Unsupervised Learning Experiments

Why did I choose these datasets? Why are they interesting?

1. **Dataset 1 - Churn for Bank Customers**

   This dataset consists of customer's data of whether they are churned or not. It consists of 13 features and one column to predict whether the customer is churned (i.e. 1) or not (i.e. 0). As there are many features it was interesting to see which feature affects churning of the customer. Also, as the dataset is not balanced, it was interesting to see how each model learns and predicts. It has 2 categorical features, it was a good example to use one hot encoding. Also one of the features "Balance" ranged from 0 to 250898. So having features with wide ranges also can affect the predictions

2. **Dataset 2 - White Wine Quality**

   This dataset consists of 11 features which together decide the quality of white wine ("quality"). It ranges from a score of 0 (lowest quality) to 10 (excellent quality). This dataset is highly unbalanced, so it was interesting to decide whether few scores needed to be grouped together or what approach is to be followed. Also unlike previous dataset, this dataset has features of chemical compositions like chlorides, citric acid etc which are not commonly known and it's hard to depict how each feature affects the predictions.

   Having different sizes of dataset was helpful to understand how the size of dataset affects the predictions. So I chose Dataset 1 with 10,000 rows and Dataset 2 with 4,898 rows.

Important steps I followed throughout:

1. As clustering is sensitive to distance between the clusters and the points inside the cluster. If you don't normalize the features, algorithms give more weight to some features than others. I used Standard Scaler to scale all the features of both the datasets.
2. Also used label encoder and one hot encoder for categorical features.
3. Also for all the clustering and dimension reduction tasks, I used the whole dataset (i.e. did not split into train and test) as clustering is unsupervised. For the Neutral Network part, I splitted the dataset into train and test and then applied dimension reduction and then applied ANN.
4. For K-means I used both the Elbow method and the Silhouette method, but the later was more useful to see how close the clusters are and silhouette also considers variance, skewness etc.
5. For Expectation Maximization I used a Gaussian Mixture Model. For choosing the number of components, I used the BIC score VS number of components and picked one with the least BIC score.
6. For feature selection, I selected ExtraTreesClassifier's feature importance and picked top 7-8 features.
7. For evaluating how well each clustering algorithm performed, I used 3 metrics. Confusion matrix to visually see distribution across all clusters. Completeness score and Homogeneity score to see how pure the clusters are and how well they are separated.
8. I also plotted the clusters after each dimension reduction to visualize the clusters and how it may have changed the representation of data points in the datasets.
9. I also used the prediction column i.e. classes (as the datasets are for supervised learning) as a way to see how well the clusters are formed.

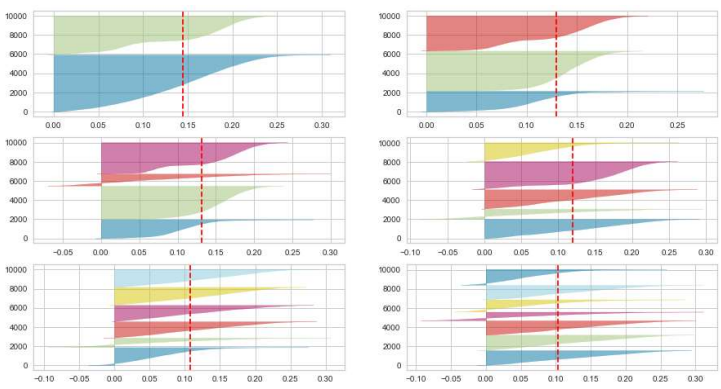
Datasets information

Dataset 1: Here I have 2 classes, 0 and 1 for churned and active

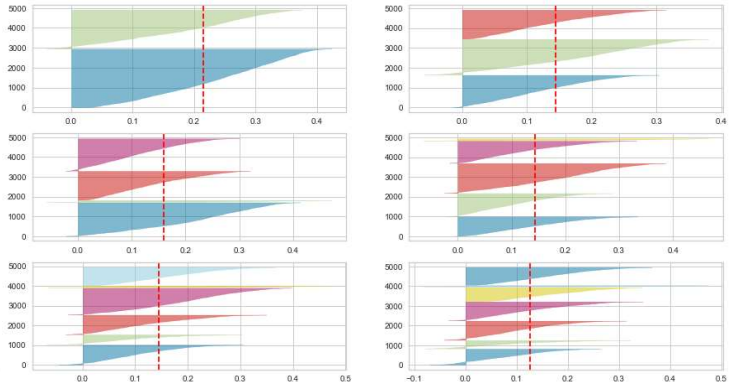
Dataset 2: Here I had 3 classes (for assignment 1 low-medium-high), I converted them into 2 classes for easy clustering and analysis as low-high i.e. 2 classes

