CS7641 Machine Learning
Assignment 3 : Unsupervised Learning

Seema Hanji (shanji3)

1 Data sets

Following datasets from Assignment-1 are used here for various experiments

1.1 Red Wine Quality

Location : winequality-red.csv : (https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv)

The data set contains various physicochemical variables (acidity, PH level etc.) as input and provides rating provided by wine experts as output. Since the ratings range between 0-10 , this data set can be challenging classification problem. Using this non-binary output , we can verify performance of various algorithms. It would be interesting to see how different algorithms apply feature selection.

This dataset has **1599-rows** with **11-features**

1.2 Credit Card Payment Default

Location : default_of_credit_card_clients.csv: (https://archive.ics.uci.edu/ml/machine-learning-databases/00350/)

The dataset is to predict if credit card clients in Taiwan, will be defaulting their next month's payment, based on various factors. These factors include client's education, gender, credit balance , past payment history etc. which act as input features to classification problem. The response variable is binary, indicating the probability of client defaulting next payment.

This dataset is fairly large, with **30,000-rows** with **24-features**

Both of these data sets differ each other in-terms of type of output, data set size as well as varying number of features. These would provide good exercise opportunities for the various algorithms we are going to apply.

2 Clustering

2.1 k-means Clustering

k-means clustering method partitions provided observations into k-clusters, where each observation belongs to the cluster with nearest mean(cluster centroid).

k-means clustering was applied to both "Red wine quality" data set and "Credit card default" data set, with varying hyper parameter n-clusters from the range of values *[1,2,5,10,15]*. Then **"inertia"** was noted down for each cluster size, where "inertia" is sum of squared distances of observations from cluster centroid. As per the plot, ***Figure-* ??** , as number of clusters increase, the inertia decreases, indicating clusters are becoming dense.

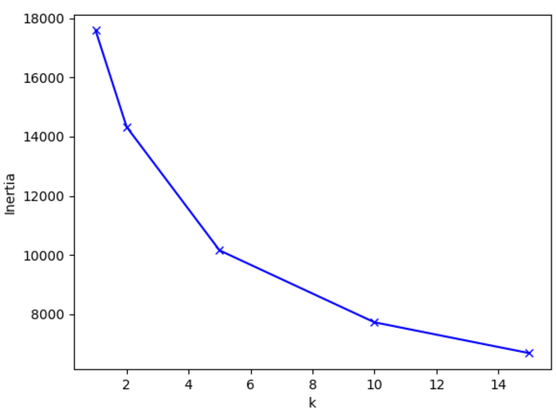


Table 1: K-means(Red Wine Dataset)

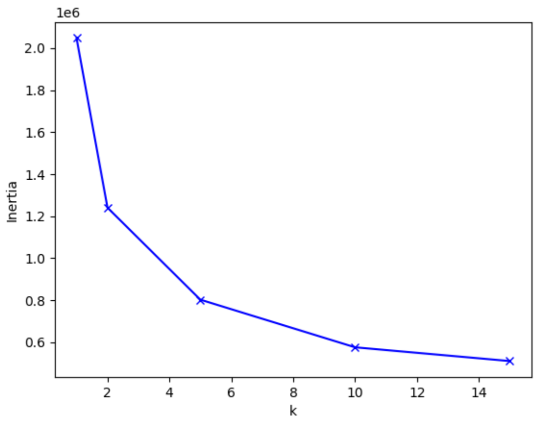


Table 2: K-means(CC Default Dataset)

Following are the cluster metrics - using v-measure, which is combination of *Homogeneity* and *Completeness* of a cluster.

Dataset	1	2	5	10	15
Red-wine quality	-6.270964e-16	0.029147	0.104306	0.138945	0.148698
CC default	3.395238e-15	0.000538	0.0171	0.065367	0.060631

Table 3

To identify optimal "k" , Elbow method is used. As per this method, optimal k is at the "elbow", at which point the inertia starts decreasing in linear way. In the case of Red wine quality data set, *k=5* appears to be optimal and in default credit card data set, *k=10* is optimal.

Based on optimal *k* , clustered points are plotted , using 2 of the prominent features from each data set.

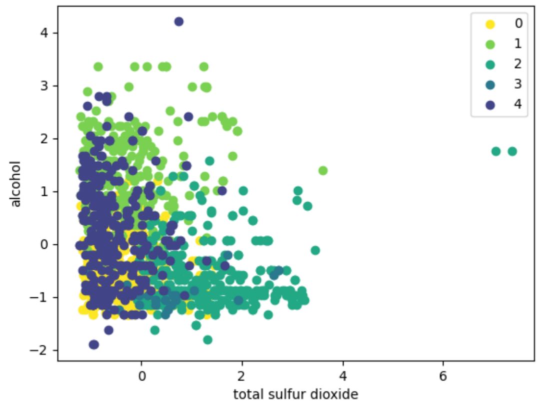


Table 4: K-means clusters(Red Wine Dataset)

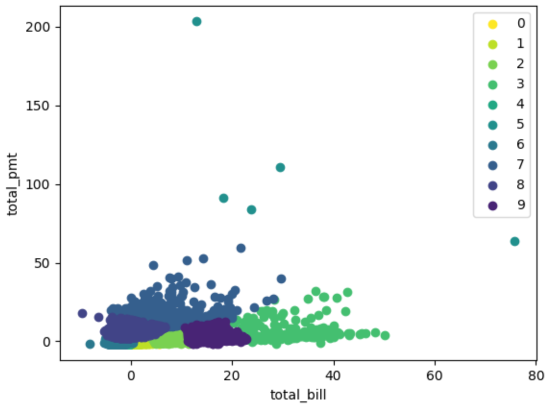


Table 5: K-means clusters(CC Default Dataset)

Looking at the v-measure metrics, both the data sets did not perform well using k-means clustering. Max v-measure scores are between 0.06 - 0.14 out of max 1.0 This indicates classes are distributed among clusters and are not grouped in same cluster, which brings down the homogeneity and completeness score.

