Unsupervised Learning: Dimensionality Reduction and Clustering – Sabhaee

Same datasets from assignment 1 were used for this report. Dataset 1 is the "Heart Disease Data Set" (1, 2), with 303 instance, 14 features and a binary target indicating a heart disorder. Second dataset is "Diabetic Retinopathy Debrecen Data Set Data Set" (3,4) containing 1151 instances and 20 features and a binary target indicating the sign of Diabetic Retinopathy (DR) in an images of patience retina. Both datasets are balanced. Considering both datasets are originally meant for a supervised learning problem it was interesting to investigate if there are additional pattern and classes in the data.

Part 1 - Clustering:

Clustering Methodology:

The K-Means and Gaussian Mixture (GM) methods from Scikit learn [cite] library was used for the clustering tasks in this project. The following methodology applies to all clustering instances performed in this report:

Selecting Number of clusters - K-Means

To select the optimum number of clusters for K-means clustering multiple metrics were considered. In the first metric were the "Elbow" method was used to for an elbow shape change in the slope of model's inertia vs number of clusters (k) curve. The Silhouette score curve vs K and the Silhouette diagram were the other two methods used for selection if number of clusters. For Silhouette score we look for the number of clusters that maximize this score, while in the Silhouette diagram we want all the clusters to pass the score line and preferring the cases were the clusters have roughly the same size.

Selecting Number of clusters - Gaussian Mixture

The Inertia and silhouette scores are not a good indicator of GM model quality when the clusters are not the same size or spherical. For this reason, in GM we are looking for the number of clusters that minimizes an information criterion. For this report I considered both Bayesian and Akaike information criterion (BIC and AIC). Best number of clusters are the ones that minimizes these two, but when they are different BIC will result in simpler model. Simultaneously the value of the covariance type hyperparameter was also investigated. For this purpose, first the BIC value was calculated for varying covariance types and number of clusters and the covariance type that resulted to the minimum BIC was selected. Next the BIC and AIC were plotted against the number of clusters to select the optimum number of clusters. Finally, for both clustering algorithms dataset was clustered into optimum number of clustered and was visualize by projecting the data into 2D using PCA and t-SNE and clusters were colored to get a better understanding of them.

Dimensionality Reduction Methodology:

PCA, ICA,RP and modified version of Locally Linear Embedding (LLE) method were used for dimensionality reduction (DR) tasks in this report. For all methods we are looking for the smallest number of dimensions and leas amount of information loss. The following methodology applies instances of DR performed in this report:

Selection Number of dimensions – PCA:

To select the optimum number of dimensions for the PCA method we want the minimum number of components that preserve the maximum amount of variance. For this purpose, the cumulative sum of the explained variance ratio[1] was plotted against the number of principal component and the minimum number of components that preserved %95 of the variance. Eigenvalues distribution was also plotted to obtain a cutoff point where the eigenvalue is not significant.

Selection Number of dimensions – ICA:

[1] The ratio of dataset variance along each principal component.

For the ICA we want our reduced dataset to be non-gaussian to this end the average kurtosis of the reduced dataset with varying number of dimension were plotted and the dimension with that resulted in the maximum kurtosis was selected.

Selection Number of dimensions –RP:

To select a proper dimension for reducing the dataset, the reconstruction error using multiple random seeds was plotted against the number of dimensions and minimum dimension with error of less than %10 was selected.

Selection Number of dimensions –LLE:

Reconstruction error using various number of neighbors and multiple random seeds plotted against the number of dimensions and the dimensional reduction that resulted in the minimum reconstruction error was selected.[2]

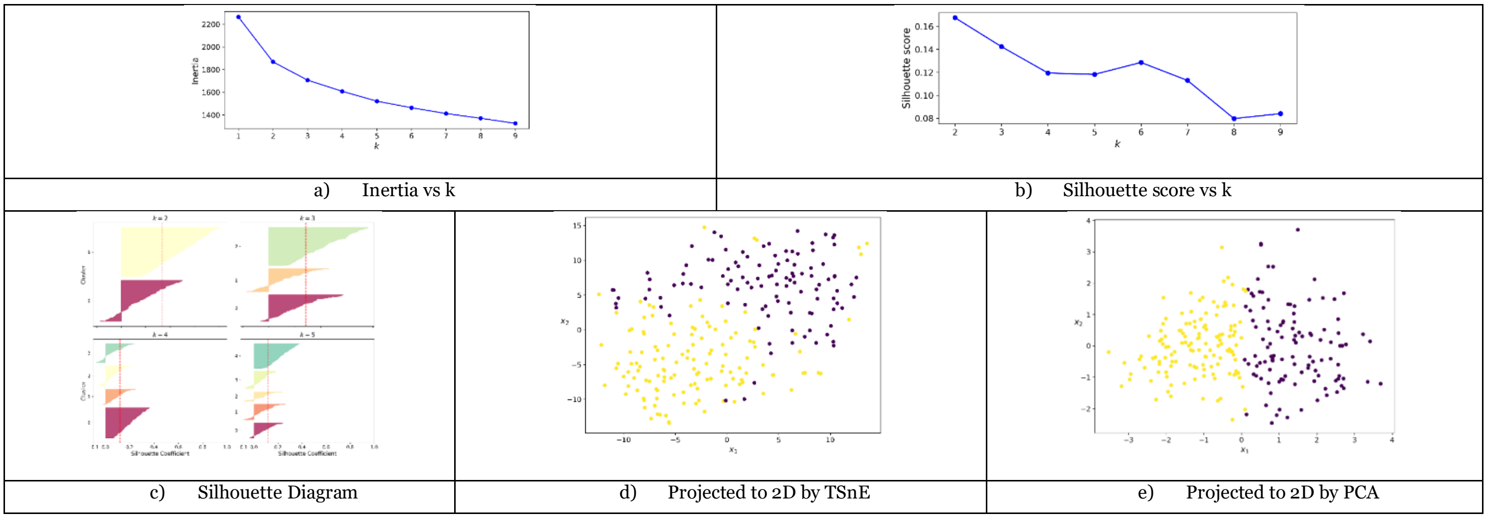
1.1 Full Dataset:

In this section the clustering algorithms were applied to the full dataset with all the original features.