K means - Before Dimensionality Reduction

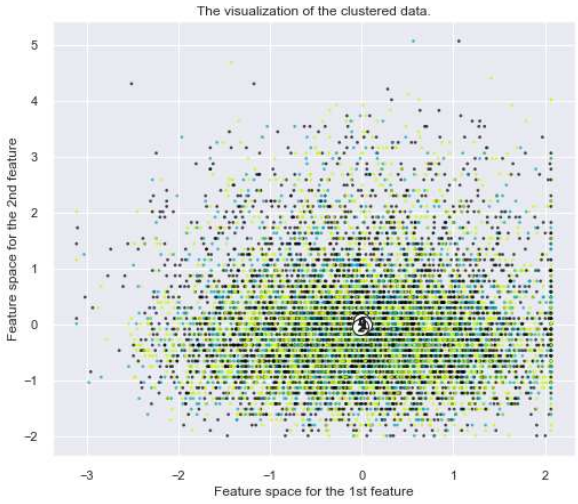
For both datasets, the optimal number of clusters were 3. As it can be seen for clusters = 3, the distribution is kind of uniform for both datasets. The inertia- i.e. how well datasets are clustered should be minimum, i.e. the sum of squared error for each cluster should be minimum. According to this condition, I chose k for not only this part but also for the rest of the analysis.



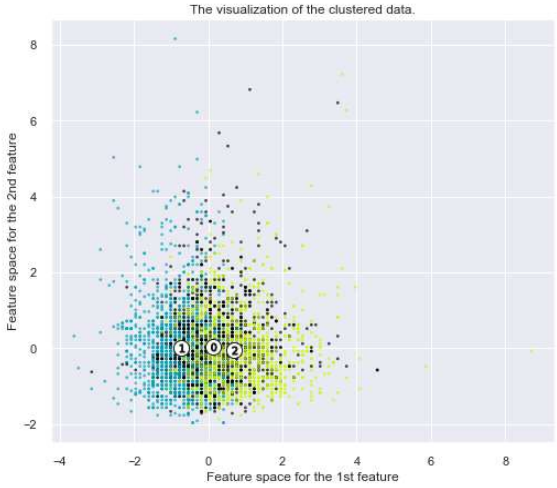
Dataset 1 clusters 2 to 7



Dataset 2 clusters 2 to 7



Clusters = 3 for Dataset 1



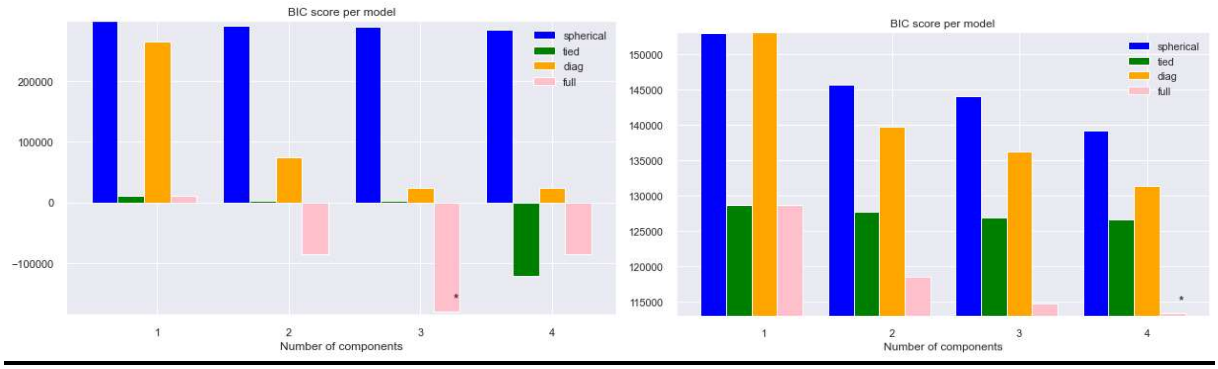
Clusters = 3 for Dataset 2

As seen in the above 2 figures, the clusters are overlapping and do not give proper separate clusters.