2.2 Expectation Maximization

Expectation maximization was applied on both the datasets, using Gaussian Mixture method. In this method, iteratively clusters are assigned based on mean, variance and density, probability of data belonging to a cluster/gaussian distribution identified and finally resulting mean, variance and density is determined. These steps are iterated until convergence point.

Following plots indicate number of clusters Vs BIC (Bayesian Information Criterion) value. Optimal number of clusters is when BIC is lowest.

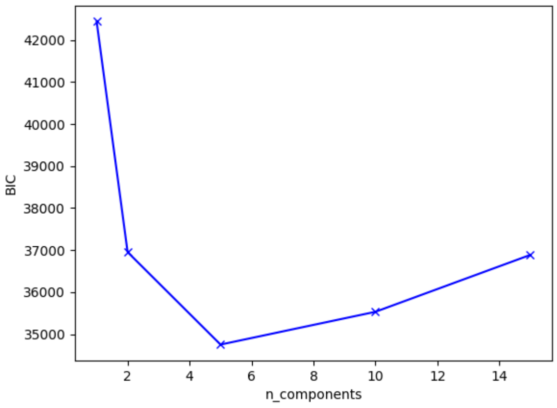


Table 6: EM clusters(Red Wine Dataset)

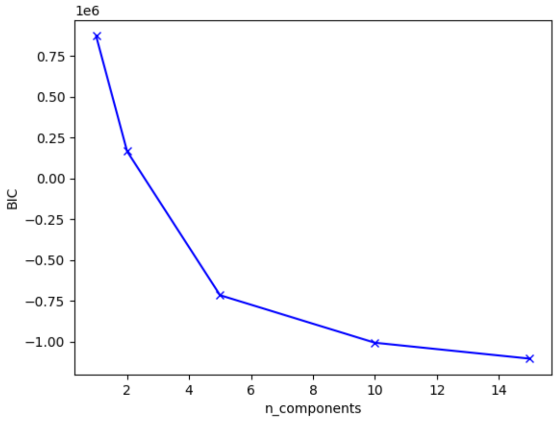


Table 7: EM clusters(CC Default Dataset)

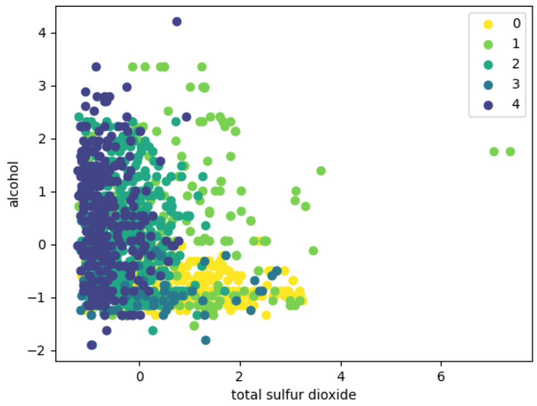
Following metric v-measure, determines homogeneity and completeness of clusters.

Dataset	1	2	5	10	15
Red-wine quality	3.395238e-15	0.00054	0.017035	0.065365	0.060583
CC default	3.395238e-15	0.015731	0.053748	0.056748	0.088744

Table 8

To identify optimal "k" , lowest BIC is considered. In the case of Red wine quality data set, *k=5* appears to be optimal and in default credit card data set, *k=5* is optimal.

Based on optimal *k* , clustered points are plotted , using 2 of the prominent features from each data set.



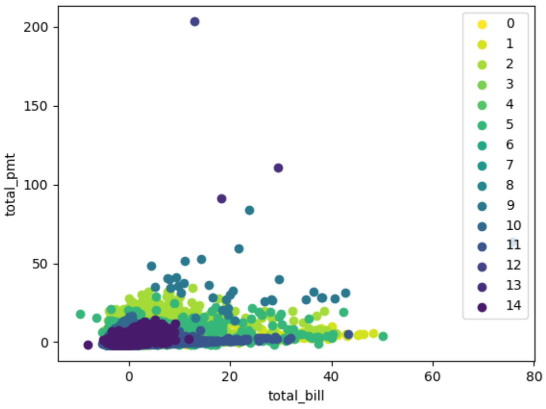


Table 9: EM clusters(Red Wine Dataset)

Table 10: EM clusters(CC Default Dataset)

