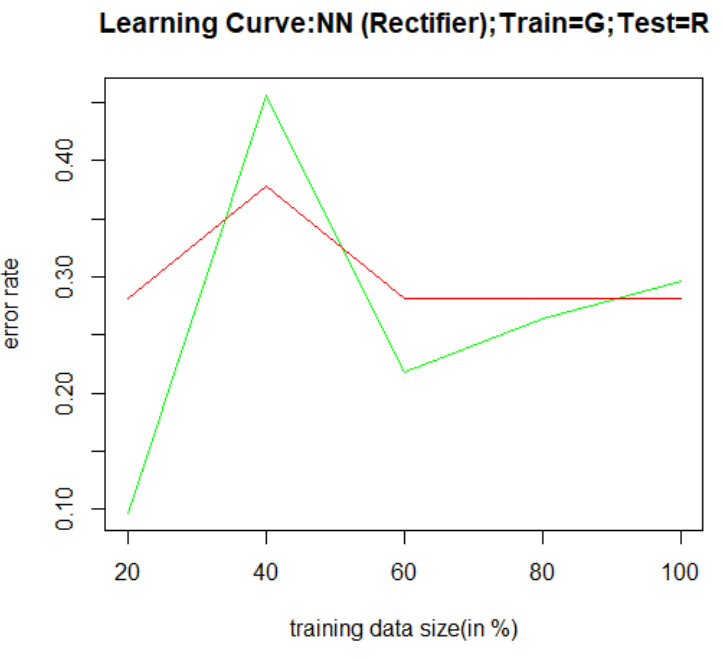
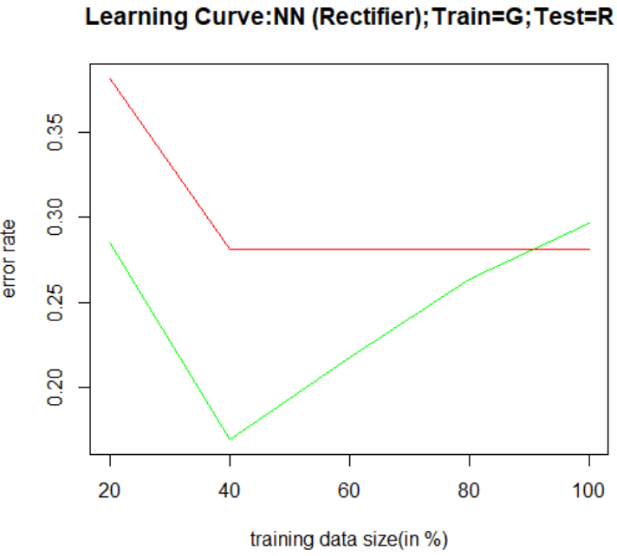


We can see the accuracy of the neural net is **94.97%** after using the 10 Principal components from RCA.

**Note** there is a dip of 1% accuracy as compared to the original model run in Project 3. No performance improvements.

## Clustering output as input features to the Neural Networks

**Neural Net on KMeans Clusters Neural Net on EM Clusters**



|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| K-Means | 0.441006 | 0.467456 | 53.25% |
| Expectation Maximization | 0.296905 | 0.281065 | 71.89% |

We can see that the EM features perform better in classifying the output labels using Neural Networks.

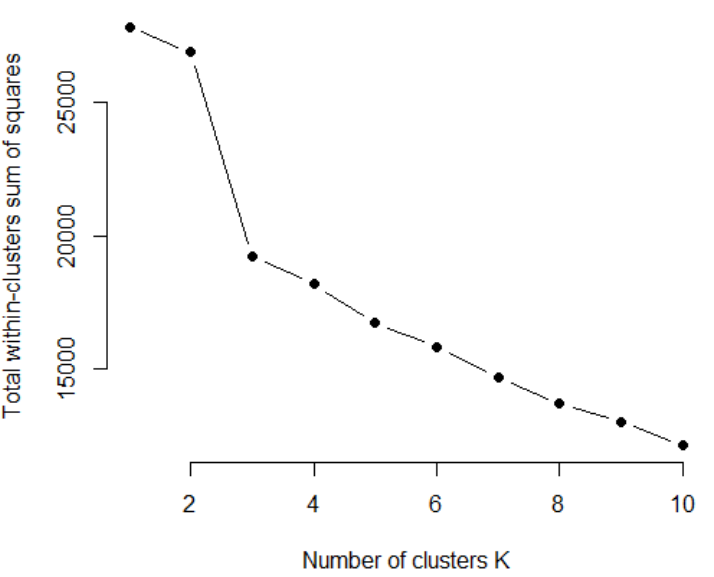
However, there is no real performance improvements in using such clustering features as input for neural networks since the original model’s accuracy was much higher.