1.1.1 Dataset 1:

a) K-Means

Based on the result of Elbow analysis (Fig 1.a) and silhouette score (Fig.1.b) k=2 is the optimum number of clusters. The projected cluster labels appear to divide the instances into tow cluster with a clear decision boundary. Since we have the true labels, we can investigate this further with additional metrics. For the current clustering we achieve an Adjusted Mutual Info of 0.21, adjusted Rand index of 0.29, Homogeneity of 0.22, completeness 0.21 and V-measure 0.22. a perfect labeling should receive a score of 1 while a bad labeling would be less than 0. In our case it appears we have multiple instance that are mistakenly labeled.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

d) Projected to 2D by TSnE

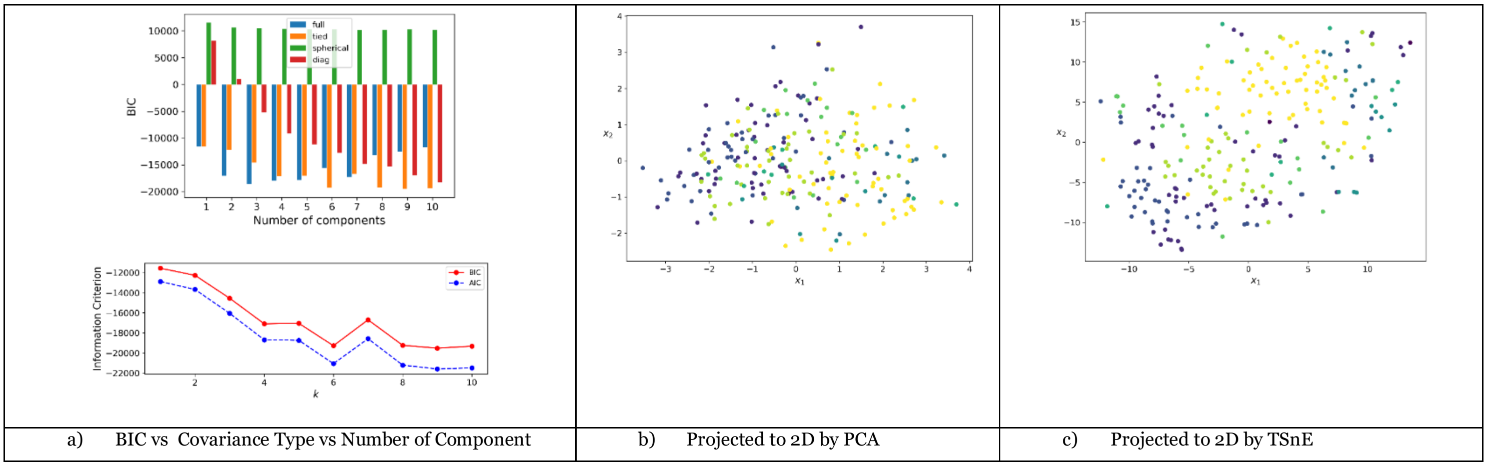
e) Projected to 2D by PCA

Figure 1- Dataset#1 – K-Means

b) Gaussian Mixture Model

Based on the Figure 2.a&b, covariance type of 'tied' with k=9 was selected. For this number of clustering the BIC and AIC both have the minimum value. The result was projected in 2D where it does not appear to capture any distinct cluster that can be investigated visually.

[2] Modified version of LLE requires the number of neighbors to be more than the number of components (dimensions) for this reason the min number of neighbors were selected as original dimension plus one.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