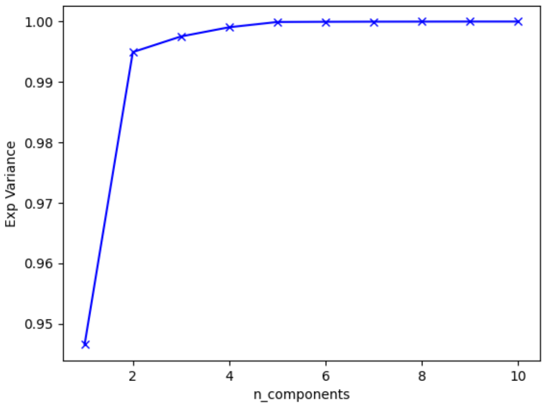
Looking at the v-measure metrics, both the data sets did not perform well using EM clustering. Max v-measure scores are between 0.06 - 0.08 out of max 1.0 This indicates classes are scattered among clusters and are not grouped in same cluster, which brings down the homogeneity and completeness score.

3 Dimension Reduction

Following dimension reduction algorithms were used to transform data from high-dimensional to low-dimensional.

3.1 PCA

Principal Component Analysis (PCA) finds directions of max variance in high dimensional and projects it into new space of equal or fewer dimensions. PCA was applied to both Red wine quality and default credit card data sets. First step was to to find out right number of PCA components to be produced. Each of the data set was analyzed by applying PCA with number of components ranging from 1 to n (where n is number of features in the data set).Then by looking at "explained variance ratio" (Ratio of variance of the component to total variance) right number of components was identified. As per the plots below, number of components ideal for *"Red Wine quality"* data set is *5* and for *"Credit card default"* data set *10*.



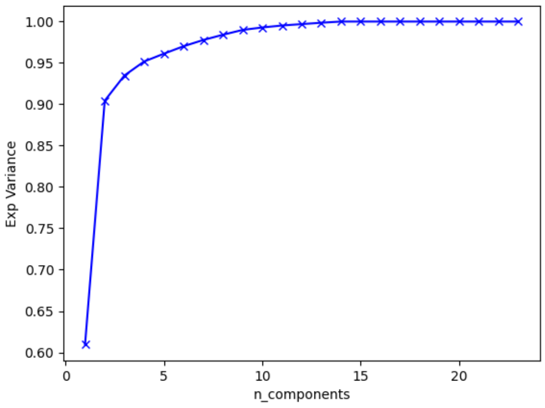


Table 11: PCA Components / Explained variation (Red Wine Dataset)

Table 12: PCA Components / Explained variation(CC Default Dataset)

Applying PCA using determined number of components on both datasets, following plots indicate the way data with different classes are clustered based on their similarity in PCA dimension(with PC1 and PC2 components as axis).

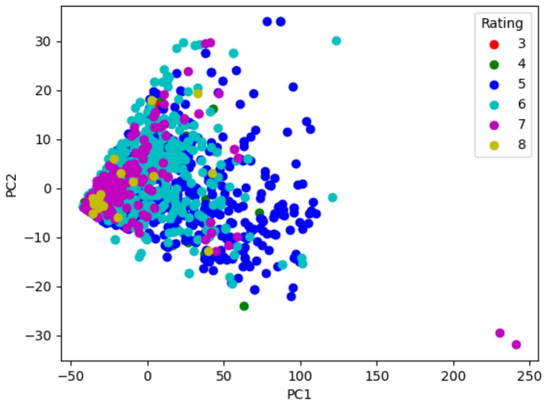


Table 13: PCA Clusters (Red Wine Dataset)

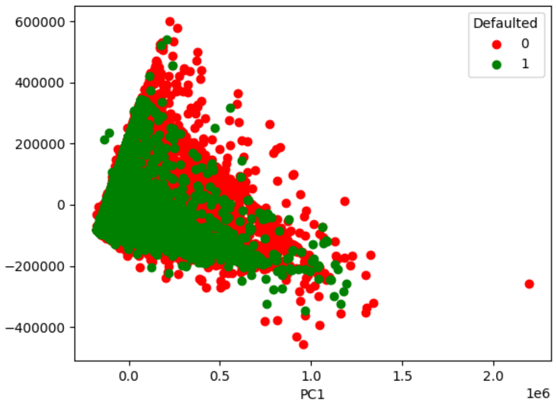


Table 14: PCA Clusters(CC Default Dataset)

These plots identify clusters in data and simplify complicated data with reduced features.

3.2 ICA

Independent Component Analysis (ICA) is linear dimension reduction method, which transforms data into columns of independent components. ICA was applied to both the data sets and to determine ideal number of components for ICA, average kurtosis was used as a measure. Following plots identify kurtosis for various number of components. The number of components with max Avg. Kurtosis is an optimal number, as maximum kurtosis indicates maximization of non-Gaussian components.

