PROJECT 4

Team Members: Aji Somaraj, Neeraja Anil

DECEMBER 8, 2017

NAVEEN JINDAL SCHOOL OF MANAGEMENT
UNIVERSITY OF TEXAS AT DALLAS

Dataset Description

Adult Income Dataset is a subset from the 1994 US Census form UCI data repository which has a label indicating if an "individual" has an income greater than $50,000 a year, along with other census data of the individual such as education, heritage and age (among others). It contains approximately 32000 observations, with 15 variables. Our objective is to implement clustering algorithms such as K-Means and Expectation Maximization to look for underlying patterns in the data as well as apply feature reduction algorithms to choose relevant features for the same, while exploring how clustering can improve the accuracy of supervised learning algorithms like Artificial Neural Networks.

Bank Marketing Dataset available at the UC Irvine Machine Learning Repository has a class label to indicate individuals who would subscribe a term deposit by using the data related with direct marketing campaigns of a Portuguese banking institution containing information such as customer data, campaign activities and social and economic environment data. It contains approximately 45000 observations with 12 categorical variables and 7 numerical variables as its features and a binary category as its response variable. Our objective is to implement clustering algorithms such as K-Means and Expectation Maximization to look for underlying patterns in the data as well as apply feature reduction algorithms to choose relevant features for the same, while exploring how clustering can improve the accuracy of supervised learning algorithms like Artificial Neural Networks.

Data Modeling

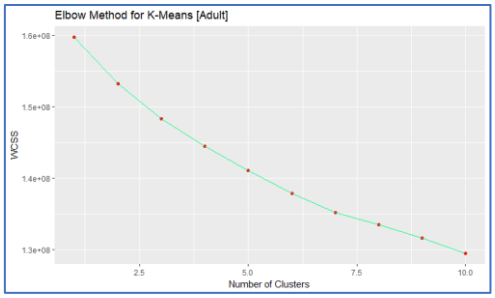
Clustering

While performing clustering analysis, one takes advantage of domain knowledge to choose the relevant features necessary for clustering. However, for this project, we experiment with the clustering algorithm with all the features in the dataset and then perform clustering again with only the features obtained after applying feature reduction algorithms like Backward Elimination, PCA, ICA and Random Projections. The clustering results are then applied to Artificial Neural Network to explore how clustering can affect the performance of supervised algorithms.

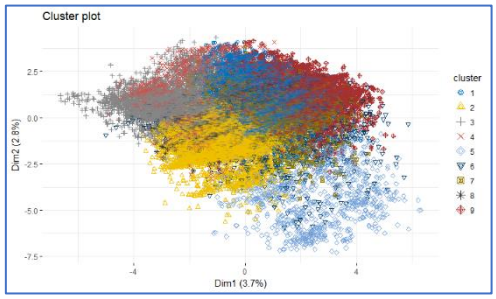
Dataset 1 [Adult]

K-Means

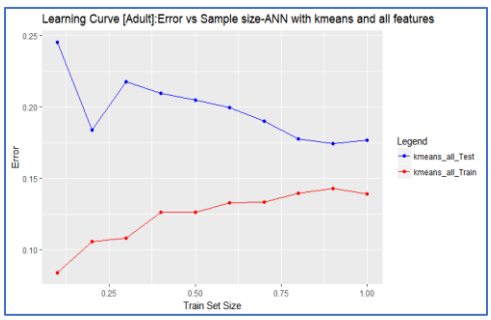
Adult dataset is mixture of numerical and categorical data. Since, K-Means can be applied only to numerical data, 1 hot encoding is performed on the categorical data to generate dummy variables to perform K-Means.



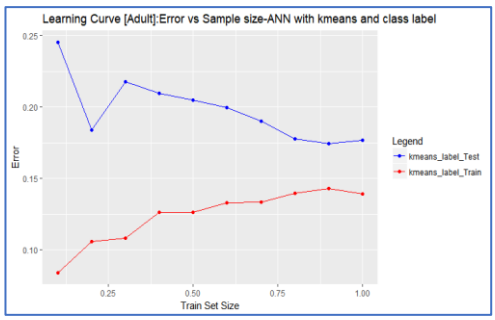
The [Adult] dataset doesn't result in well-defined clusters after performing K-Means. From the elbow plot, a dip in Within Cluster Squared Error is seen at k=9, which is taken for further clustering analysis. A scatter plot of the clustering in 1st two dimensions (explains 6.5% of variance) is shown above. The clusters overlap with each other and explains a very low percentage of variance in the first 2 principal components. This might be due to the mix of categorical and numerical data found in the dataset. A better approach to do clustering is to take k-prototypes which combines k-modes and k-means and can cluster mixed numerical / categorical data.



Artificial Neural Network On K-Means results

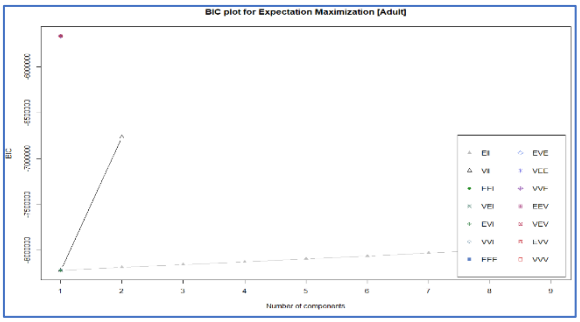


The results from K-Means clustering including all the features, cluster labels and class labels was applied to ANN model with 1 hidden layer and 8 nodes and the learning curves are shown on left. Moreover, K-Means cluster labels and class labels alone were applied as input to the same Neural Network and the learning curves are plotted as shown on right. The learning curve for error vs Sample Size for ANN with all features on K-Means results as well as ANN on only the cluster labels show similar performance. Initially when the train size is low, we can see that the errors are high for test and low for train and as the size increases, the train error goes up slightly and the test error comes down before reaching a plateau. As compared to the learning curves for ANN on original dataset, the fluctuations in the new learning curves have been smoothened out which implies a reduction in loss values and greater learning accuracy. The accuracy on ANN utilizing clustering results has gone up to 84.03% from 83.85% and AUC remains the same at 0.909 and the execution time required has gone down. Accuracy for ANN with only K-Means cluster labels and class labels is 70.64% and AUC is 0.7985.