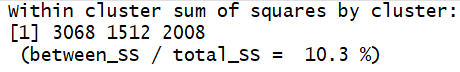
## Final Comparison of Models:

|  |  |
| --- | --- |
| **Model** | **Best Version** |
| Kmeans (Highest Variance Explained) | RCA Features (45.6%) |
| EM (High BIC Value) | RCA Features (-18486) |
| Neural Net (High Accuracy) | ICA Features (96.15%) |

## Data-Set 2: HR Analytics Dataset:

**K-MEANS:**





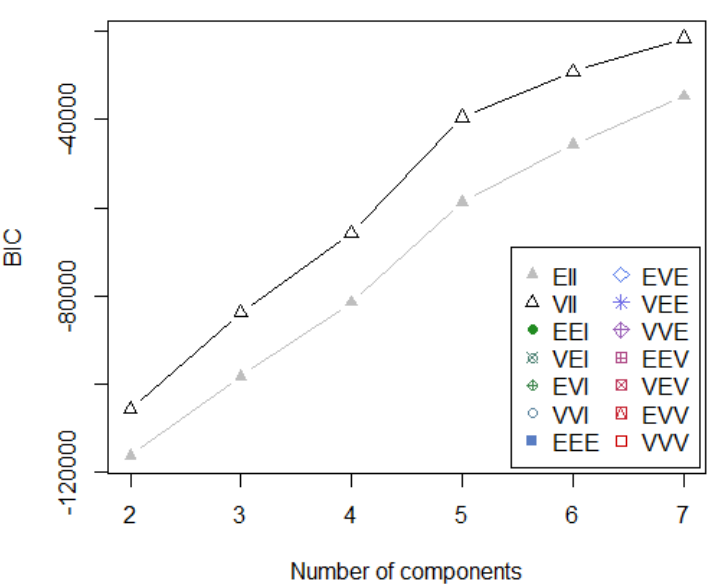
**Choosing K:** Optimum number of clusters using both the methods is **3**. Using Elbow Curve, we notice that the total WSS value doesn’t decrease drastically with K > 3. We see that about **10%** of the variance in the Data has been explained by this model.

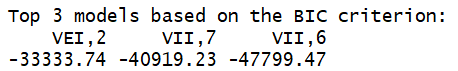
**The within cluster sum-of-squares is a measure of the variability of observations within each cluster. In general, the cluster that has a small sum-of squares is more compact than a cluster that has a large sum-of-squares.**

**CLUSTER 2 is more COMPACT**

**Expectation Maximization:**

As the distribution of this data is Uniform, MClust returned 1 Cluster as Optimum. So, we ran the algorithm by passing K values explicitly ranging from 2 to 7. Based on the highest BIC value, we obtained 7 as the optimum number of clusters.





**Cluster Statistics**

Within cluster ss = 17500.46

Average silhouette width = 0.2707224

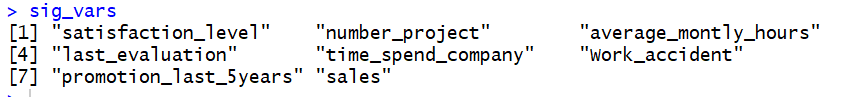
Average between = 1.979309

Average within = 1.448658

**Note:** The clusters did not line-up as per the actual class labels. They were formed naturally.

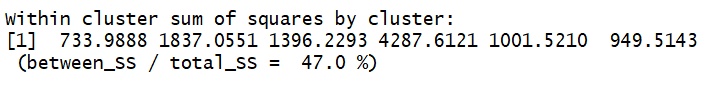
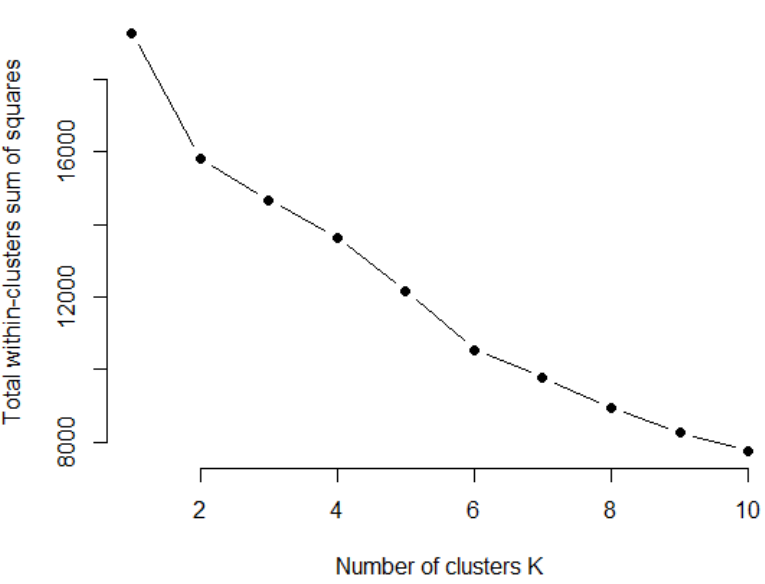
## Dimensionality Selection