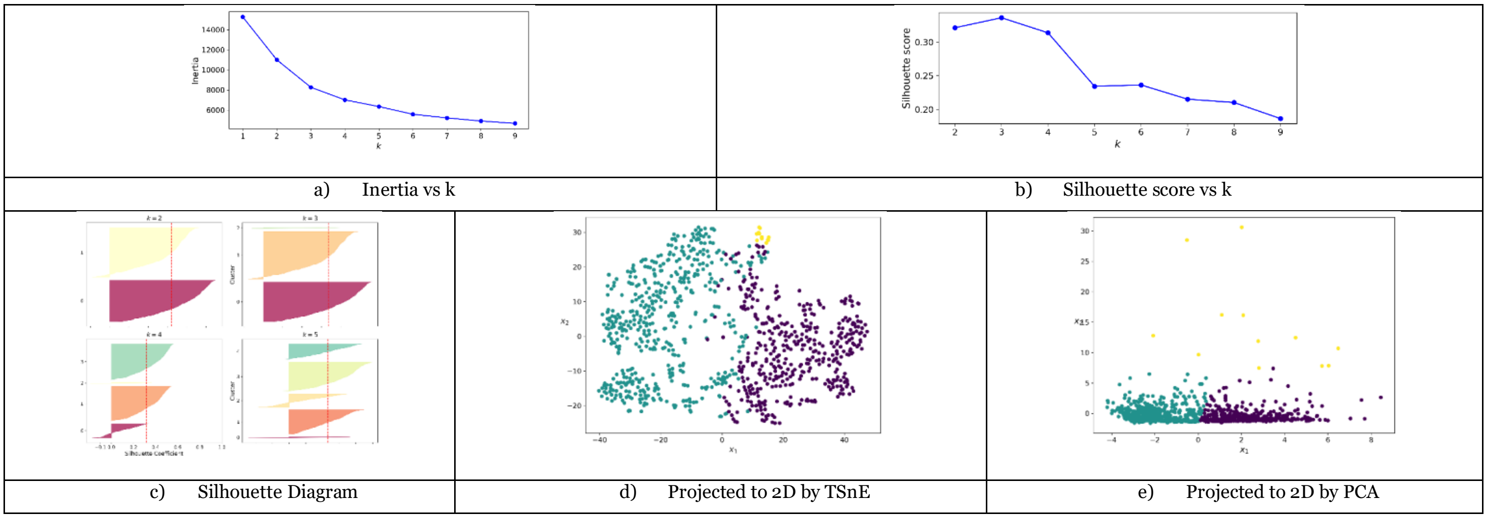
c) Projected to 2D by TSnE

Figure 2 - Dataset#1 GM

1.1.2 Dataset 2:

a) K-Means

Result of Elbow analysis and silhouette score suggest k=3 is the optimum number of clusters. Based on the silhouette diagram it appears k=2 and k=3 will result in better clusters. for the k=4 we have clusters with silhouette coefficients less than the silhouette score. Based on this it still appears k=3 to result in a better clustering. The projected result shows an interesting result where two of the clusters are almost similar in sizes while there is third small cluster that based on the PCA projection appears to include the outliers. Figure 3.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

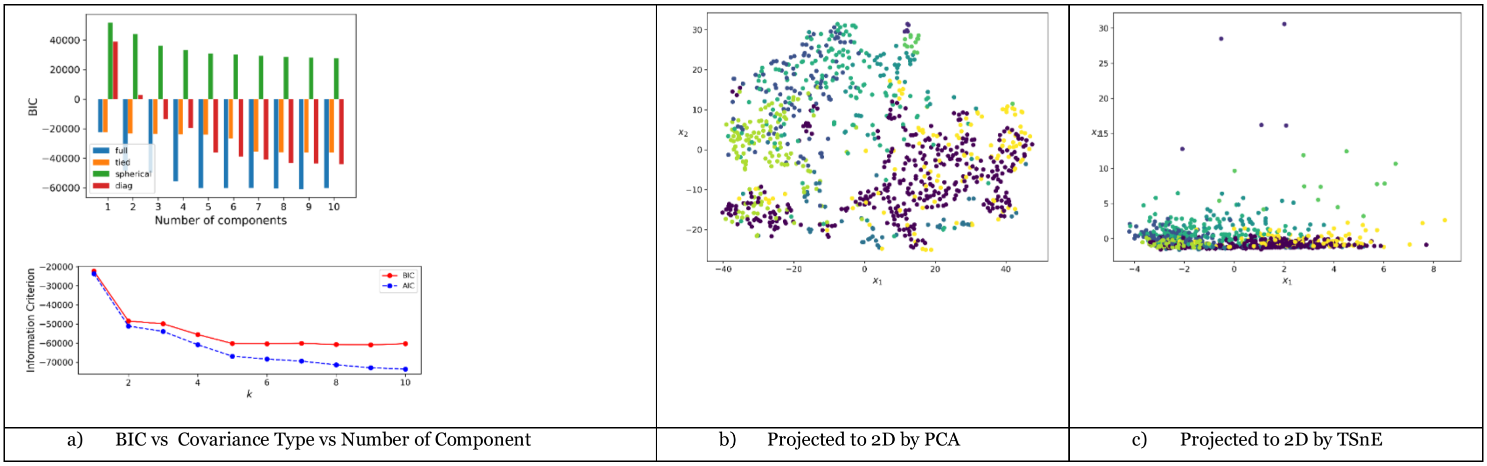
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 3 - Dataset#2 Full - KMean

b) Gaussian Mixture Model

Based on the Figure 2.a&b, covariance type of 'Full' of k=9 was selected. In this case the AIC continues to drop while the minimum of BIC is at k=9. As mentioned earlier the BIC will typically lead to simpler model and is preferred here. Projected data to 2D does not appear to capture different clustering very well.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

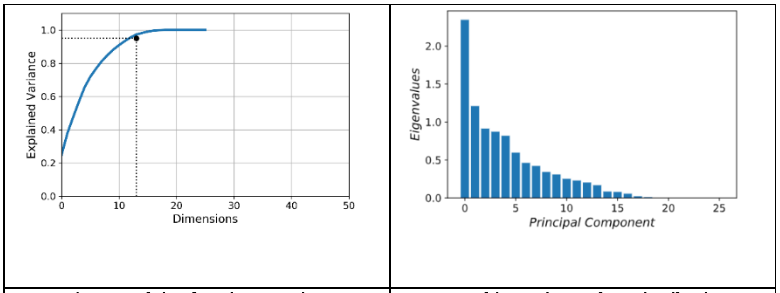
Figure 4 - Dataset#2 - Gaussian Mixture

1.2 Dimensionality Reduction:

1.2.1 Dataset 1:

1.2.1.A-PCA

Based on analysis of explained variance ratio and the eigen value distribution, 13 number of components was selected to create the reduce dataset using the PCA method.