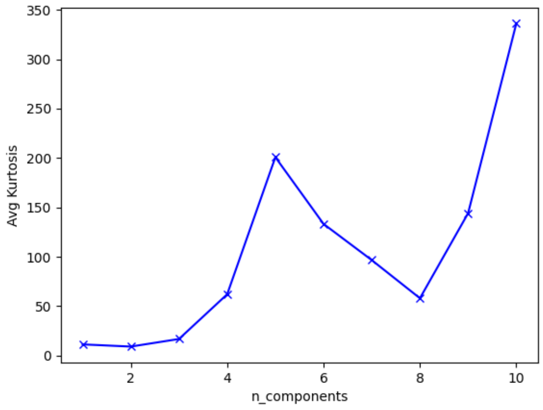


Table 15: ICA Components (Red Wine Dataset)

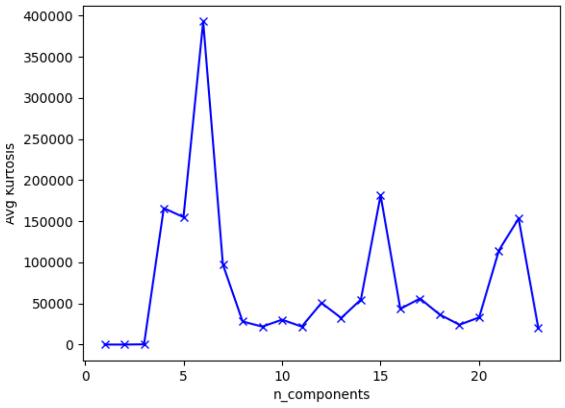


Table 16: ICA Compoenents(CC Default Dataset)

With optimal number of components *n=11* for Wine dataset and *n=6* for credit card default dataset, ICA was applied on data and following plots are result of ICA for the first 2 components space.

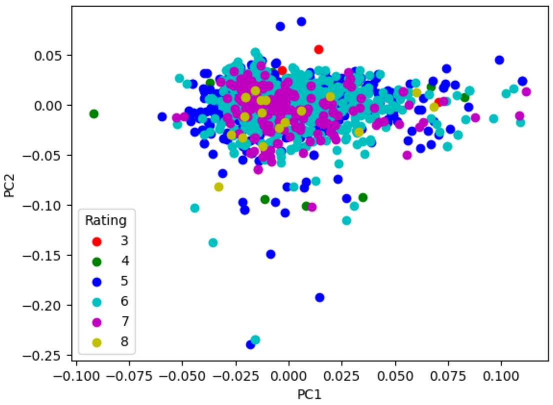


Table 17: ICA Clusters (Red Wine Dataset)

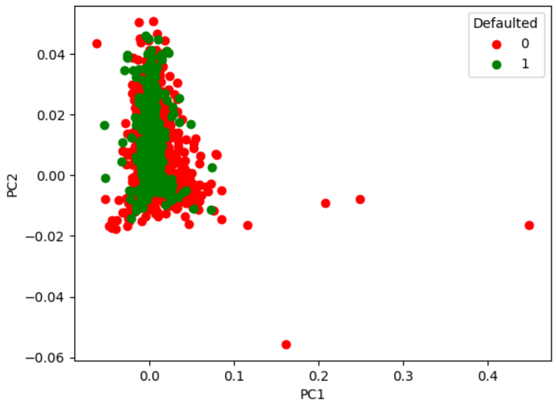


Table 18: ICA Clusters(CC Default Dataset)

3.3 Randomized Projections

Randomized Projections (RP) is a method of dimension reduction where random set of dimensions are selected and projected on to lower dimension space.

Experiment was run over both data sets with different number of components and based on minimum reconstruction error, optimal number was found. Following plots are from the experiment. In this case optimal values *n=11* and *n=24* were selected for respective data sets.

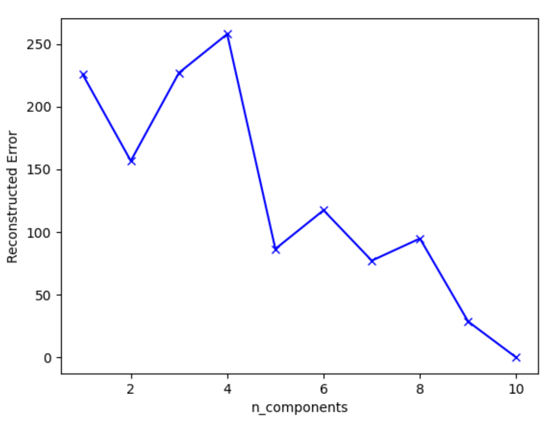


Table 19: Random Projections (Red Wine Dataset)

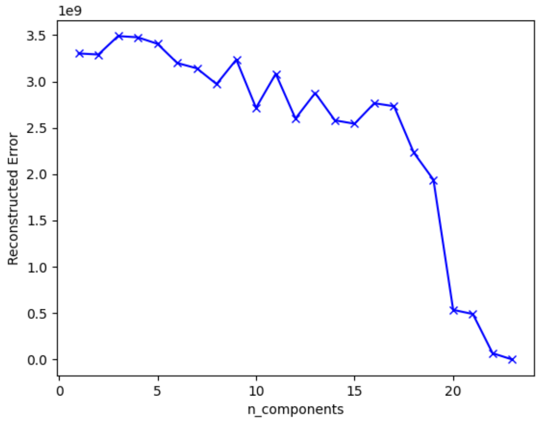


Table 20: Random Projections(CC Default Dataset)

This indicates, in this method as number of components increase, the performance over the data set gets better.