We used **Decision Tree for feature selection** algorithm. It helped us to find the significant variables that was enough to get the desired output.



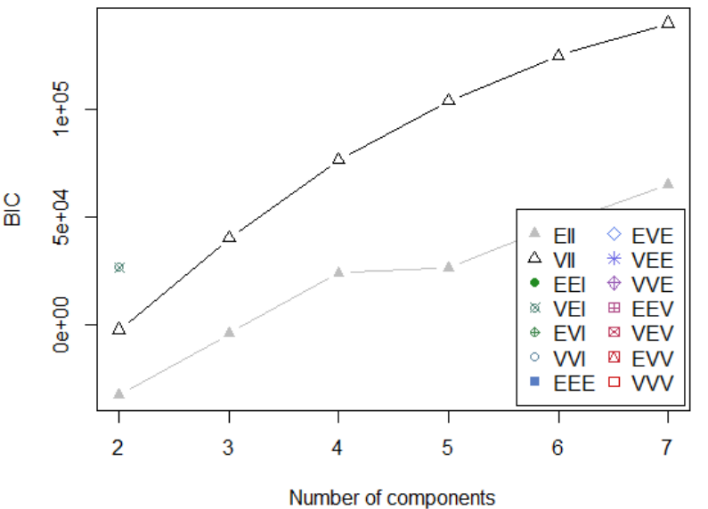
**K-Means using Feature Selection**

Next, we ran K-means using the significant variables that we got from the decision tree.



We can see that the optimal cluster K = 6.