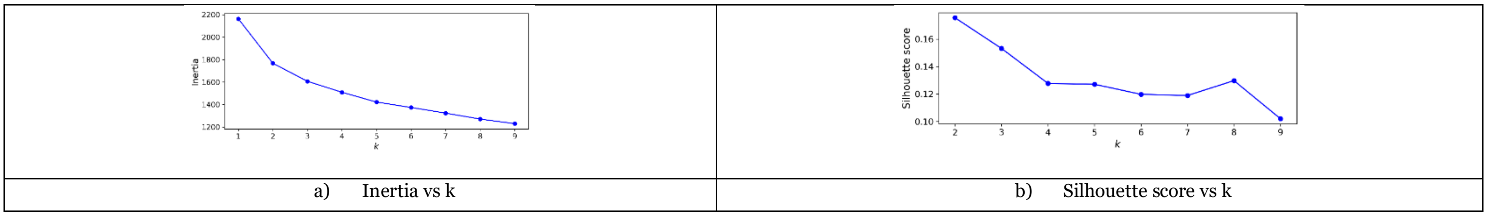
a) Explained variance ratio

b) Eigenvalue Distribution

Figure 5 – Dataset#1 – PCA – Number of components

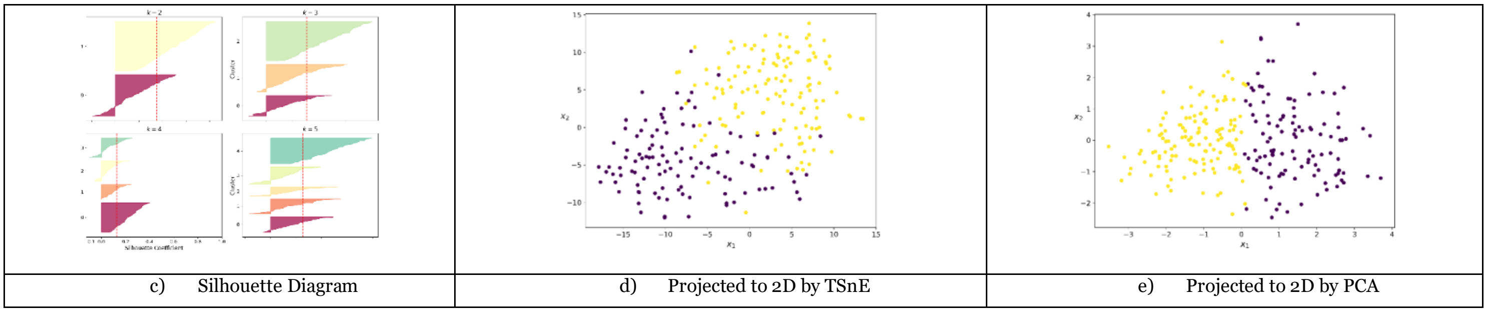
a) K-Means

Based on the result of Elbow analysis and silhouette score k=2 is the optimum number of clusters. Optimum number of clustering is similar as the case of the full dataset earlier but comparing the projected results it appear the reduced data resulted in a better clustering with a more clear decision boundary, while for the full dataset there were some instances falling too close to the other cluster. Comparing the clustering with the result obtained for the full dataset we can see the optimum. Since we have the true labels and we have only two clusters we can further evaluate our clustering. For the current clustering we achieve an Adjusted Mutual Info of 0.21, adjusted Rand index of 0.29, Homogeneity of 0.22, completeness 0.21 and V-measure 0.22. the result appears to be identical which can indicate we have preserve necessary information while reducing the number of dimension to about half.



a) Inertia vs k

b) Silhouette score vs k



c) Silhouette Diagram

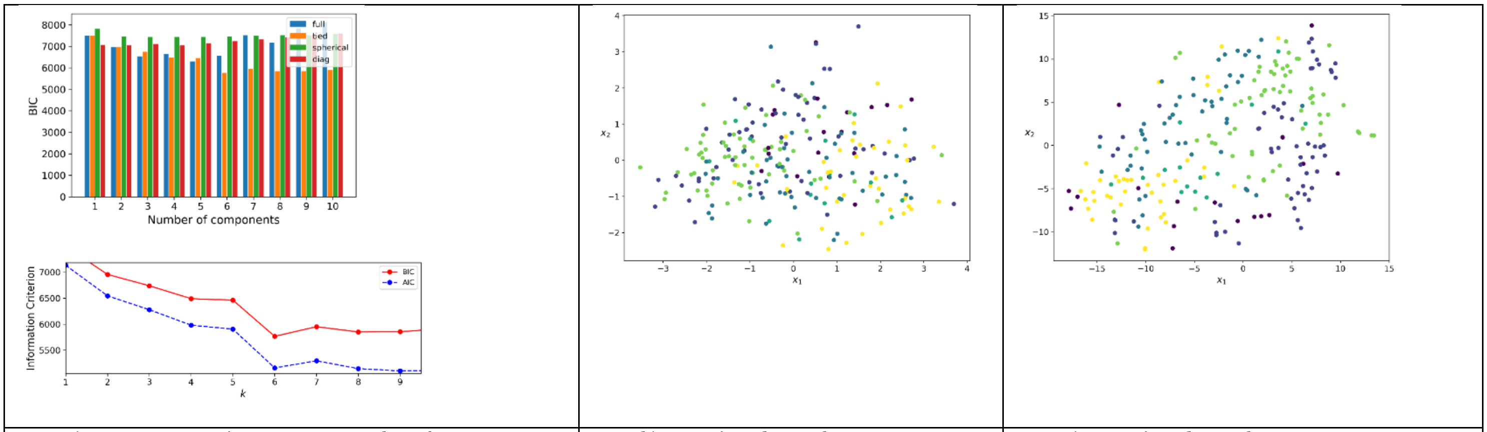
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 6 - Dataset#1 - PCA - KMeans

b) Gaussian Mixture Model

Based on Figure 7.a&b it appears the minimum value for both AIC and BIC agrees at k=6 while for AIC the k=9 result to a slightly lower value it is reasonable to go with the k=6 and a 'tied' covariance type which is less than the number of clusters suggested by analysis for the full dataset. Figure 7.c&d shows the clusters in a 2D space.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 7 - Dataset#1 - PCA - Gaussian Mixture

1.2.1.B-ICA

Based on the plot of the kurtosis value in Figure 8, ICA projection using 19 component results in the maximum kurtosis.