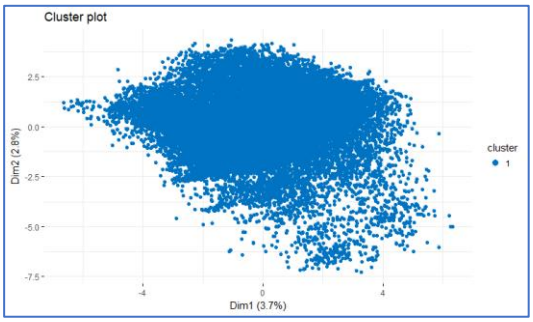


K-Means gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

Expectation Maximization

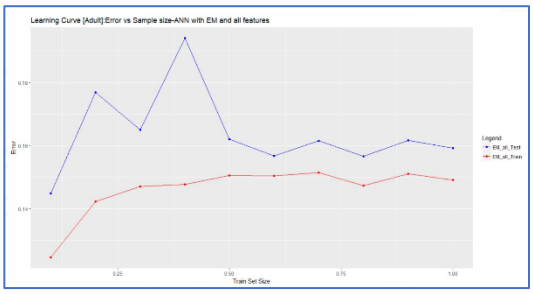


Expectation-maximization clustering probabilistically assigns data to different clusters. Ellipsoidal multivariate normal (XXX) with 1 component is found to be the optimal clustering model from BIC plot after finding out the BIC for various other models. Cluster plot for 1 cluster is shown.



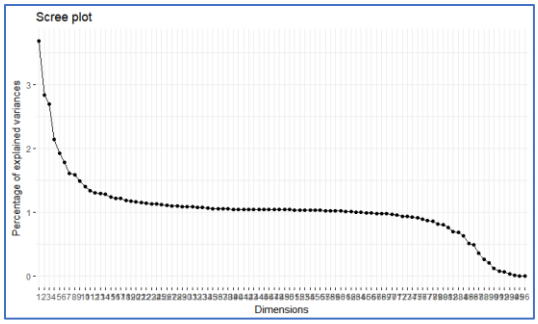
ANN on Expectation Maximization

Since EM model resulted in only 1 cluster, ANN with only cluster labels cannot be performed. However, on implementing, ANN with all features and cluster labels resulted in an increase in accuracy and AUC over ANN with original dataset. The learning curve on right shows that with increase in sample size, the test error comes down significantly as the algorithm becomes more confident in predicting correct output. An accuracy of 84.08% and AUC of 0.909 is obtained. EM gives only one cluster, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

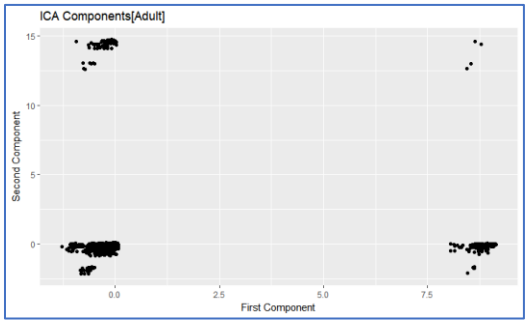


Dimensionality Reduction Algorithms

Feature selection using backward elimination as well as feature transformation using PCA, ICA and Random Projection was done on the dataset to reduce the number of dimensions of the dataset to improve interpretability, accuracy and visualization. Feature selection with backward elimination with logistic regression learner resulted in 14 features out of the total 15 features.



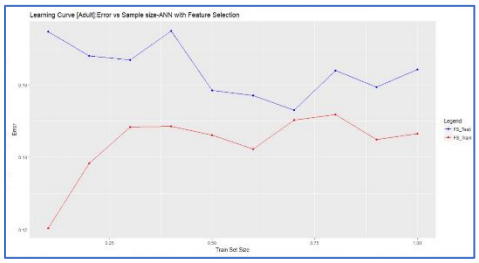
PCA transforms the variables to a new set of variables, which are orthogonal and ordered such that the retention of variation present in the original variables decreases as we move down in the order. PCA transforms the 96 features into 96 dimensions which explains 100% of the variance. From the scree plot for PCA shown on right, it can be obtained that 64 dimensions explains 80% of the total variation which is taken as the reduced number of dimensions from PCA for further analysis.

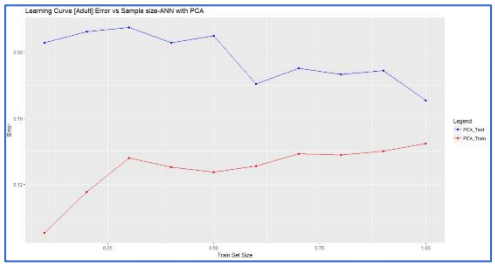


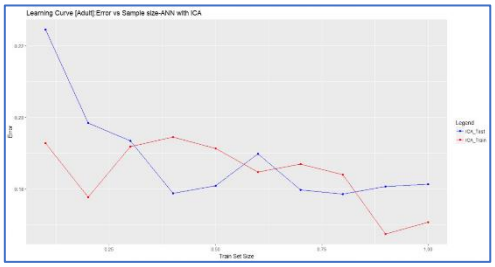
The purpose of ICA is to find independent components in the data. After analyzing the kurtosis information for the dataset with 96 components, all the components with kurtosis>100 are chosen, i.e., 52 components. The scatter plot for ICA on 2 dimensions is shown.