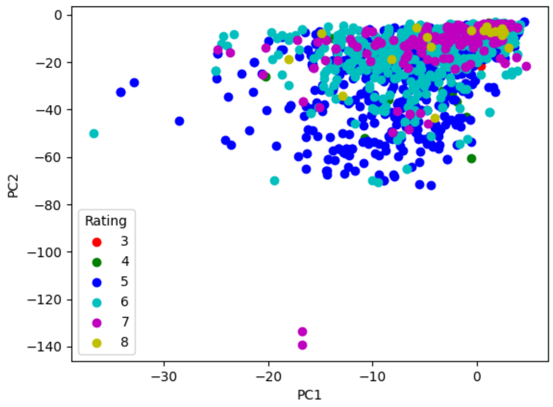


Table 21: RP Clusters (Red Wine Dataset)

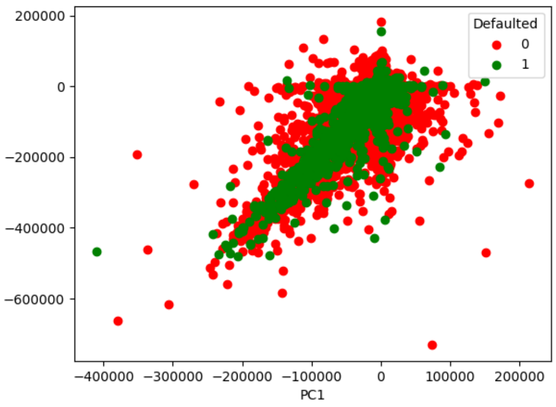


Table 22: RP Clusters(CC Default Dataset)

3.4 Select K Best

Select K Best method is to select features according to the k highest scores. To select k best features for both the data sets, grid search is used with different k values and optimal k is found based on the scores calculated using scoring method *"f_classif"* - which is an ANOVA Fvalue between label/feature for classification tasks.

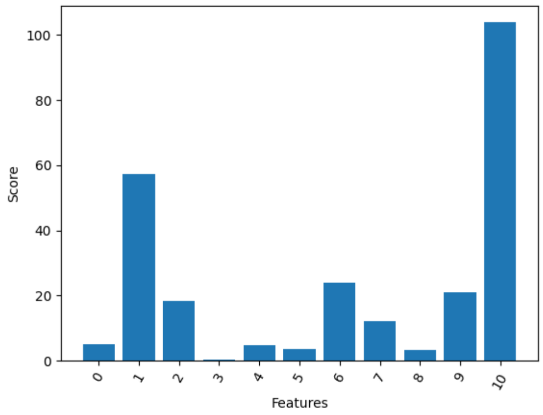


Table 23: Select K Best (Red Wine Dataset)

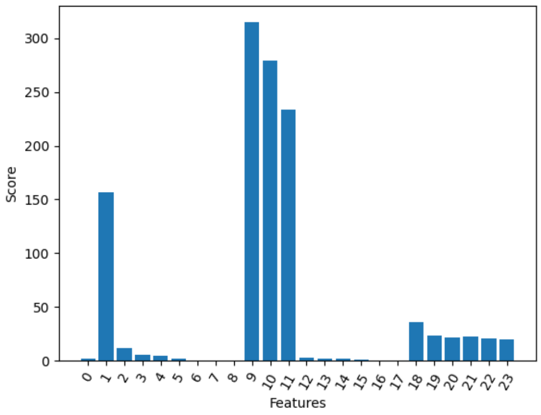


Table 24: Select K Best(CC Default Dataset)

For *Red Wine Quality data set* - optimal k : 8
CC Default data set - optimal k : 4

4 Experiments

4.1 Clustering With Dimensionality Reduction

DataSet - Red Wine Quality

After applying each of the dimensionality reduction algorithms (PCA, ICA, Randomized Projections and Select-K best) on Red Wine quality dataset, k-Means clustering was applied. Following plots show the result of clustering. Visual inspection of the clusters indicate that k-means clustering performs better grouping of data after dimension reduction using PCA and Randomized Projections algorithms. Comparing mutual information score of clusters formed by k-Means clustering without dimension reduction and k-Means clustering with dimension reduction, Randomized projection and Select-K best features perform better as shown in **Table- 41** , score is highlighted with green when it is better than without dim. reduction.