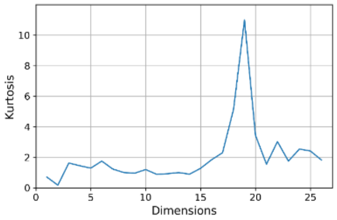
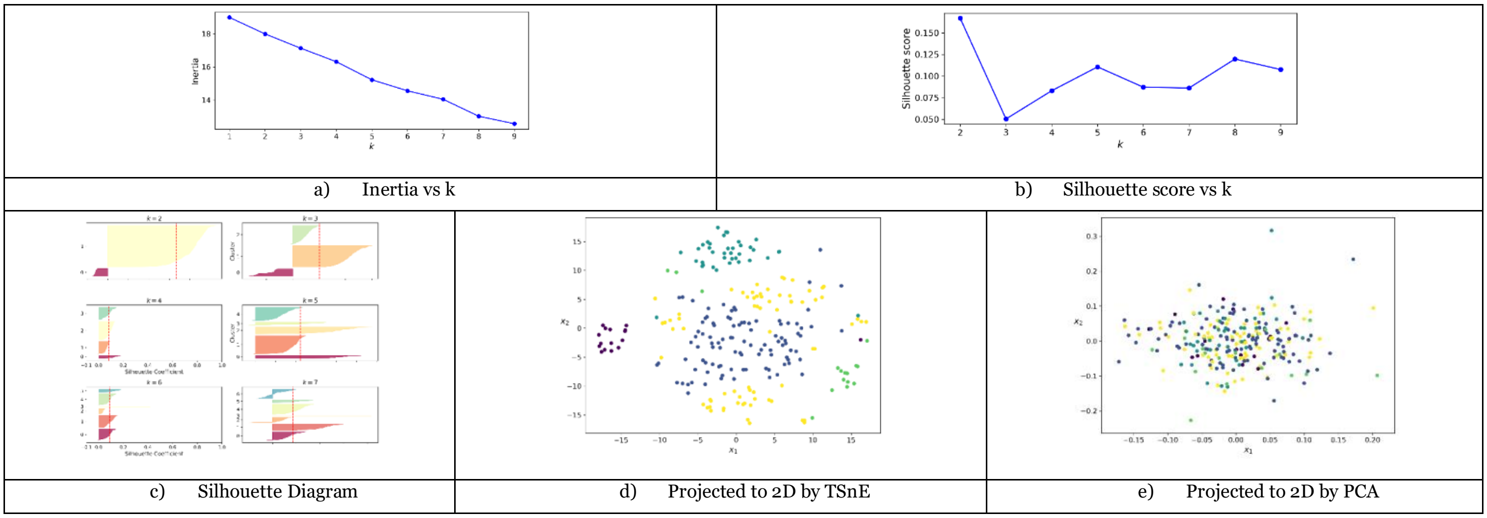


Figure 8 - Dataset#1 - ICA - Kurtosis vs num_component

a) K-Means

Based on the result in Figure 8.a -c there is no clear elbow in inertia plot but the silhouette score for k=2 will result in the max score based on these result we consider k=2 as the optimum number of clusters. silhouette diagrams suggest k=4 and k=5 will result in better clusters while k=1 and 2 have clusters with negative coefficient which indicate the higher probability of the instances assigned to the wrong cluster. For the k=6 we have clusters with silhouette coefficients less than the silhouette score. Based on these k=5 was selected. ICA projection results in higher number of clusters compares to both full dataset and PCA but the clusters visualized using t-SNE shows 4 clear clusters (Purple, Green and Navy) and one sparse yellow cluster, which was not clear in the previous clustering.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

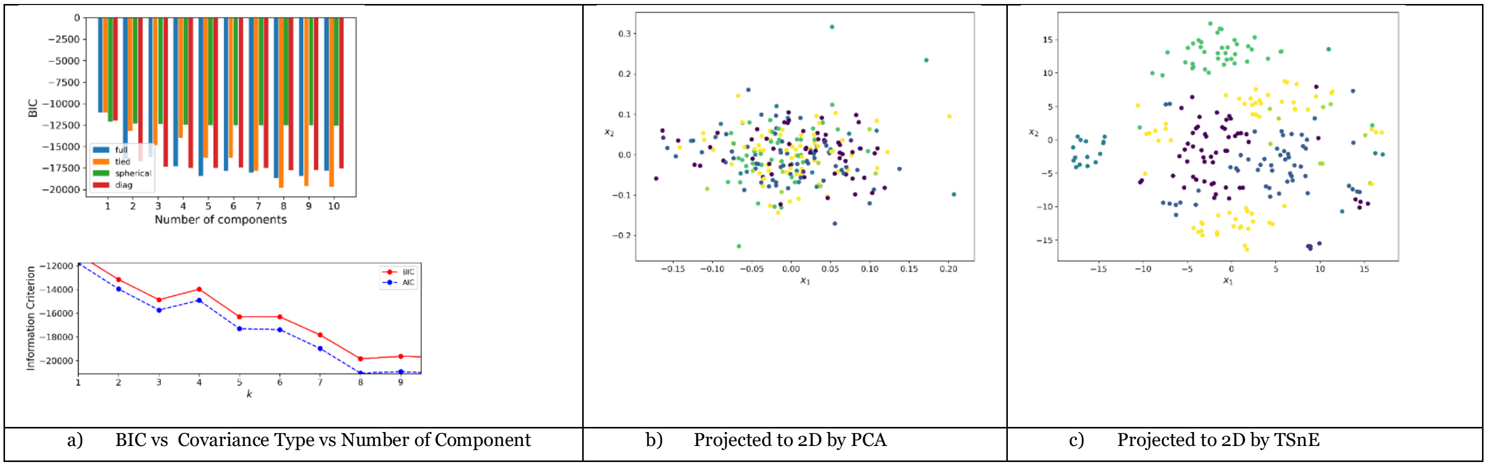
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 9 - Dataset#1 - ICA - KMeans