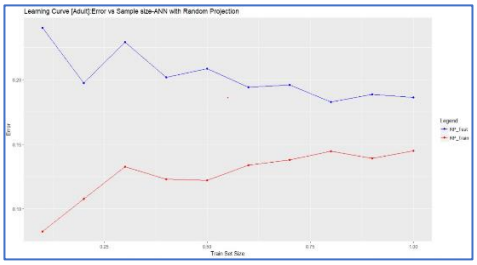
Random Projection implements a simple and computationally efficient way to reduce the dimensionality of the data by trading a controlled amount of accuracy (as additional variance) for faster processing times and smaller model sizes. Here, sparse random projection is implemented. It reduces the dimensionality by projecting the original input space using a sparse random matrix. Sparse random matrices are an alternative to dense Gaussian random projection matrix that guarantees similar embedding quality while being much more memory efficient and allowing faster computation of the projected data. By experimentation, it was realized that a random projection with sample size of 10 gives a sparse matrix of 55 components which used to implement RP

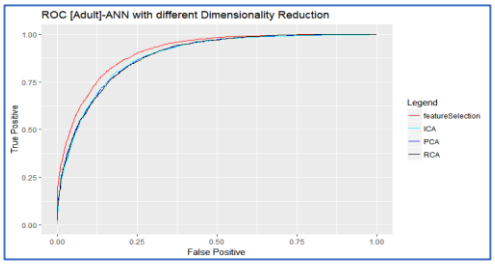
ANN applied to the Feature Reduction Methods







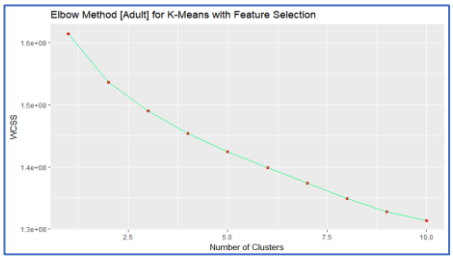




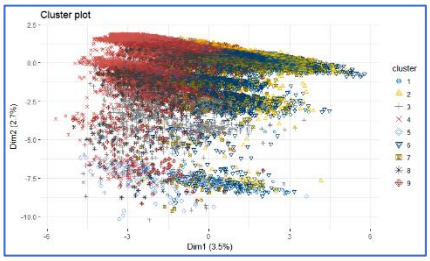
Feature Reduction Algorithm	Accuracy	AUC
Backward Elimination	84.14%	0.9096
PCA	81.85%	0.886
ICA	81.53%	0.8851
RP	81.48%	0.8863

The learning curves of sample size vs error for all feature reduction algorithms are given above. On considering the algorithm performance with respect to Accuracy, AUC and ROC plot shown on left, Backward Elimination outperforms the others. From the learning curves, we can find that apart from ICA, the others show a similar trend in the plot i.e., train and test errors plateaus at an elevated level of error indicating overfitting. However, ICA plot plateaus at a lower error, which indicates an ideal error curve.

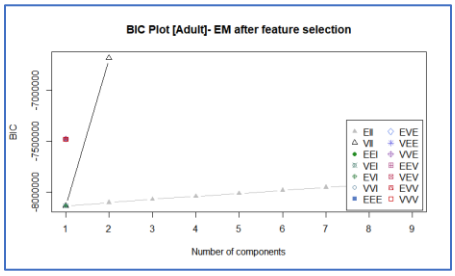
K-Means after Feature Selection using Backward Elimination



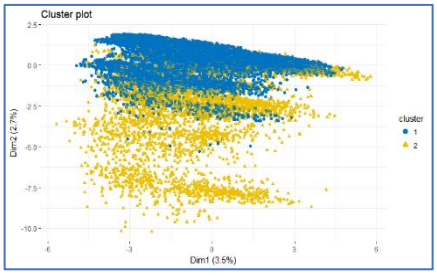
K-Means was applied to reduced dataset after feature selection with k=9, as obtained from scree plot shown below on left. The cluster plot obtained after feature selection shows an improvement in performance from previous K-Means implementation. K-Means gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.



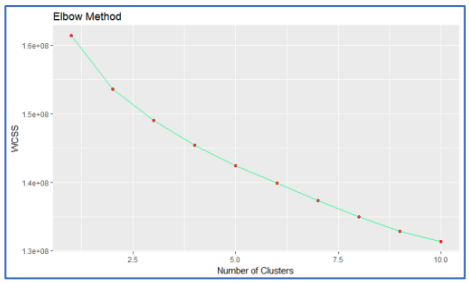
Expectation Maximization after Feature Selection using Backward Elimination



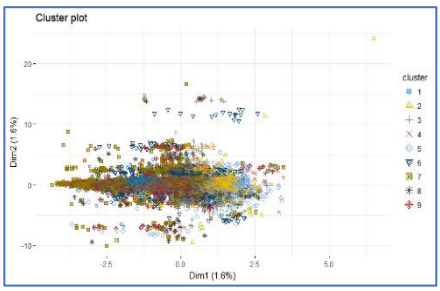
From BIC plot, it was obtained that a spherical, varying volume (VII) with 2 components is the most appropriate model. This is significantly different from previous result that resulted in a single cluster. The cluster plot for the same, in 2 dimensions is shown above. EM gives 2 clusters, which align with the class label of 2. From, the confusion matrix we can find that the cluster results are 68.62% accurate with respect to the class labels.



K-Means after Feature Transformation using PCA

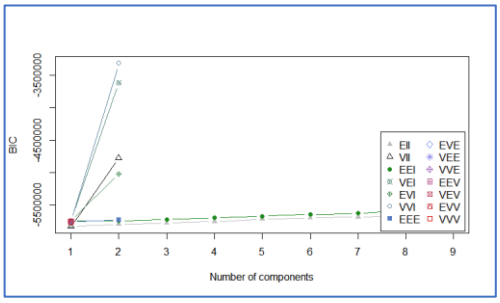


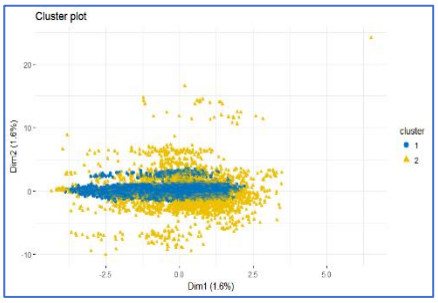
From scree plot, k=9 is chosen for clustering on the 64 PCA components due to a dip in WCSS. The cluster plot shows that only 3% of variance is explained by first 2 components. Moreover, the system performance time goes up by a big margin with PCA. It can be concluded that PCA hampers the performance of the supervised learning method for this dataset and shouldn't be applied for feature reduction. K-Means gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.