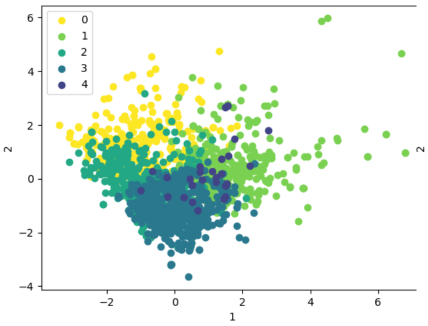


Table 25: K-Means/PCA

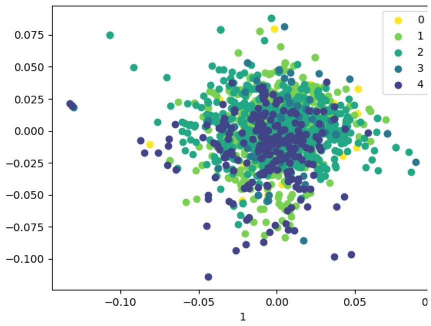


Table 26: K-Means/ICA

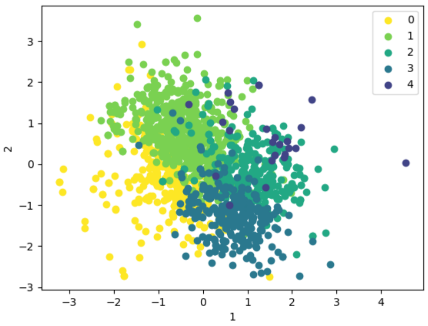


Table 27: K-Means/RP

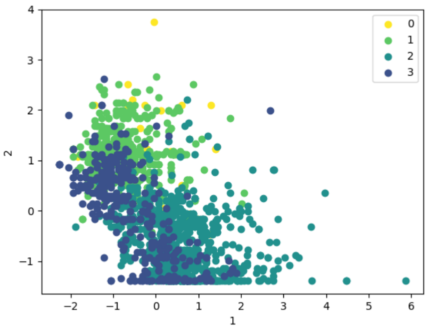


Table 28: K-Means/SK

Applying Expectation Maximization(EM) clustering after dimensional reduction applied, clusters as formed as shown in plots below. Visually clusters seemed to have better grouped for EM with PCA dimension reduction. Comparing mutual info. score of each of dim. reduction algorithms with EM, to score with only applying EM (as shown in **Table- 41** , EM scores better with PCA.

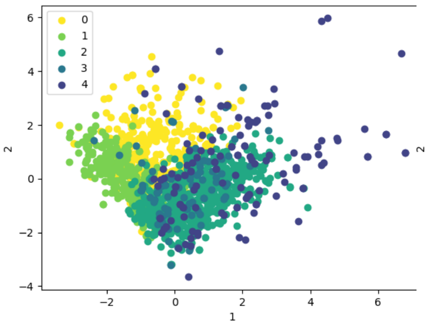


Table 29: EM/PCA

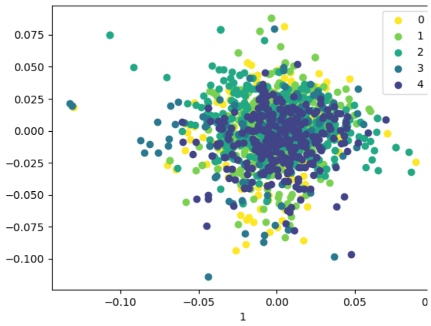


Table 30: EM/ICA

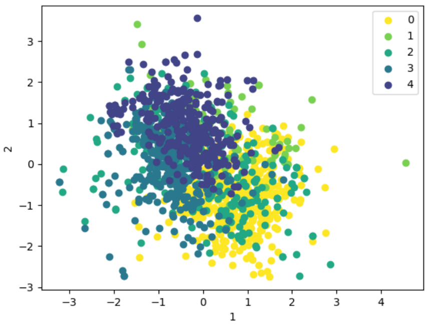


Table 31: EM/RP

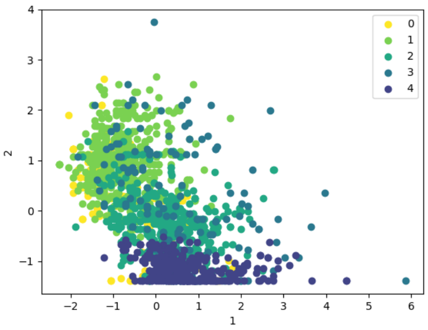


Table 32: EM/SK

DataSet - CreditCard Defaults

After applying dimensionality reduction algorithms (PCA, ICA, Randomized Projections and Select-K best) on CC default dataset, k-Means clustering was applied. Following plots show the result of clustering. Visual inspection of the clusters indicate that k-means clustering performs better grouping of data after dimension reduction using PCA and ICA algorithms. Comparing MI score of clusters formed by k-Means clustering without dimension reduction and k-Means clustering with dimension reduction, PCA and Select-K best features perform better as shown in **Table- 41** , score is highlighted with green when it is better than without dim. reduction.