b) Gaussian Mixture

Based on Figure 10.a&b it appears the minimum value for both AIC and BIC agree at k=8 and a 'tied' covariance type. This number of clusters is which is less than the one for the full dataset which indicate there might have been noise in the original data that was resulting to additional clustering. Figure 7.c&d shows the clusters in a 2D space. The t-SNE again shows the isolated clusters identified by the GM method but compare to the K-Mean it seems it was the KMean that resulted in a better clustering.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 10 - Dataset#1 - ICA - Gaussian Mixture

1.2.1.C -RP

As explained in the methodology section, based on the reconstruction error threshold of %10, 20 was selected as minimum number of dimensions. Figure 11

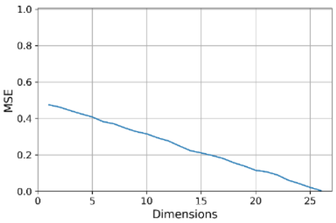
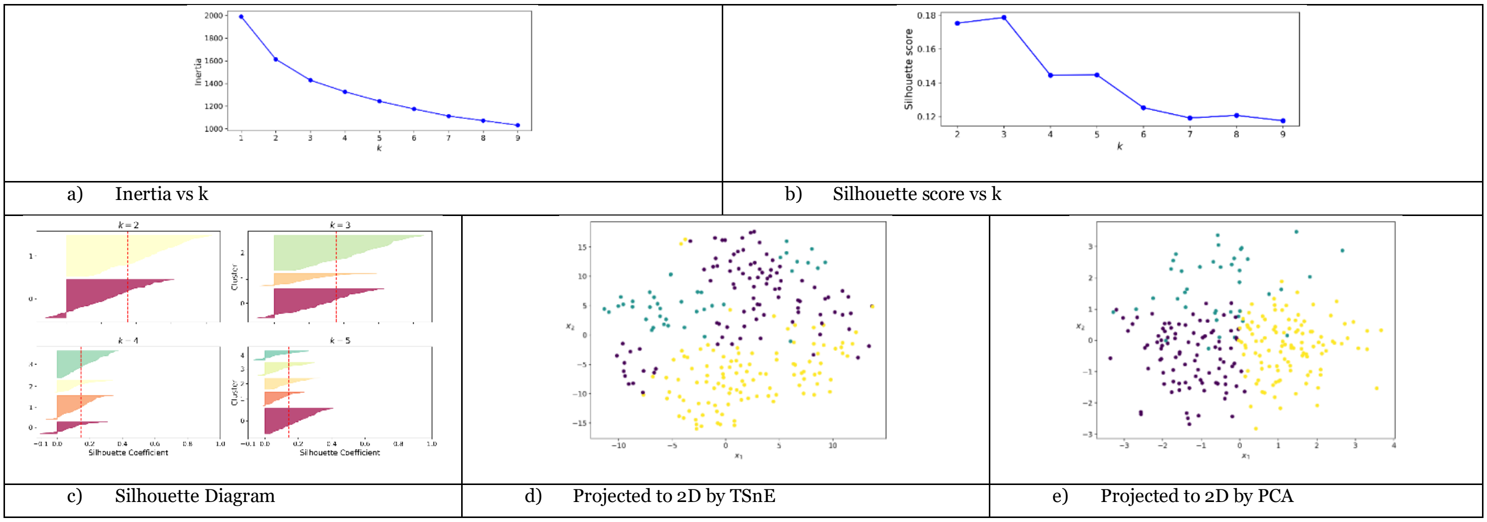


Figure 11 - Dataset#1 - RP

a) K-Means

Based on the result there is not clear Elbow in inertia plot but the silhouette score for k=3 will result in the max score based on these results we consider k=2 as the optimum number of clusters. The visualization in 2D space shows a clear clustering compare to the previous methods for dataset 1. Figure 12



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

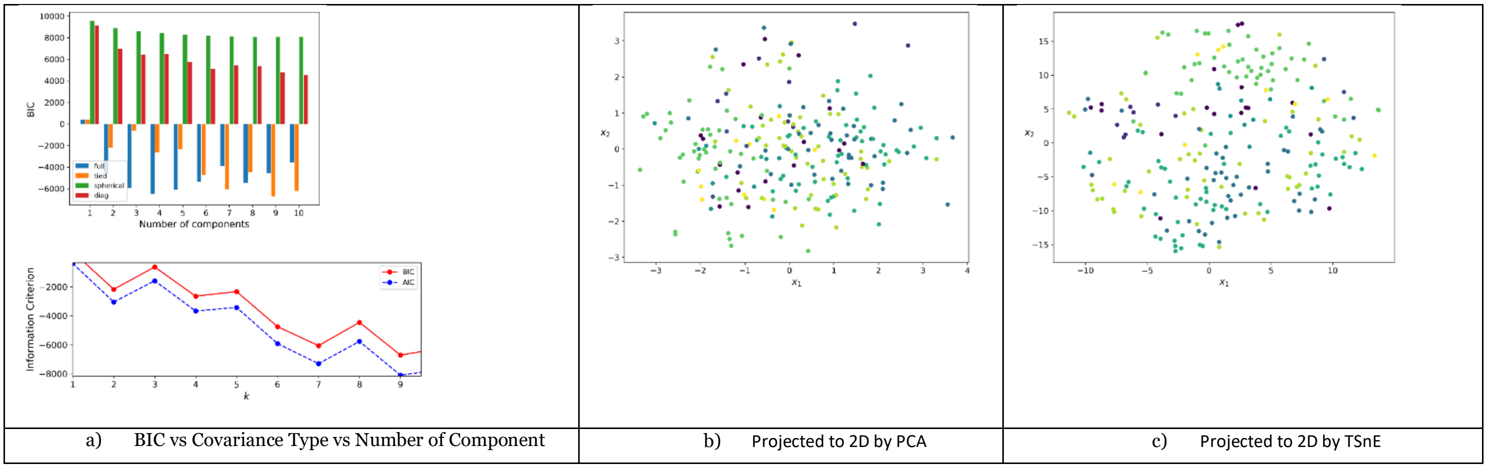
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 12 - Dataset#1 - RP - KMeans