For Dataset1 and Dataset 2, the clusters overlapped a lot and the scores were very low.

Dataset	completeness_score	homogeneity_score
1	0.00794	0.01672
2	0.02343	0.04025

Expectation Maximization (Gaussian Mixture Model) - Before Dimensionality Reduction



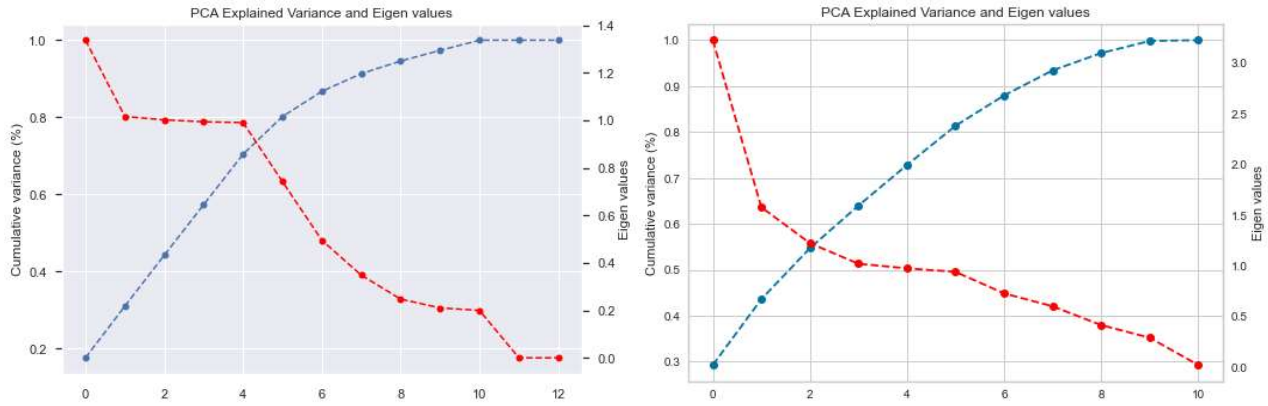
Dataset 1 *Dataset 2*

For EM, after choosing the lowest BIC scores and choosing the appropriate covariance_type and n_components, For Dataset1 and Dataset 2, the clusters overlapped a lot again

Dataset	completeness_score	homogeneity_score
1	0.2521	0.62883
2	0.3780	0.2521

Even after multiple tries, the scores did not increase. I also tried to use different data scaling techniques before applying clustering algorithms, but the best performing was the StandardScaler.