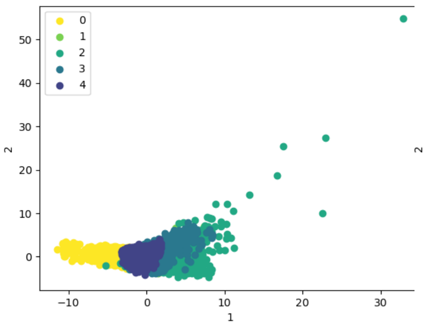


Table 33: K-Means/PCA

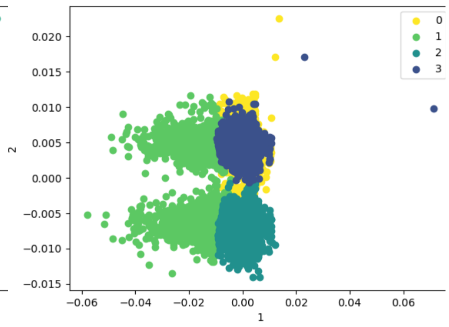


Table 34: K-Means/ICA

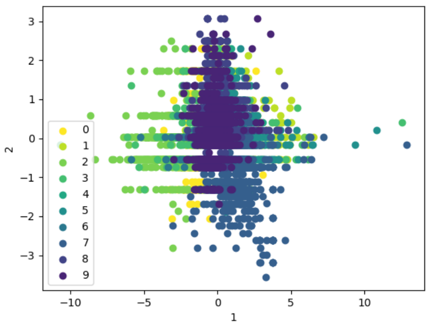


Table 35: K-Means/RP

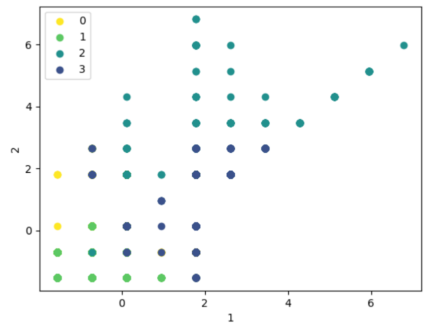


Table 36: K-Means/SK

Applying Expectation Maximization(EM) clustering after dimensional reduction applied, clusters as formed as shown in plots below. Visually clusters seemed to have better grouped for EM with PCA and ICA dimension reduction. Comparing mutual info. score of each of dim. reduction algorithms with EM, to score with only applying EM (as shown in **Table- 41** , EM scores better with PCA.

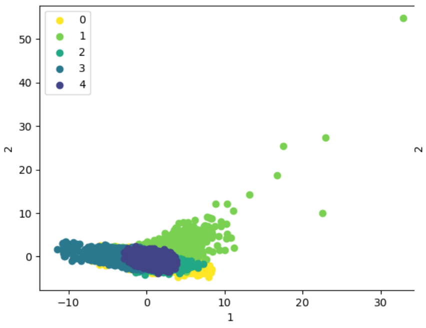


Table 37: EM/PCA

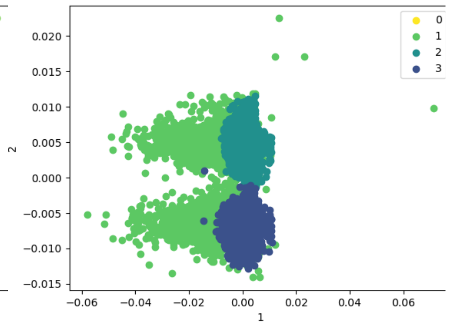


Table 38: EM/ICA

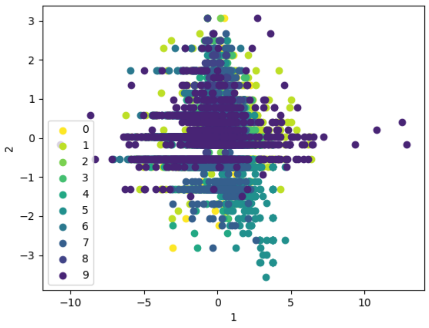


Table 39: EM/RP

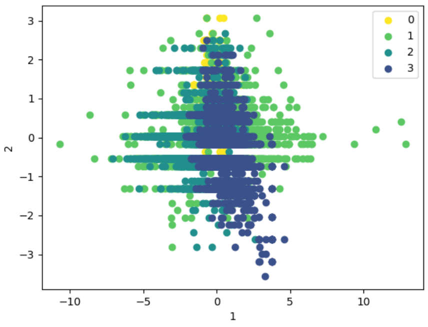


Table 40: EM/SK

Data Set	Dim. Red. Algo	Cluster Algo.	# Clusters	MI Score	MI Score (without dim red.)
Wine Quality	PCA	k-Means	4	0.087345	0.090455
Wine Quality	PCA	Exp. Maximization	5	0.101862	0.078003
Wine Quality	ICA	k-Means	5	0.063289	0.090455
Wine Quality	ICA	Exp. Maximization	5	0.071175	0.078003
Wine Quality	RP	k-Means	5	0.100406	0.090455
Wine Quality	RP	Exp. Maximization	5	0.078104	0.078003
Wine Quality	Select-K	k-Means	4	0.114089	0.090455
Wine Quality	Select-K	Exp. Maximization	5	0.098254	0.078003
CC Default	PCA	k-Means	5	0.043768	0.030610
CC Default	PCA	Exp. Maximization	5	0.036609	0.034857
CC Default	ICA	k-Means	6	0.001128	0.030610
CC Default	ICA	Exp. Maximization	4	0.005200	0.034857
CC Default	RP	k-Means	10	0.025736	0.030610
CC Default	RP	Exp. Maximization	10	0.032683	0.034857
CC Default	Select-K	k-Means	4	0.085360	0.030610
CC Default	Select-K	Exp. Maximization	4	0.019972	0.034857

Table 41: Clustering with Dim. reduction results

Above **Table- 41** lists mutual information scores for various experiments performed with clustering with dimension reduction algorithms. Also the scores without dim. reduction and only clustering are listed to compare. The scores which are better are highlighted in green.

4.2 Neural Network with Dimensionality Reduction

Each of the dimensionality reduction algorithms were applied and resulting data was used as input for Neural network classification for *Credit Card defaults* dataset. For this experiment, 30% of the data was held out from training set and used as test data. Hyper parameter used for neural-network : *hidden_layer_sizes=(5,),activation=’logistic’,solver=’adam’*

Accuracy metric is compared with original neural network result without dimension reduction. As shown in **Figure- 1** , Original Neural network accuracy still looks better and accuracy after dimension reduction using Randomized Projections and Select-K best algorithms match the original, if not better. This is not surprising, since with dimension reduction there is no information lost. Though the accuracy is worse in case of ICA , but still it is high accuracy of greater than 70%.