Expectation Maximization after Feature Transformation using PCA

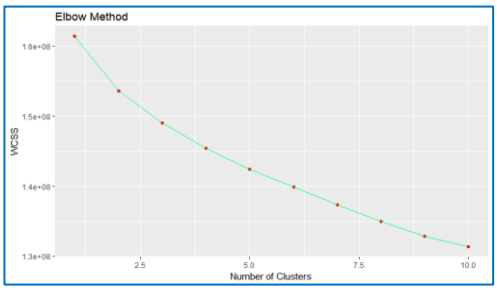
From BIC plot, it was obtained that a spherical, varying volume (VII) with 2 components is the most appropriate model. This is significantly different from EM without PCA that resulted in a single cluster. EM gives 2 clusters, which align with the class label of 2. The cluster plot for the same, in 2 dimensions is shown on right.



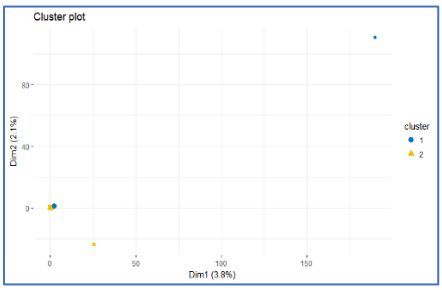


From, the confusion matrix we can find that the cluster results are 69.31 % accurate with respect to the class labels.

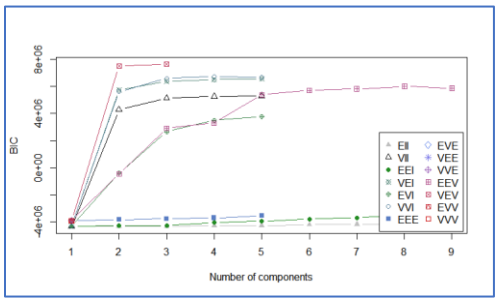
K-Means after Feature Transformation using ICA



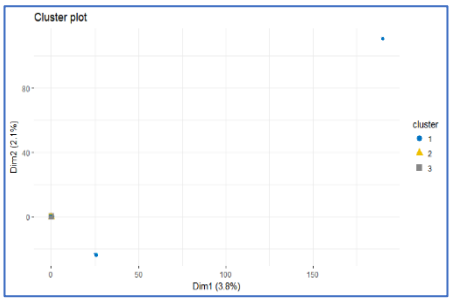
From scree plot, k=2 is chosen for clustering on the 52 ICA components due to a dip in WCSS. The cluster plot shows that only 5.9% of variance is explained by first 2 components. ICA hampers the performance of the supervised learning method for this dataset and shouldn't be applied for feature reduction. K-means gives 2 clusters, which align with the class label of 2. From, the confusion matrix we can find that the cluster results are 24.98% accurate with respect to the class labels.



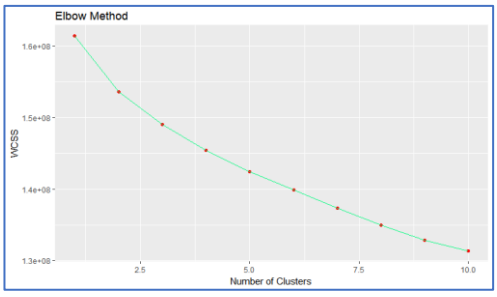
Expectation Maximization after Feature Transformation using ICA



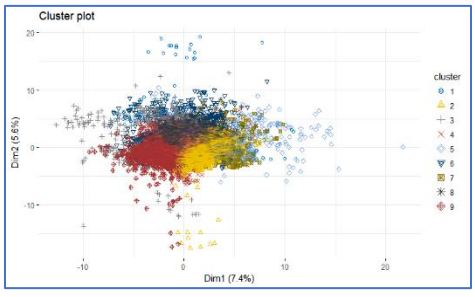
From BIC plot, it was obtained that an ellipsoidal, equal shape (VEV) with 3 components is the most appropriate model. This is significantly different from EM without ICA that resulted in a single cluster. The cluster plot for the same, in 2 dimensions is shown. EM improves the performance of ANN with ICA. EM gives 3 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to label.



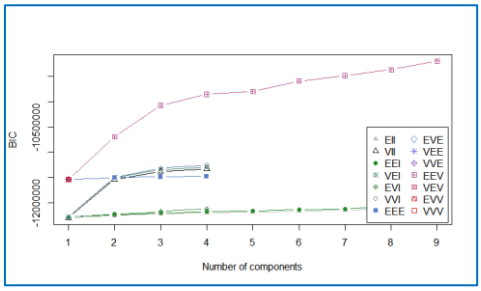
K-Means after Feature Transformation using RP



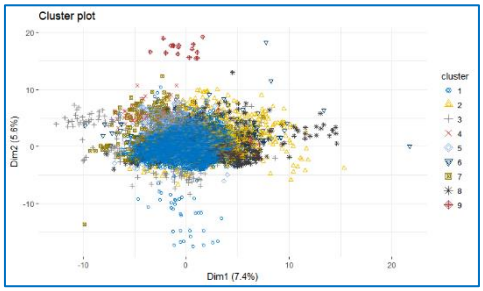
From scree plot, k=9 is chosen for clustering on the 10 RP components due to a dip in WCSS. It can be concluded that RP doesn't increase the performance of the supervised learning method for this dataset and shouldn't be applied for feature reduction for K-Means. K-Means gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.