b) Gaussian Mixture Model

It seems the minimum value for both agrees at k=9 this number of clusters match the result for full dataset and might suggest the distribution of variance in the reduced dataset is preserved in a way that still result in the same number of clusters. Visualization in 2D does not appear to present any clear cluster structure.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 13 - Dataset#1 – RP - Gaussian Mixture

1.2.1.D-LLE

The dimension (n_component=15) that resulted in the minimum value of the reconstruction error was selected for to reduce the dimensionality of the original dataset.

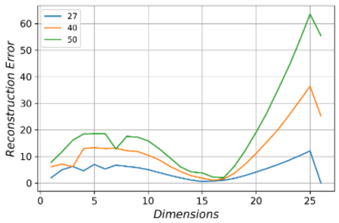
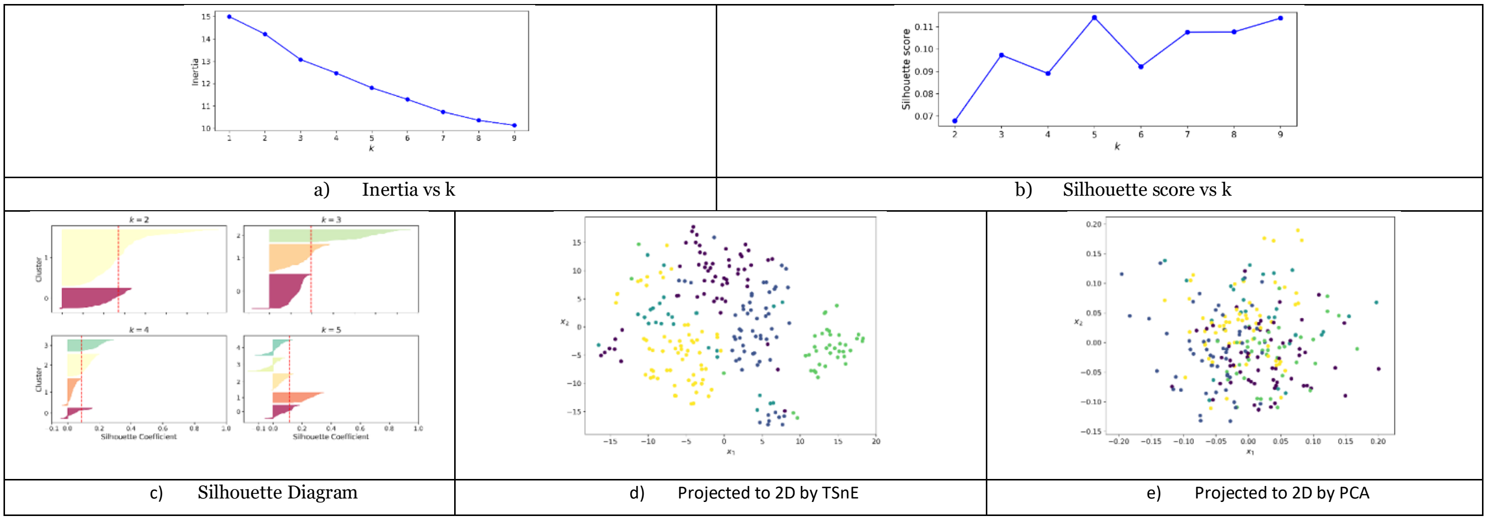


Figure 14 - Dataset#1 - Reconstruction Error

a) K-Means

Based on the result there is not clear elbow in inertia plot but the silhouette score for k=5 and k=9 will result in the max score. Silhouette diagram also suggest the K=3. I consider k=3 as the optimum number of clusters. Similar to the RP method we ended up with 3 clusters. But clusters shown in Figure 15.d shows an interesting isolated green cluster.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