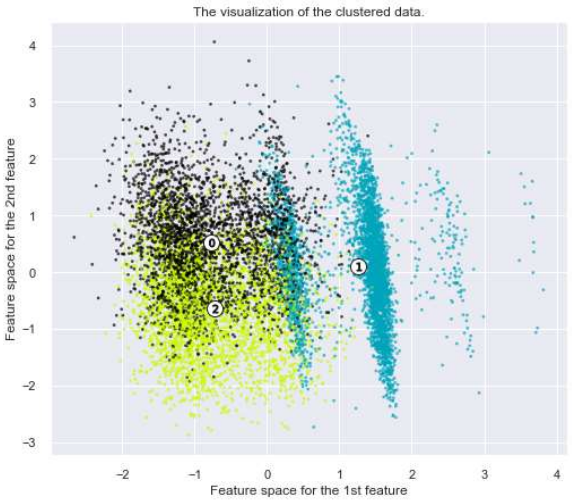
PCA



Eigenvalues and variance for Dataset 1 *Eigenvalues and variance for Dataset 1*

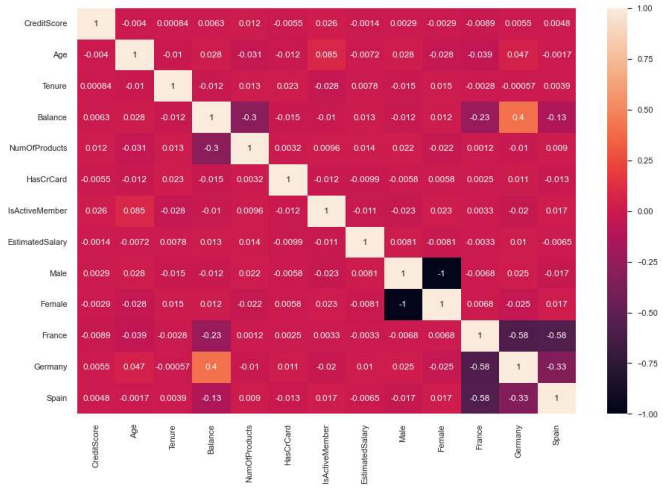
For PCA, the goal was to capture about 90-95% variance without losing a lot of information. For Dataset 1 it was around 7 components and for Dataset 2 it was also 7 components where PCA captured around 90-95% variance.
After applying PCA on both the datasets, the graphs for the Silhouette method for k means and the BIC score for EM were very similar to that of the above graphs. The clusters were still 3 for both datasets.

After visualizing 2 components of the dataset 1 transformed using PCA, we can clearly see that the clusters have been separated a little but compared to the first where PCA was not applied. This shows the transformation allowed the datasets to get separated and we can at least start to see clusters a little more clearly. It had a similar effect on dataset 2.



Dataset with PCA applied	completeness_score	homogeneity_score
1	0.078	0.16
2	0.12	0.23

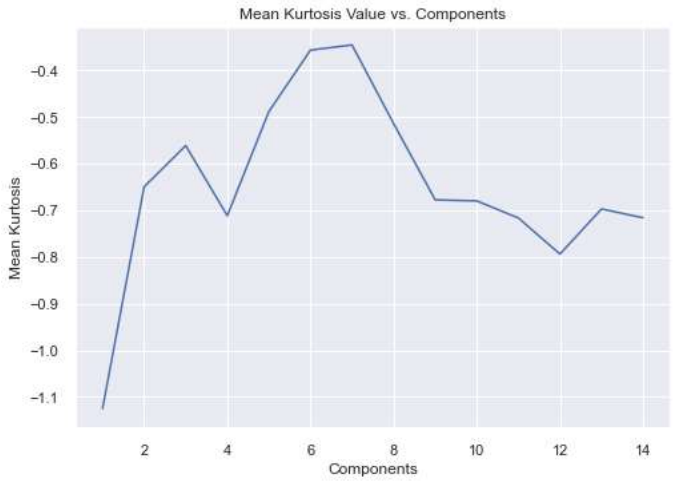
The reason why PCA is not working very well on both of the datasets is that, PCA may not always work on hot encoded features, it is mainly meant for continuous variables. Also, it works well on features which are strongly correlated. If I plot the correlation matrix for one of the datasets, the features are not very strongly correlated. This might be an issue as to why PCA is not increasing the performance of k-means and EM.