Expectation Maximization after Feature Transformation using RP



From BIC plot, it was obtained that an ellipsoidal, equal shape (VEV) with 9 components is the most appropriate model. This is significantly different from EM without RP that resulted in a single cluster. The cluster plot for the same, in 2 dimensions is shown above. EM improves the performance of ANN with RP. EM gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

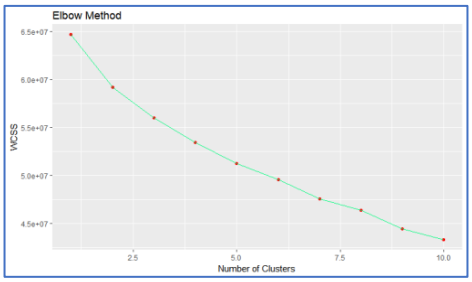


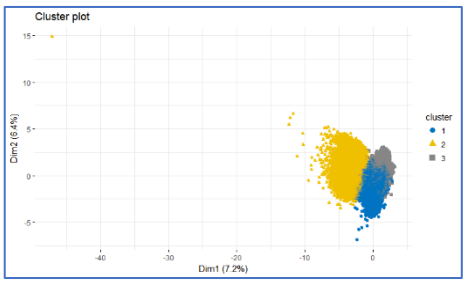
Dataset 2 [Bank]

K-Means

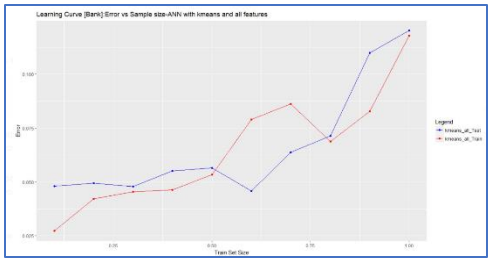
Bank dataset is mixture of numerical and categorical data. Since, K-Means can be applied only to numerical data, 1 hot encoding is performed on the categorical data to generate dummy variables to perform K-Means.

The [Bank] dataset results in somewhat well-defined clusters after performing K-Means. From the elbow plot, a dip in Within Cluster Squared Error is seen at k=3, which is taken for further clustering analysis. A scatter plot of the clustering in 1st two dimensions (explains 13.6% of variance) is shown above. The clusters overlap with each other and explains a very low percentage of variance in the first 2 principal components. This might be due to the mix of categorical and numerical data found in the dataset. A better approach to do clustering is to take k-prototypes which combines k-modes and k-means and can cluster mixed numerical / categorical data.



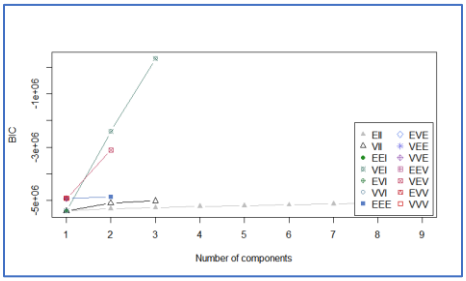


Artificial Neural Network On K-Means results

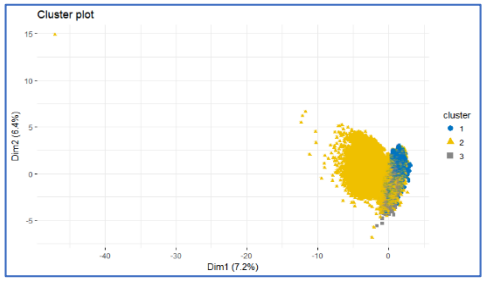


The results from K-Means clustering including all the features, cluster labels and class labels was applied to ANN model with 1 hidden layer and 10 nodes and the learning curves are shown on left. Moreover, K-Means cluster labels and class labels alone were applied as input to the same Neural Network. But ANN found the features to be insignificant and failed to work. The Error vs sample size plot shows a similar trend for training and test data sets. Learning Curve of error vs size shows high bias which implies the model underfits the data and model performs poorly for large sample size. As compared to the learning curves for ANN on original dataset, the fluctuations in the new learning curves have been smoothened out which implies a reduction in loss values and greater learning accuracy. The accuracy on ANN utilizing clustering results has gone down to 83.59% from 87.96% and AUC has gone up to 0.9010. Performing K-Means on the dataset before ANN hampers the performance of ANN. This might be because the dataset doesn't show clustering tendency. K-Means gives 3 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

Expectation Maximization

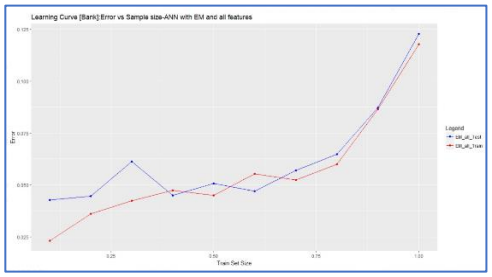


Expectation-maximization clustering probabilistically assigns data to different clusters. Diagonal, equal shape(VEI) with 3 components is found to be the optimal clustering model from BIC plot after finding out the BIC for various other models. Cluster plot for 3 clusters is shown.



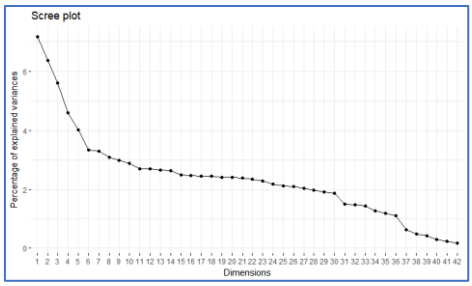
ANN on Expectation Maximization

On implementing, ANN with all features and cluster labels resulted in an accuracy of 84.5% and AUC of 0.909 over ANN with original dataset. The Error vs sample size plot shows a similar trend for training and test data sets. Learning Curve of error vs size shows high bias which implies the model underfits the data and model performs poorly for large sample size. As compared to the learning curves for ANN on original dataset, the fluctuations in the new learning curves have been smoothened out which implies a reduction in loss values and greater learning accuracy. Even though, EM shows a marked improvement over K-Means, it cannot cluster the dataset properly. EM gives 3 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