**EM Using Feature Selection**

**BIC Curve**

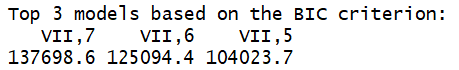
**Cluster Statistics:**

Within cluster ss = 11619.38

Average silhouette width = 0.1866322

Average between = 1.616654

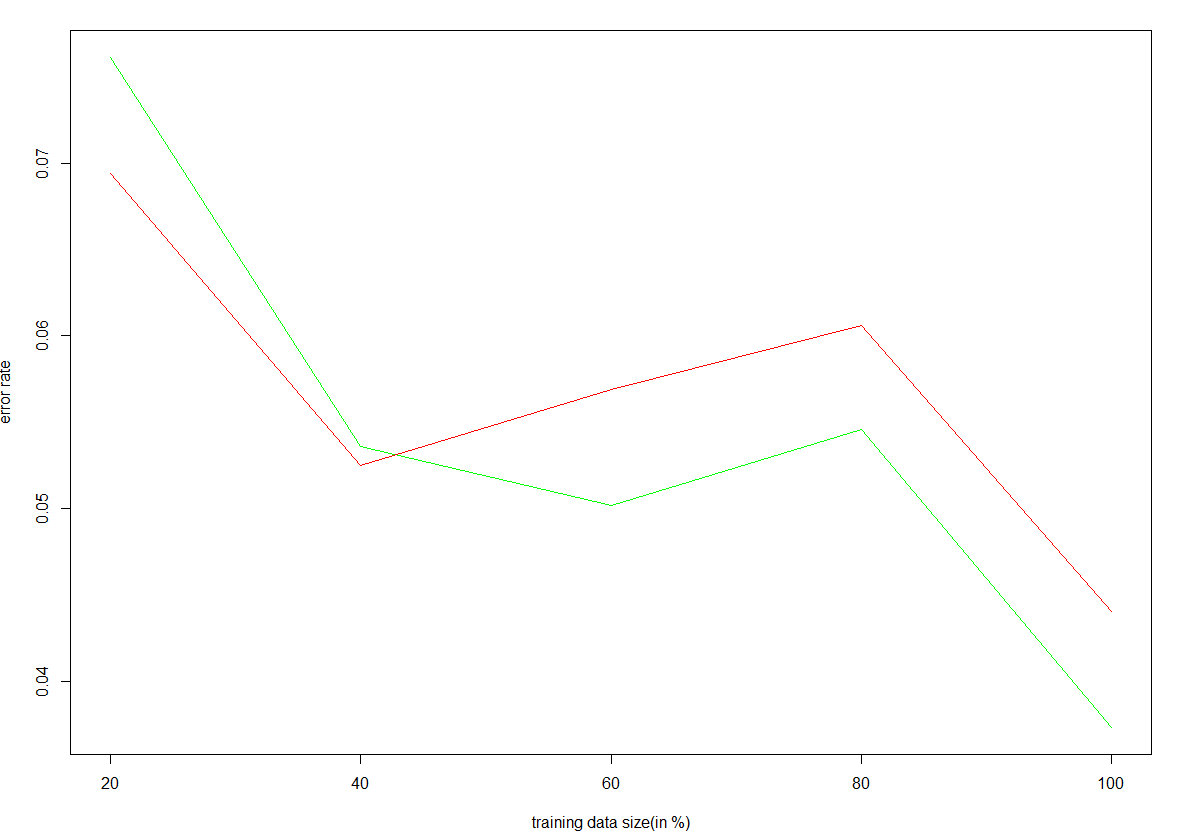
Average within = 1.400536



We chose the model with 7 clusters. We see that the BIC here is maximum.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| Feature Selection | 0.036226 | 0.037928 |  |

**Neural network Using Feature Selection**

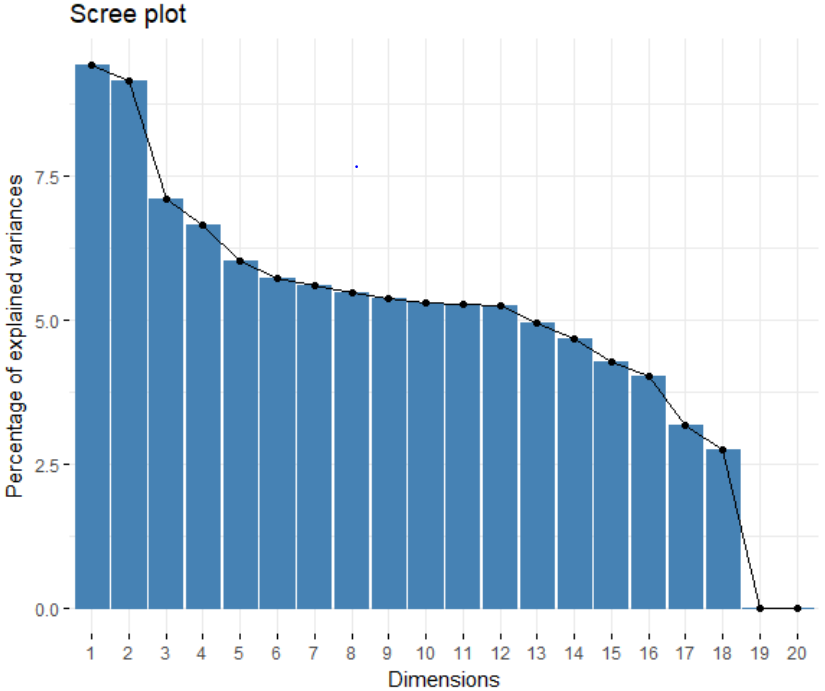


|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| Feature Selection | 0.036226 | 0.037928 | 96.21% |

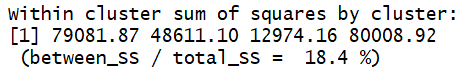
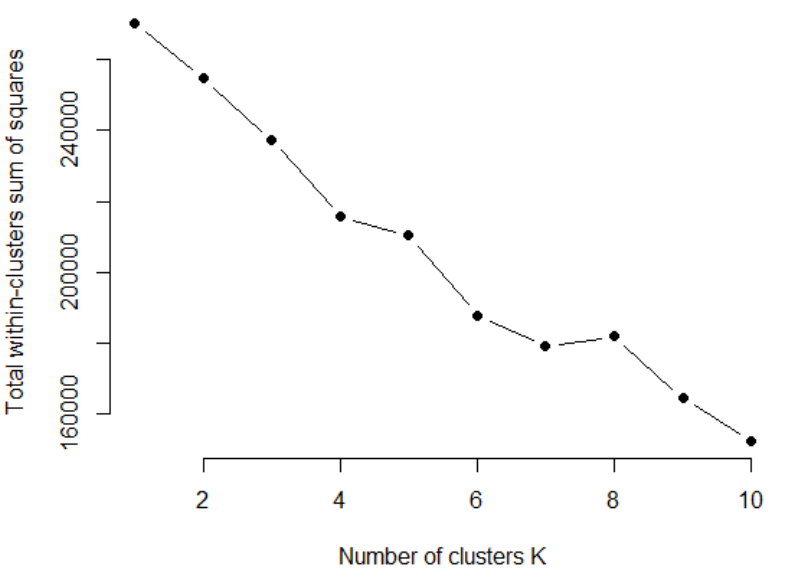
We see that there is no change in the accuracy of the neural network model even after using the significant variables obtained through Feature Selection.