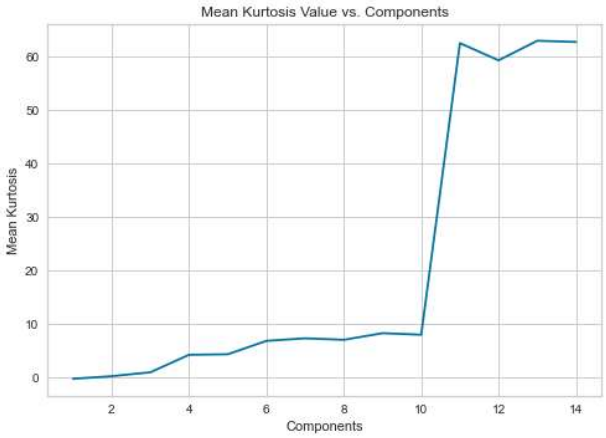
Correlation matrix for Dataset1

ICA

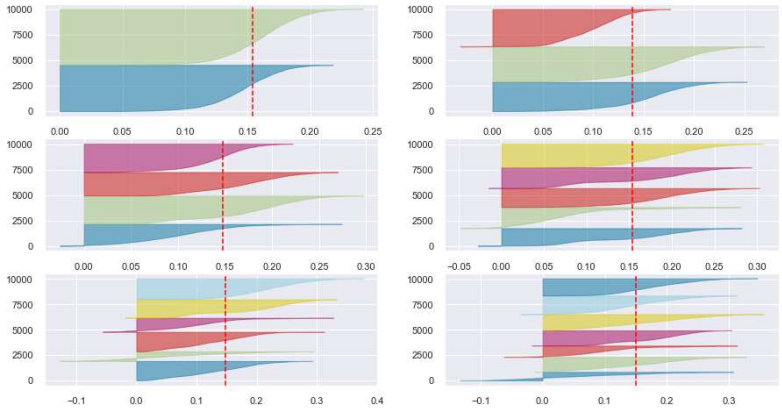
The goal of ICA is to maximize non-gaussianity (kurtosis) which gives independent components. For Dataset1 it is 7 components and for Dataset2 it is 11 component where the kurtosis was maximum.



Components vs mean kurtosis for Dataset 1

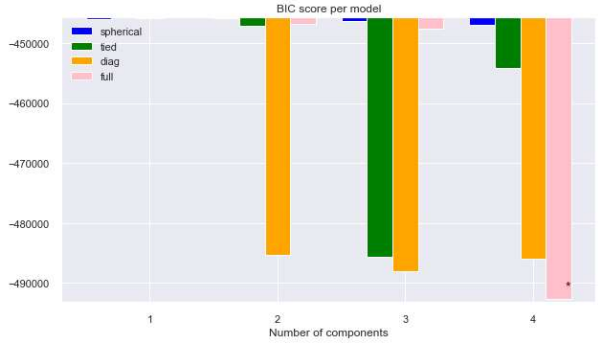


Components vs mean kurtosis for Dataset 2

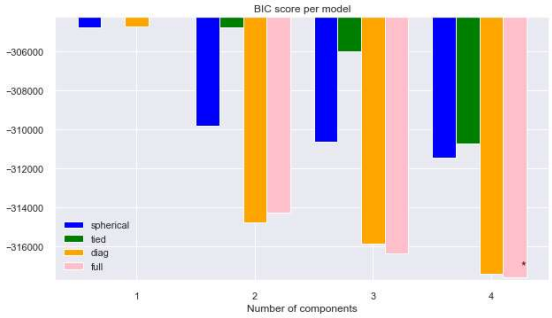


After transforming the data using ICA, and again using silhouette method, the clusters for k means changed from 3 to 4 as it was more evenly distributed and even in the elbow method the elbow was at 4 clusters. The figure at left is similar for both datasets so included only 1 figure.

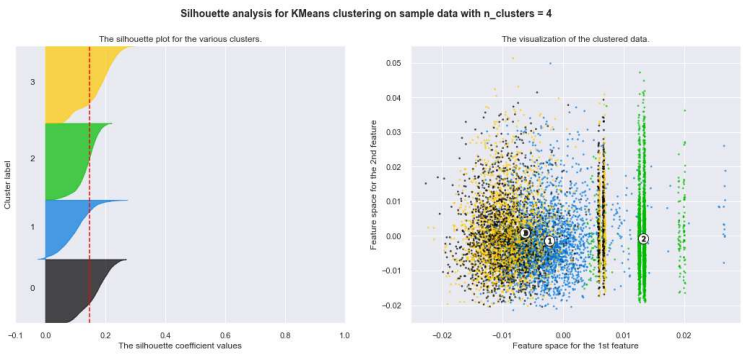
For EM the BIC values were negative, the ultimate goal is to minimize BIC so for that for both datasets, 4 components were chosen.



BIC values for Dataset 1

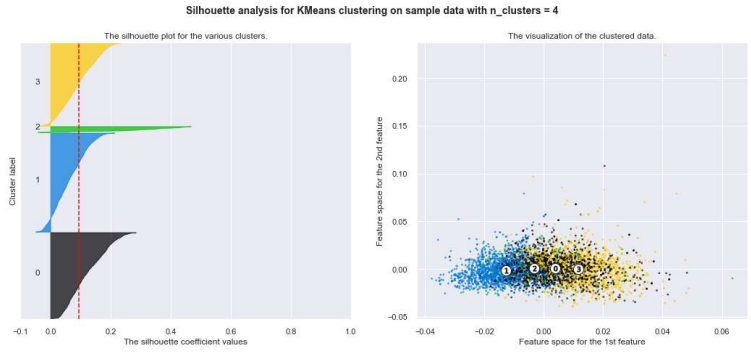


BIC values for Dataset 2



The clusters for dataset 1 are still a little crowded after applying ICA and clusters = 4. It did not change much, only for cluster 2 the data points are more evident