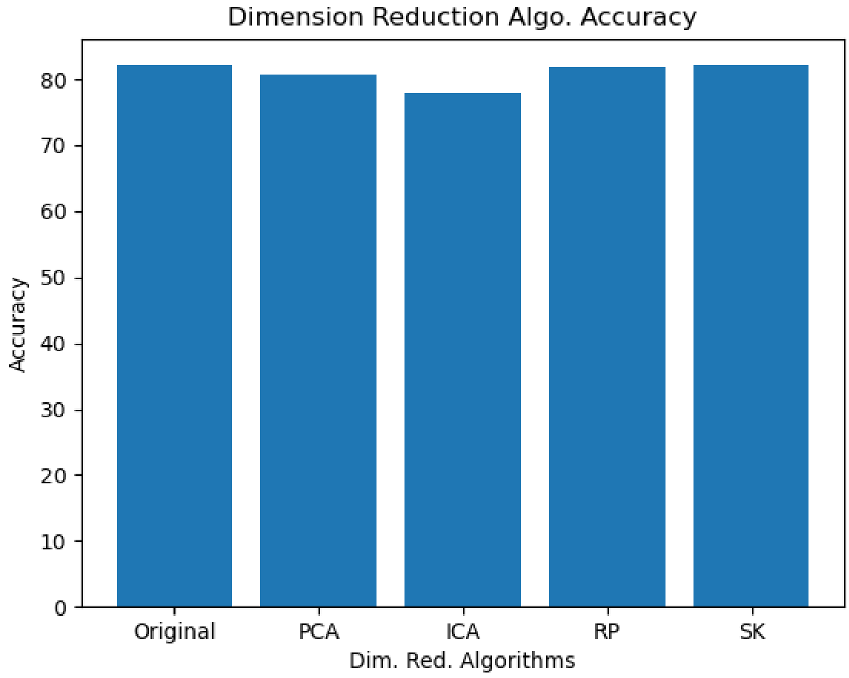


Figure 1: Accuracy after dimension reduction

4.3 Neural Network with Clustering

Each of the clustering algorithms were applied and resulting data was used as input for Neural network classification for *Credit Card defaults* dataset. For this experiment, 30% of the data was held out from training set and used as test data. Hyper parameter used for neural-network : *hidden_layer_sizes=(5,),activation=’logistic’,solver=’adam’*

Accuracy metric is compared with original neural network result clustering. As shown in **Figure- 2** , Original Neural network accuracy still looks better, accuracy after clustering using Expectation Maximization algorithm match the original, if not better. k-Means accuracy looks worse comparatively , but it is still high accuracy.

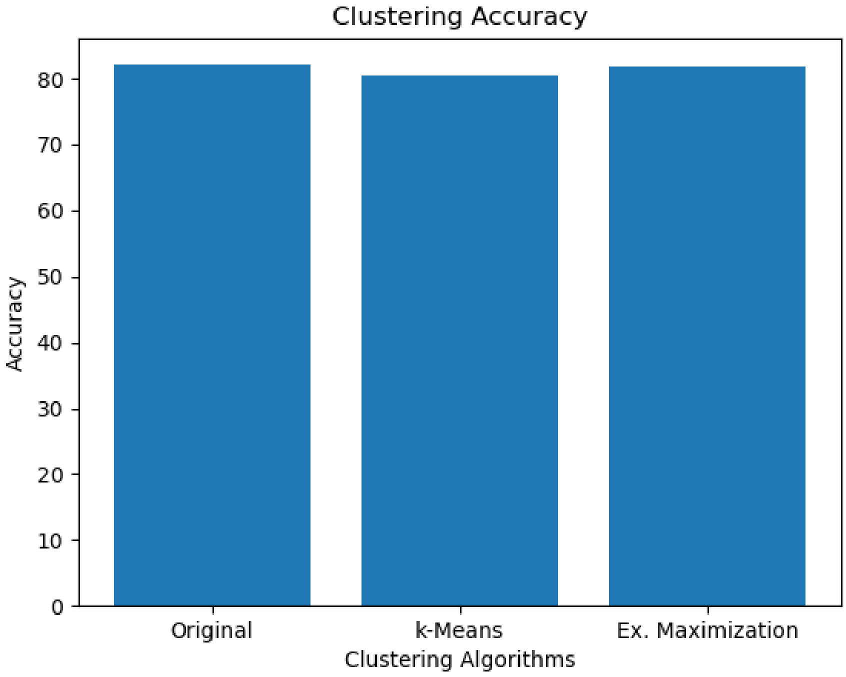


Figure 2: Accuracy after Clustering

5 Conclusion

Working through all the various experiments of clustering, dimension reduction on different types of data sets, I have observed different kinds results. The performance of an algorithm depends on the problem context. Here problem context is either data set, type of learning (unsupervised or supervised) done or other algorithms used in combination. Unsupervised algorithms in combination with supervised, can improve performance or degrade performance based on the context. In summary , unsupervised algorithms definitely help reduce the complexity of a problem and hence reduce issues found due to curse of dimensionality.

References