## Dimensionality Transformation:

**PCA:**

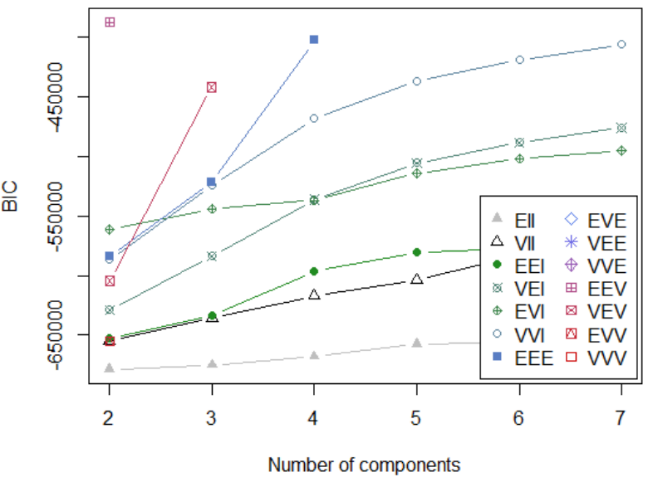
From the scree plot, the number of principal components chosen is **15**. These 14 components explain the maximum variance in data is about 90%

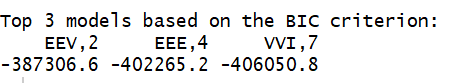
**PCA on Kmeans**



Optimal K=4. The number of clusters is increased by 1 compared to the original features, when run on the reduced dimensions after PCA.

**Expectation Maximization on PCA dimensions**

**BIC Curve**



**Cluster Statistics:**

Within cluster ss = 253338.7

average silhouette width = 0.1349801

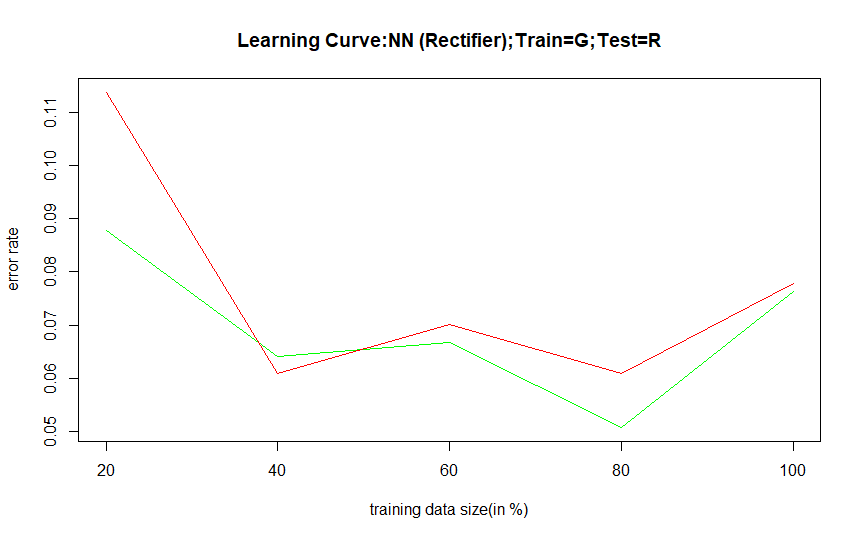
Average between = 6.362609

Average within = 5.319648

We chose the model with 7 clusters. We see that the BIC here is maximum.

**Neural network Using PCA**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| PCA | 0.048339 | 0.048764 | 95.12% |



We can see the accuracy of the neural net is **95.12%** after using the first 15 Principal components from PCA.