Dataset 1- ICA transformed- 4 clusters

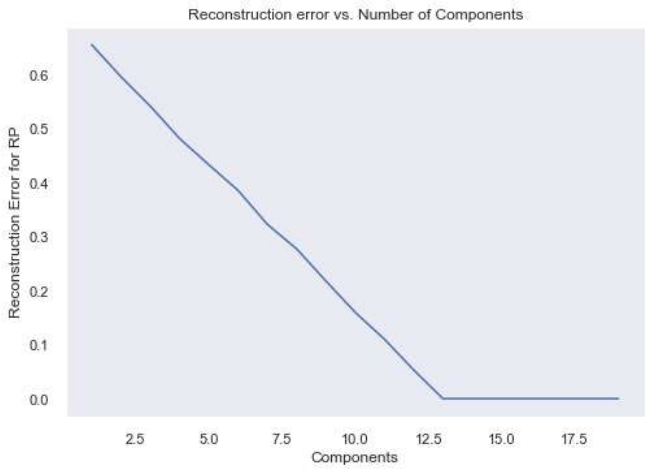


For dataset 2, ICA did not help much, the clusters are still crowded.

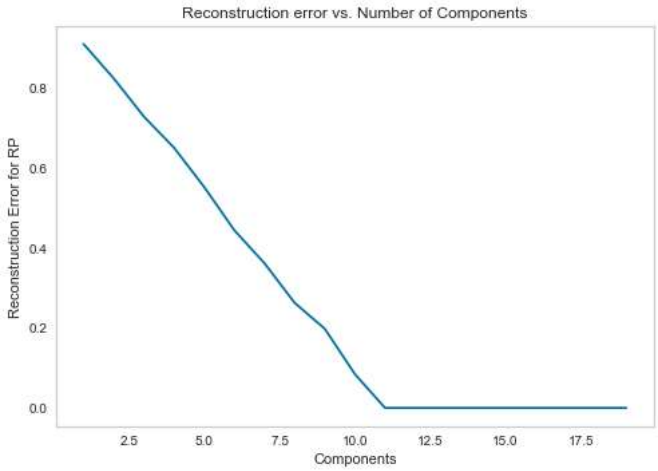
Dataset 2- ICA transformed- 4 clusters

Dataset with ICA applied	completeness_score	homogeneity_score
1	0.222	0.162
2	0.153	0.178