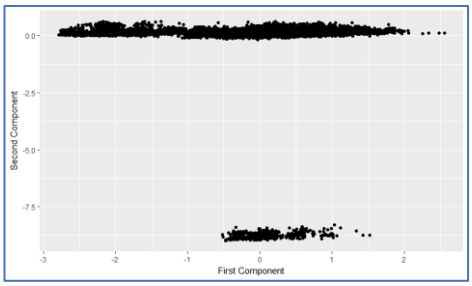


Dimensionality Reduction Algorithms

Feature selection using backward elimination as well as feature transformation using PCA, ICA and Random Projection was done on the dataset to reduce the number of dimensions of the dataset to improve interpretability, accuracy and visualization. Feature selection with backward elimination with logistic regression learner resulted in 13 features out of the total 15 features. PCA transforms the 42 features into 42 dimensions which explains 100% of the variance. From the scree plot for PCA shown on right, it can be obtained that 25 dimensions explains 80% of the total variation which is taken as the reduced number of dimensions from PCA for further

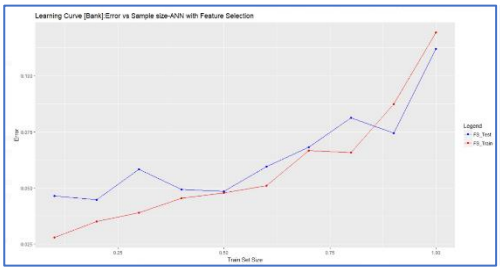


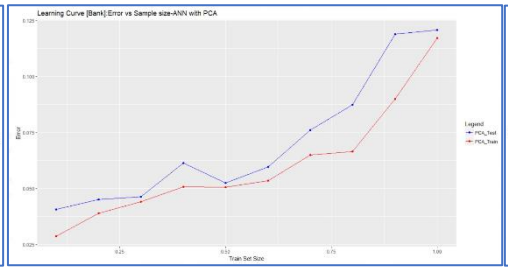
analysis. The purpose of ICA is to find independent components in the data. After analyzing the kurtosis information for the dataset with 42 components, all the components with kurtosis>10 are chosen, i.e., 32 components. The scatter plot for ICA on 2 dimensions is shown.

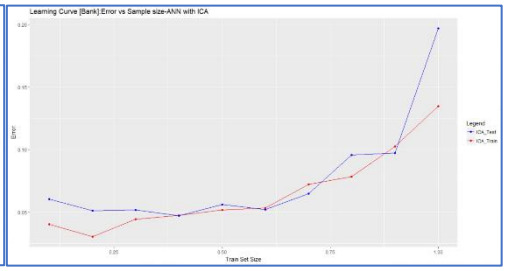


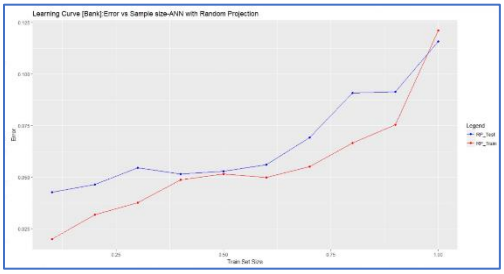
Random Projection using sparse matrix is implemented. By experimentation, it was realized that a random projection with sample size of 4 gives a sparse matrix of 33 components which used to implement RP.

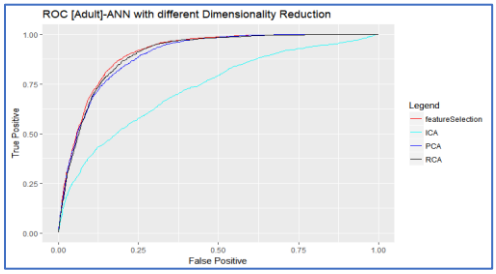
ANN applied to the Feature Reduction Methods









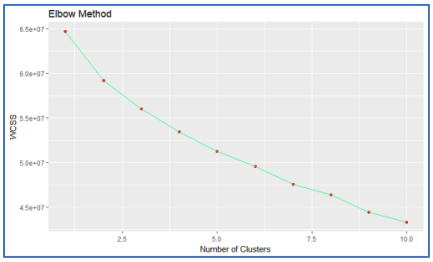


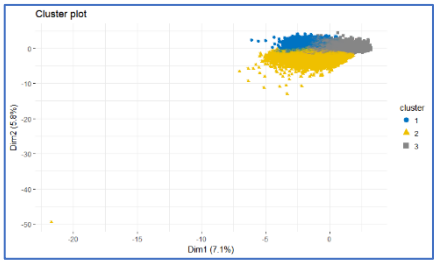
Feature Reduction Algorithm	Accuracy	AUC
Backward Elimination	87.81%	0.9072
PCA	86.62%	0.8967
ICA	82.58%	0.7294
RP	86.95%	0.9009

The learning curves of sample size vs error for all feature reduction algorithms are given above. On considering the algorithm performance with respect to Accuracy, AUC and ROC plot shown above, Backward Elimination outperforms the others. From the learning curves, we can find that all the plots show a similar trend i.e., train and test errors plateaus at an elevated level of error indicating overfitting.

K-Means after Feature Selection using Backward Elimination

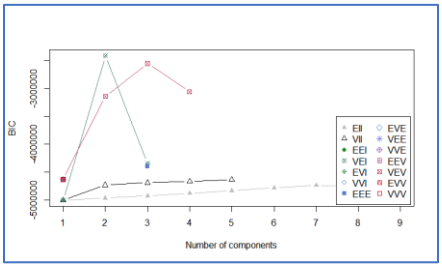
K-Means was applied to reduced dataset after feature selection with k=3, as obtained from scree plot shown below on left (same number of clusters as before). The cluster plot obtained after feature selection shows an improvement in performance from previous K-Means implementation. K-Means gives 3 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