**Note** that there is **a 1% dip** in the accuracy as compared to the original model run in Project 3.

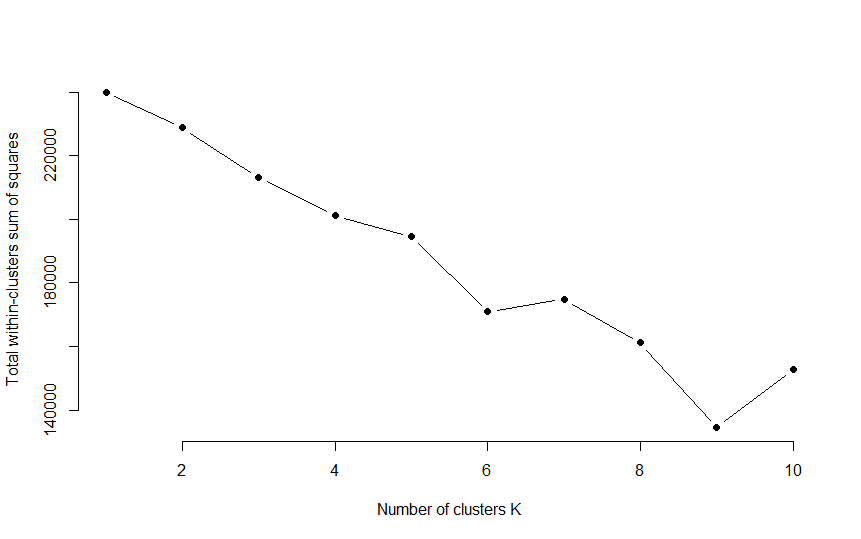
### ICA:

We used the package **fastICA** to compute the Independent components from the entire set of original features.

And then ran decision tree over the generated ICs to find the significant subset of features.

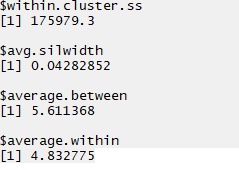
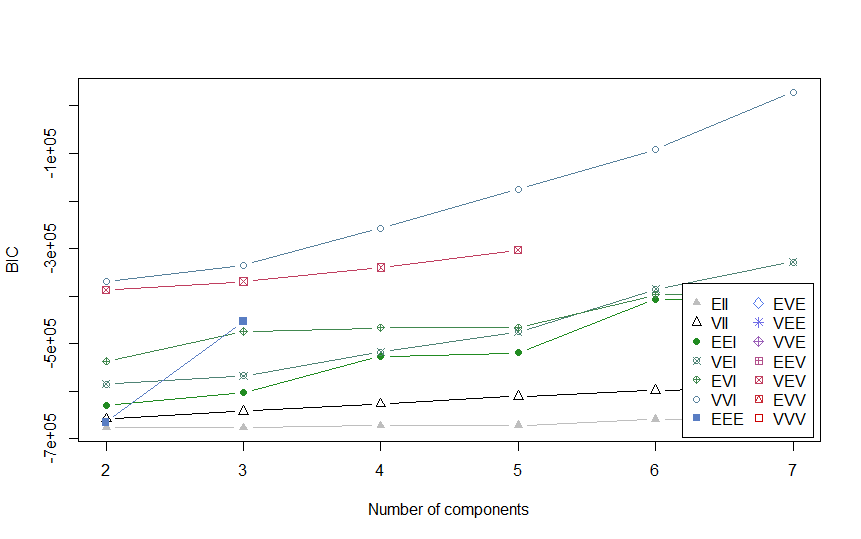
Chose **16 components** after using feature selection.

**K-Means using ICA**



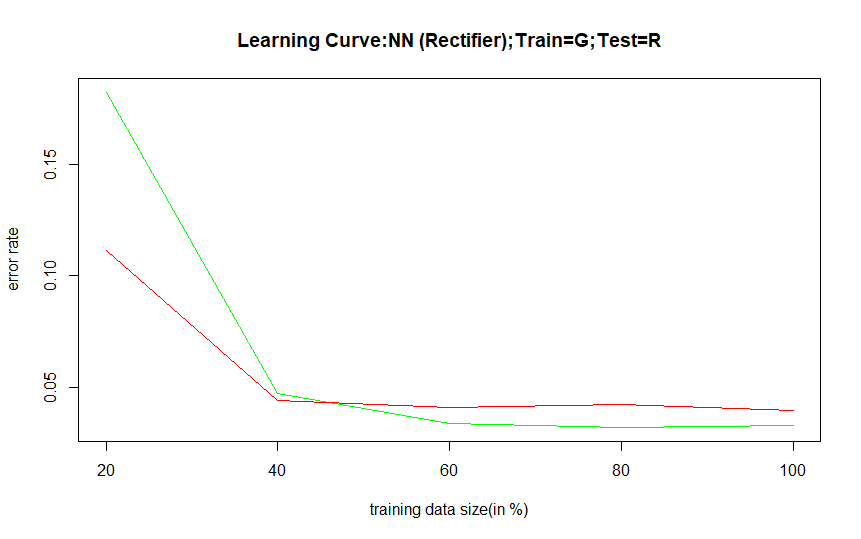
From the above elbow curve, we observe that the optimal number of clusters is **6** as in original KMeans model.