GaussianRandomProjection

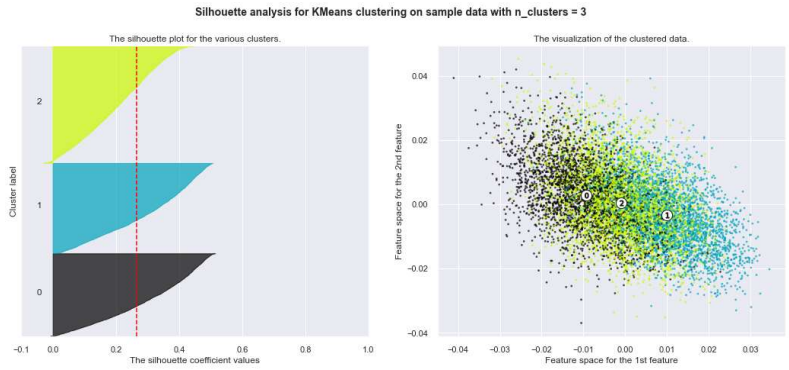


Reconstruction error for Dataset 1



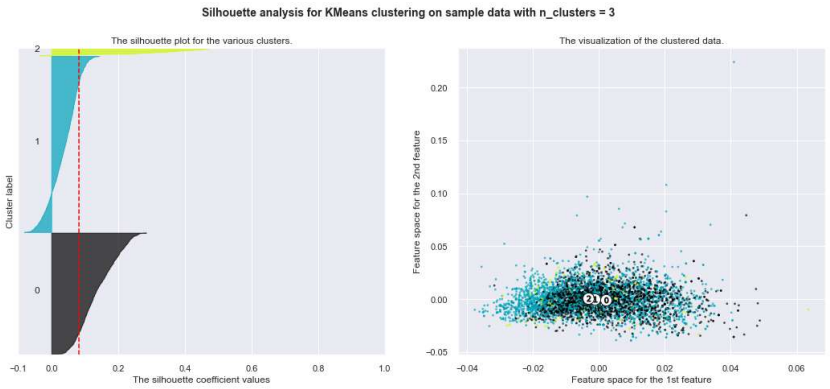
Reconstruction error for Dataset 2

After running RP several times, it is observed that after certain components the RP tends to be zero. The reconstruction error must be as little as possible but still be at a good level.
The clusters formed for k means and the n_components using BIC values for EM were very similar to that of PCA which has 3 clusters for k means and 4 components for EM.



For Dataset 1, the clusters are overlapping but a little scattered than PCA. But still is not able to have separate evident clusters.

Dataset 1- RP transformed data-3 clusters



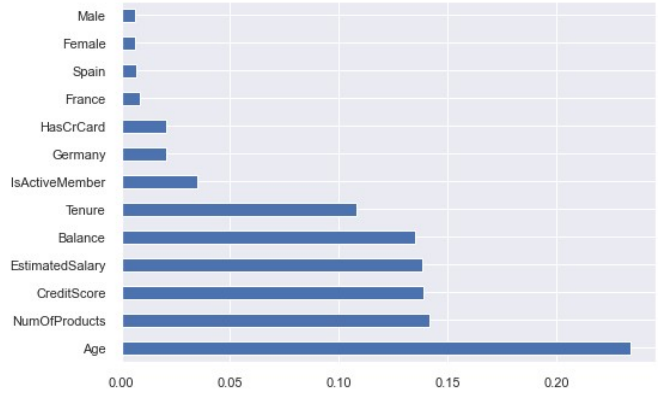
For Dataset 2, RP is not working well as that of previous PCA and ICA,. The clusters are again overlapping

Dataset 2- RP transformed data-3 clusters

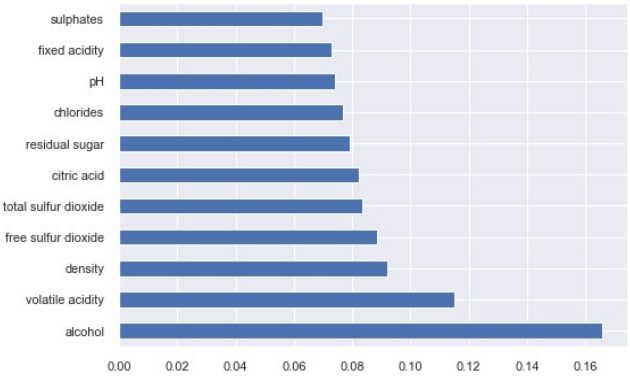
Dataset with RP applied	completeness_score	homogeneity_score
1	0.155	0.126
2	0.164	0.29

Feature selection - ExtraTreesClassifier

I used the feature importance from this classifier and ranked them from low importance to high importance and picked the most important features and removed the less important ones.



Feature importance for Dataset 1



Feature importance for Dataset 2

This feature selection technique did not change the number of clusters for k-means (3 as for PCA and RP) or components for EM by much. The graphs and clusters are similar to that of RP.

Reasons why clustering did not work well on my chosen datasets:

1. As the datasets chosen were classification problems, using them for clustering may have reduced the accuracy of the clustering model as the datasets are meant for supervised learning.
2. The datasets used had categorical data as well as continuous data. Clustering does not always work well with categorical data, as distance function between these will not give any meaningful output. I did not change from k means to KModes, as there were important continuous features present.
3. The datasets chosen from assignment 1 are not such that can be separated by clusters using any transformations.
4. For all the 4 dimension reduction techniques, when applied clustering, the clusters did not appear to be distinct and homogenous, as the datasets are not meant for clustering.
5. The correlation between features of both datasets is not strong. This affects the way PCA and RP works.