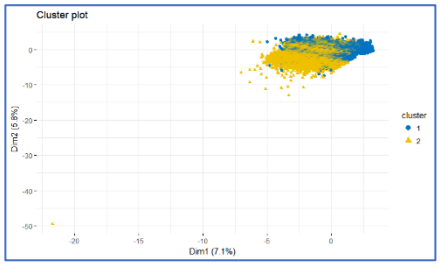




Expectation Maximization after Feature Selection using Backward Elimination

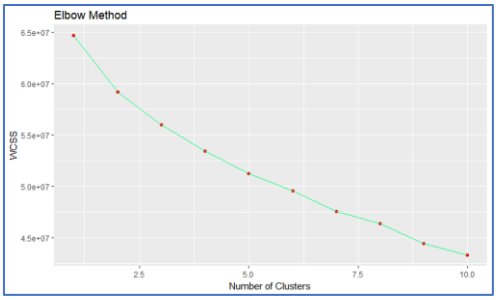
From BIC plot, it was obtained that a diagonal, equal shape (VEI) with 2 components is the most appropriate model. This is significantly different from previous result that resulted in a single cluster. The cluster plot for the same, in 2 dimensions is shown above. EM gives 2 clusters, which align with the class label of 2. From, the confusion matrix we can find that the cluster results are 58.87 % accurate with respect to the class labels.

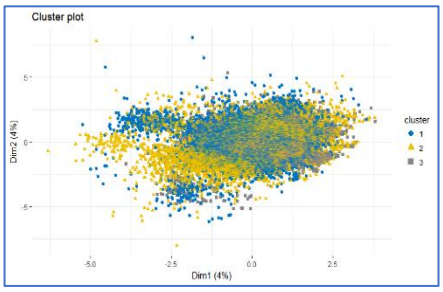




K-Means after Feature Transformation using PCA

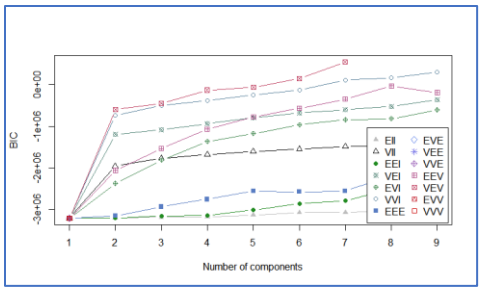
From scree plot, k=3 is chosen for clustering on the 25 PCA components due to a dip in WCSS. The cluster plot shows that only 8% of variance is explained by first 2 components. Moreover, the system performance time goes up by a big margin with PCA. It can be concluded that PCA hampers the performance of the supervised learning method for this dataset and shouldn't be applied for feature reduction. K-Means gives 3 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

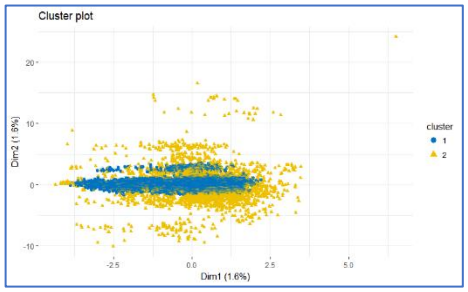




Expectation Maximization after Feature Transformation using PCA

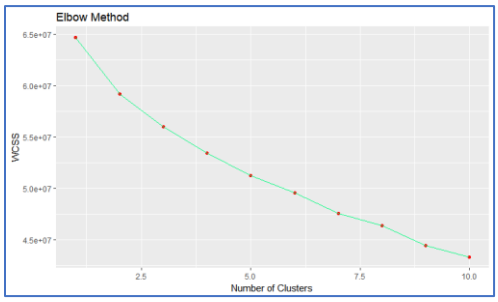
From BIC plot, it was obtained that a ellipsoidal, equal shape (VEI) with 2 components is the most appropriate model. This is significantly different from EM without PCA that resulted in 3 clusters. EM gives 2 clusters, which align with the class label of 2. From, the confusion matrix we can find that the cluster results are 70.186% accurate with respect to the class labels. The cluster plot for the same, in 2 dimensions is shown above.

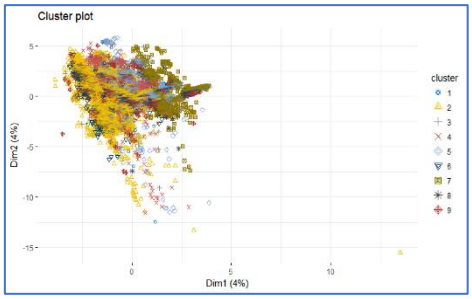




K-Means after Feature Transformation using ICA

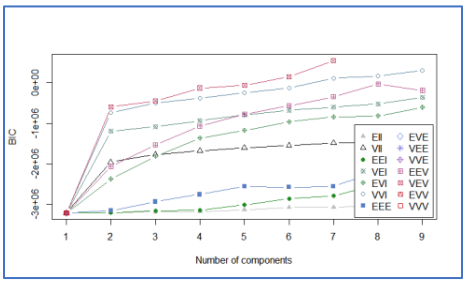
From scree plot, k=9 is chosen for clustering on the 52 ICA components due to a dip in WCSS. The cluster plot shows that only 8% of variance is explained by first 2 components. ICA hampers the performance of the supervised learning method for this dataset and shouldn't be applied for feature reduction. K-Means gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

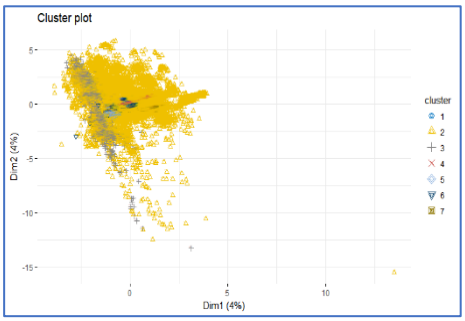




Expectation Maximization after Feature Transformation using ICA

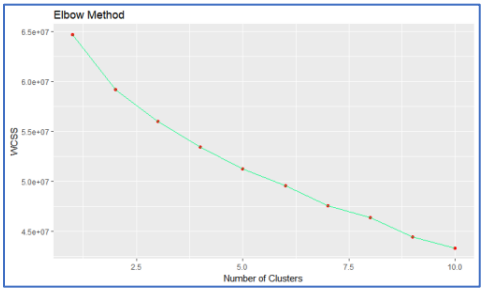
From BIC plot, it was obtained that an ellipsoidal, equal shape (VEV) with 7 components is the most appropriate model. This is significantly different from EM without ICA that resulted in 3 clusters. The cluster plot for the same, in 2 dimensions is shown. EM improves the performance of ANN with ICA. EM gives 7 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