**Expectation Maximization on ICA dimensions**



We chose the model with 7 clusters. We see that the BIC here is maximum.

**Neural network Using ICA**



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| PCA | 0.032448 | 0.033864 | 96.61% |

We can see the accuracy of the neural net is **96.61%** after using the 16 components from ICA.

**Note** that this is slight improvement as compared to the original model run in Project 3.

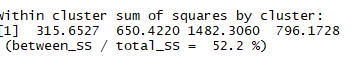
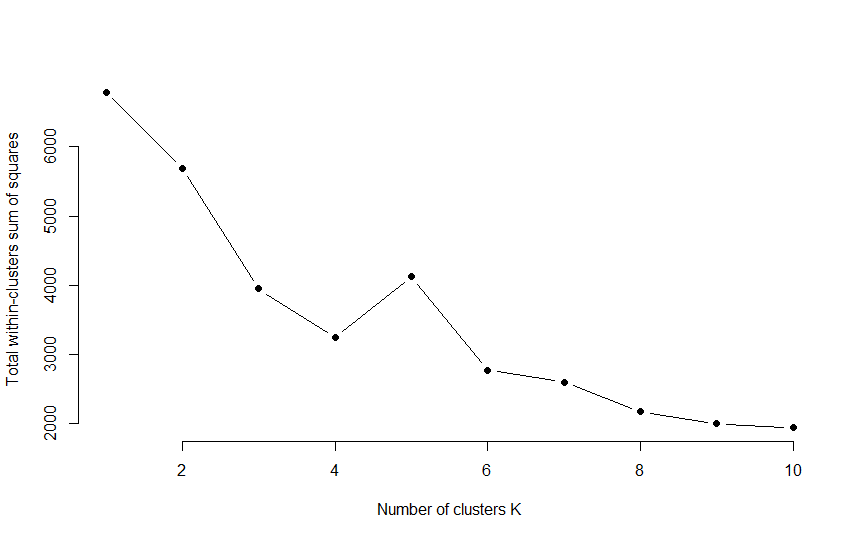
### **RCA:**

We implemented the code for RCA and generated same number of Random Components from the original features set.

And then ran decision tree over the generated RCs to find the significant subset of features.

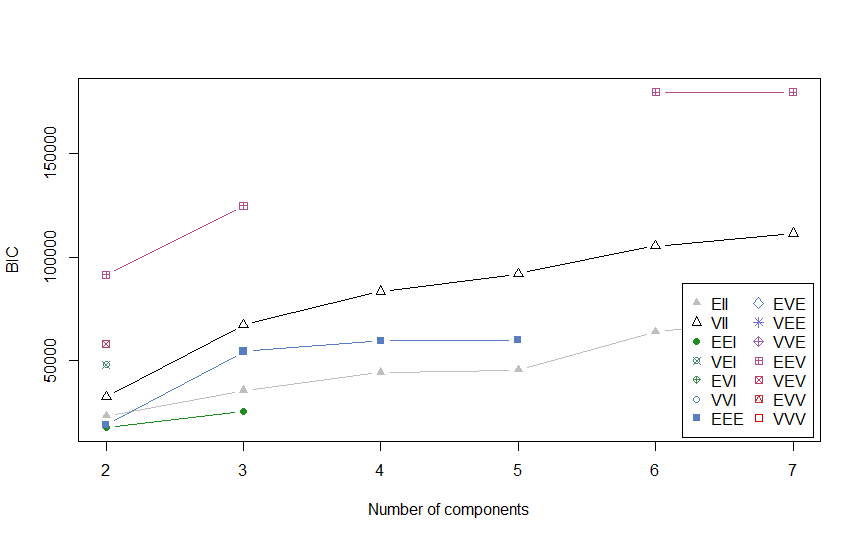
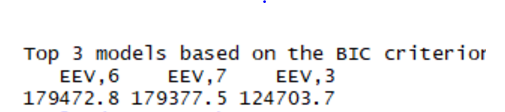
Chose **7 components** after using feature selection.