Neural Network on Dataset1 (dimensionally reduced datasets)

I chose dataset 1 to apply ANN on dimensionally reduced datasets. The accuracy of all the reduced datasets was similar but little less to that I achieved in Assignment 1. The difference was in the speed of the algorithm for finding the optimal hyperparameters which was much faster than the Assignment 1.
***For hyperparameter optimization, I had 4 activation functions + 11 different hidden layers to choose from.*

	Training Accuracy	Testing Accuracy	Hyperparameter optimization speed (in minutes)	Remarks
Assignment 1	86.93%	86.15%	10.32 mins	
PCA	85.35%	84.55%	5.55 mins	
ICA	84.01%	83.85%	6.23 mis	
RP	85.78%	86.2%	6.12 mins	
Feature Selection	84.37%	84.6%	7.89 mins	

Neural Network on both datasets (using clusters as features):

For all different combinations of data, I used cluster numbers as an additional feature and added that in the original dataset. I also one hot encoded the cluster numbers as a higher cluster number can mean more weightage if not one hot encoded.
I expected the accuracy to decrease, as clustering did not work well on both datasets. And the results also matched my expectations.

The speed for choosing hyperparameters increased but the accuracy decreased as clustering was not good which affected the way the ANN worked and decreased the accuracy by 1-2%.

Original dataset + clusters as new feature	Training accuracy (Dataset 1)	Test accuracy (Dataset 1)	Training accuracy (Dataset 2)	Test accuracy (Dataset 2)

Assignment 1	86.93%	86.15%	76.49%	75.40%
PCA + Kmeans	82.55%	82.91%	75.08%	75.40%
PCA + EM	83.77%	82.17%	74.97%	71.11%
ICA + Kmeans	82.17%	81.65%	72.23%	71.22%
ICA + EM	81.23%	81.55%	71.12%	69.28%
RP + Kmeans	82.36%	81.68%	71.8%	70.01%
RP + EM	81.52%	80.44%	70.11%	71.79%
Feature selection + Kmeans	83.88%	82.72%	70.97%	70.04%
Feature selection + EM	82.87%	82.19%	70.25%	69.90%

Conclusion:

1. K means and EM used before using dimension reduction has very low scores of completeness and homogeneity. Scaling the datasets plays an important role in clustering.
2. The Elbow method and the Silhouette method helped to choose a good k value. It is observed that Inertia should be as low as possible and Silhouette score as high as possible keeping the cluster sizes uniform and distinct from one another.
3. Plotting the components against the BIC score helped choosing the component value with lowest BIC score for Gaussian Mixture Model (EM).
4. All 4 dimension reduction techniques have different types of clusters formed even if the number of clusters were 3 or 4. For PCA variance (covering 90-95% variance) and eigenvalues are important. For ICA a high mean kurtosis is important. For RP reconstruction error is important.
5. The scores of completeness and homogeneity also increased a bit as shown in the tables throughout the analysis, which shows that the techniques had a positive impact on the datasets.
6. The Neural Networks speed increased for dimensionally reduced Dataset 1 but the accuracy was very similar to that of Assignment1.
7. The Neural Networks accuracy decreased by 2-3% where I used cluster numbers as new features, as clusters were not of a good quality.
8. I also observed that clustering did not work well on my chosen datasets as they were meant for classification problems.
9. The speed for choosing hyperparameters and the speed for training increased on the reduced datasets.

- **References**

1. https://www.codecademy.com/learn/machine-learning/modules/dspath-clustering/cheatsheet#:~:text=Inertia%20measures%20how%20well%20a,number%20of%20clusters%20(%20K%20).
2. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
3. https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
4. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
5. https://scikit-learn.org/stable/modules/clustering.html#k-means
6. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
7. https://www.kaggle.com/code/abhishekyadav5/kmeans-clustering-with-elbow-method-and-silhouette/notebook
8. https://blogs.sas.com/content/iml/2019/11/04/interpret-graphs-principal-components.html
9. https://vitalflux.com/elbow-method-silhouette-score-which-better/#:~:text=The%20elbow%20method%20is%20used,cluster%20or%20across%20different%20clusters.
10. https://www.kaggle.com/vipulgandhi/pca-beginner-s-guide-to-dimensionality-reduction/notebook
11. https://en.wikipedia.org/wiki/Cluster_analysis#External_evaluation
12. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
13. https://towardsdatascience.com/clustering-how-to-find-hyperparameters-using-inertia-b0343c6fe819