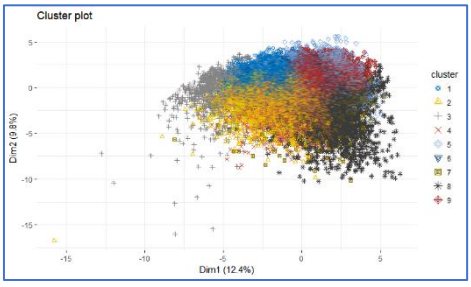




K-Means after Feature Transformation using RP

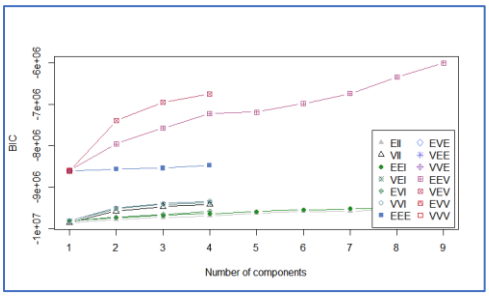
From scree plot, k=9 is chosen for clustering on the 10 RP components due to a dip in WCSS. It can be concluded that RP doesn't increase the performance of the supervised learning method for this dataset and shouldn't be applied for feature reduction for K-Means. K-Means gives 9 clusters, which doesn't align with



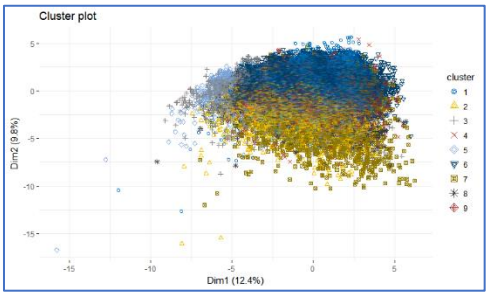


the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.

Expectation Maximization after Feature Transformation using RP



From BIC plot, it was obtained that an ellipsoidal, equal volume and shape (EEV) with 9 components is the most appropriate model. This is significantly different from EM without RP that resulted in a single cluster. The cluster plot for the same, in 2 dimensions is shown above. EM improves the performance of ANN with RP. K-Means gives 9 clusters, which doesn't align with the class label of 2. So, it's difficult to evaluate the cluster performance with respect to class label.



Performance Comparison

Adult

Algorithm	Accuracy	AUC
ANN	83.85%	0.9086
ANN with K-Means data	84.03%	0.909
ANN on EM data	84.08%	0.909
Backward Elimination	84.14%	0.9096
PCA	81.85%	0.886
ICA	81.53%	0.8851
RP	81.48%	0.8863

Cluster sizes for various algorithms

Adult Dataset	Size	1	2	3	4	5	6	7	8	9
ALL	K-Means	1232	6875	3383	1973	1126	644	8478	1238	5213
	EM	30162								
Backward Elimination	K-Means	10612	1674	3252	1523	641	893	375	799	10393
	EM	26483	3679							
PCA	K-Means	1444	3890	3114	1238	991	667	7670	6170	4978
	EM	26972	3190							
ICA	K-Means	87	30075							
	EM	21441	5913	2808						
RCA	K-Means	118	8366	471	3900	421	1330	694	6691	8171
	EM	25796	771	473	76	1473	56	384	1037	96

Bank

Algorithm	Accuracy	AUC
ANN	89.85%	0.8767
ANN with K-Means data	83.59%	0.9010
ANN on EM data	84.5%	0.909
Backward Elimination	87.81%	0.9072
PCA	86.62%	0.8967
ICA	82.58%	0.7294
RP	86.95%	0.9009

Cluster sizes for various algorithms

BANK Dataset	Size	1	2	3	4	5	6	7	8	9
ALL	K-Means	20971	8184	16056						
	EM	18141	22692	4378						
Backward Elimination	K-Means	11936	7408	25867						
	EM	25244	19967							
PCA	K-Means	20948	8135	16128						
	EM	31083	14128							
ICA	K-Means	1280	31783	2163	579	2379	888	2407	1565	2167
	EM	8373	18170	2201	10831	804	3822	1010		
RCA	K-Means	8941	4132	981	6808	1307	4162	1485	5069	12326
	EM	6337	2154	747	556	986	24964	2409	3423	3635

Summary

In this project, we implemented clustering algorithms such as K-Means and Expectation Maximization to look for underlying patterns in the data as well as applied feature reduction algorithms to choose relevant features for the same, while exploring how clustering can improve the accuracy of supervised learning algorithms like Artificial Neural Networks using two datasets, "Adult" and "Bank".

In this project, from the unsupervised clustering methods like K-Means and EM, we try to attain cluster labels that could be fed into ANN. This eliminates the necessity to feed all the features of our dataset into the supervised algorithm, which could increase the computational speed and prediction accuracy. However, out datasets were found to lack clustering tendency, evident from the clusters. The scatter plots of clustering algorithms shown in this project didn't give us much insight on how the data patterns are project onto the new spaces. Hence, we couldn't improve the performance of ANN through clustering in this project.

In "Bank" dataset, the clustering algorithms performed very badly. When ANN was fed with these clustering results, ANN couldn't perform well on these as the H2O package which runs the deep learning algorithm found these features to be insignificant and algorithm didn't converge. This implies bank data cannot be clustered. ANN on "Adult" dataset with all features took 3s and with only cluster labels and class labels took 1.6s. i.e., clustering significantly improves the speed of ANN at the cost of accuracy.

Among the clustering algorithms, EM was found to be better than K-Means. Among all the dimensionality reduction algorithms, Backward Elimination was found to improve performance. It was found that PCA, ICA and RP hamper the performance of ANN and shouldn't be used on the above datasets.