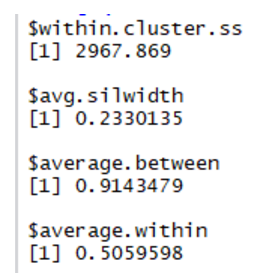
**Kmeans using RCA**



From the above elbow curve, we observe that the optimal number of clusters is **4** as opposed to 6 in original KMeans model.

**Expectation Maximization for RCA**

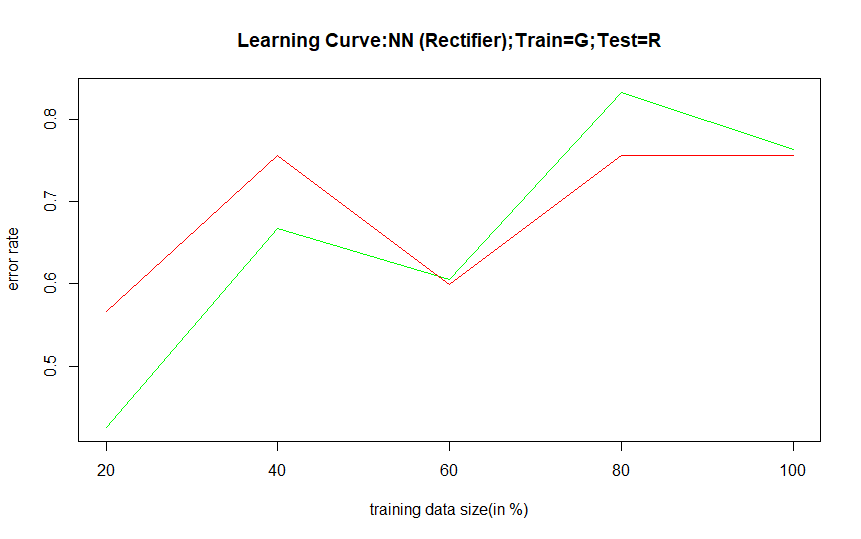
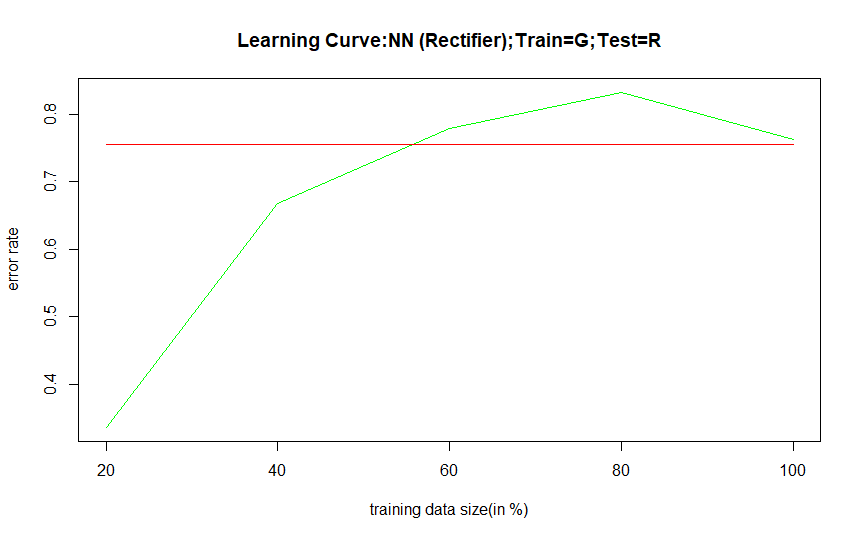
 



We chose the model with 7 clusters. We see that the BIC here is maximum.

# Clustering output as input features to Neural Network

**Neural Network on Kmeans Neural Network on EM**



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Train Error** | **Test Error** | **Accuracy on Test Set** |
| Kmeans features | 0.762974 | 0.755164 | 24.49% |
| Expectation Maximization | 0.629848 | 0.636302 | 36.37% |

We see that the Kmeans features perform better than EM features in classifying the output labels using Neural Network. However, there is no real performance improvement using clustering features as input for neural networks since the original model accuracy was much higher.

## Final Comparison of Models:

|  |  |
| --- | --- |
| **Model** | **Best Version** |
| Kmeans (Highest Variance Explained) | RCA Features (52.2 %) |
| EM (High BIC Value) | RCA Features (179472.8) |
| Neural Net (High Accuracy) | Feature Selection (96.21%) |