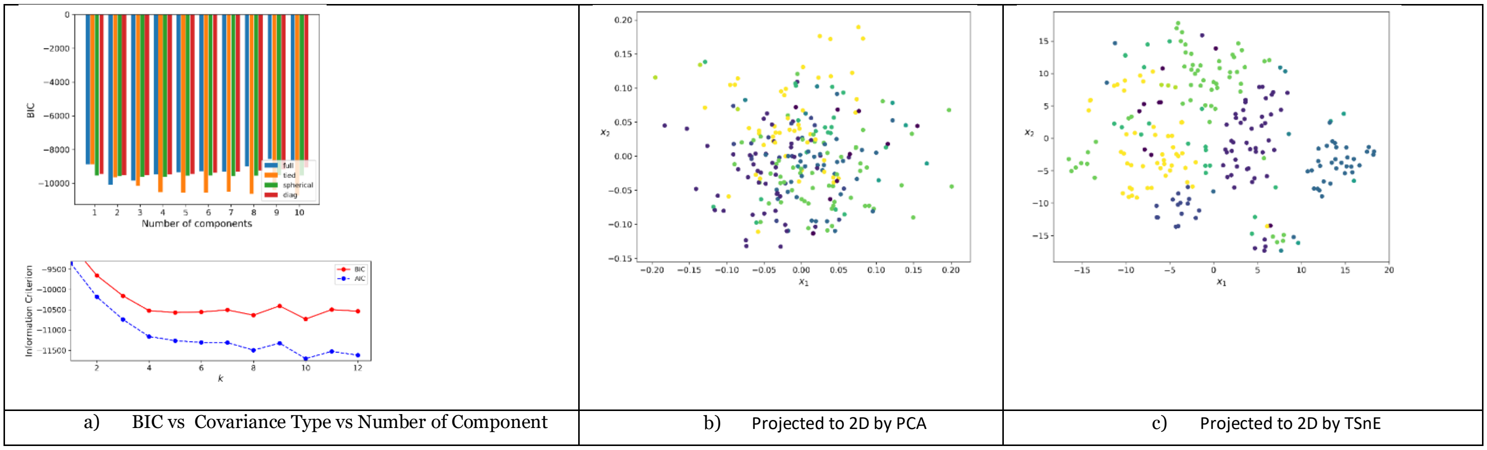
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 15 - Dataset#1 - LLE - KMeans

b) Gaussian Mixture Model

Based on Figure 16.as&b It seems the minimum value for both BIC and AIC agree at k=10 with a covariance type of 'full'. Although the optimum k was selected as 10 but based on the visualization it appears most of the instances belong to 4 main clusters.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

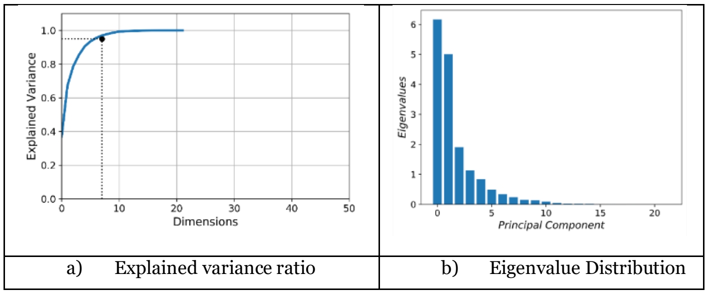
c) Projected to 2D by TSnE

Figure 16 - Dataset#1 - LLE - Gaussian Mixture

1.2.2 Dataset 2:

1.2.2.A-PCA

Based on analysis of explained variance ratio and the eigen value distribution, 7 components was selected to create the reduce dataset using the PCA method.



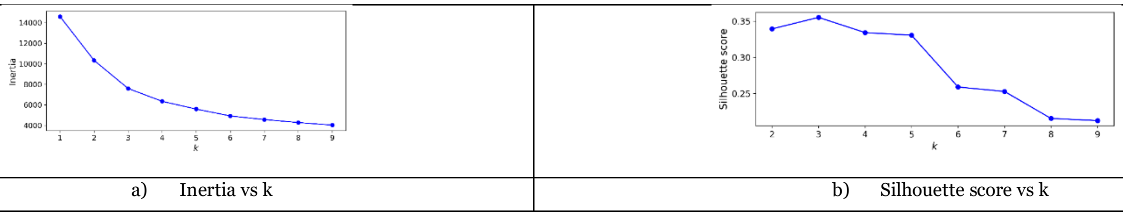
a) Explained variance ratio

b) Eigenvalue Distribution

Figure 17 - Dataset#2 – PCA – Number of components

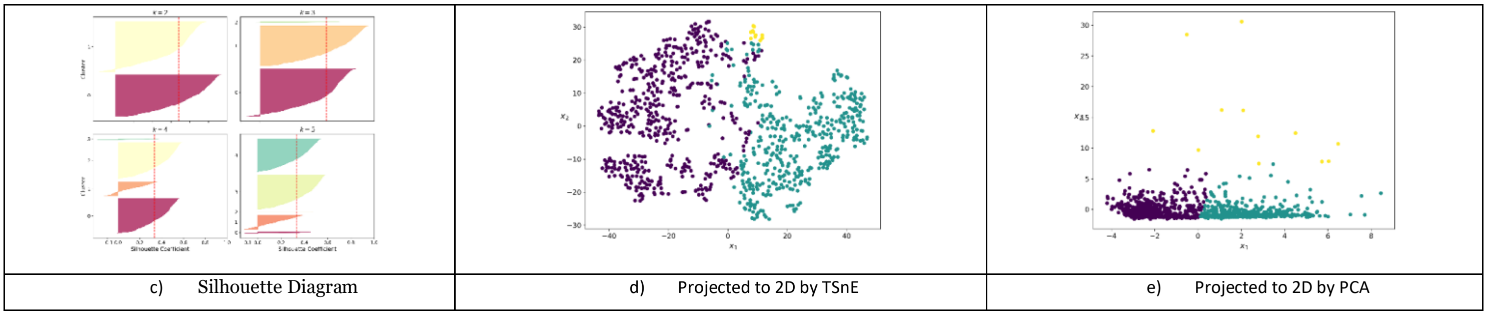
a) K-Means

Based on the result of Elbow analysis and silhouette (Figure 18.a-c) score k=3 is the optimum number of clusters. Observing Figure 18.d-e, we can see the clustering is like the clustering we achieved for the full dataset, this indicate that PCA were able to preserve the sufficient amount of variance. We can clearly see 3 separate clusters in Figure 18.d. Interestingly this might suggest there is a third class in our dataset that is not captured by original binary true labels.



a) Inertia vs k

b) Silhouette score vs k



c) Silhouette Diagram

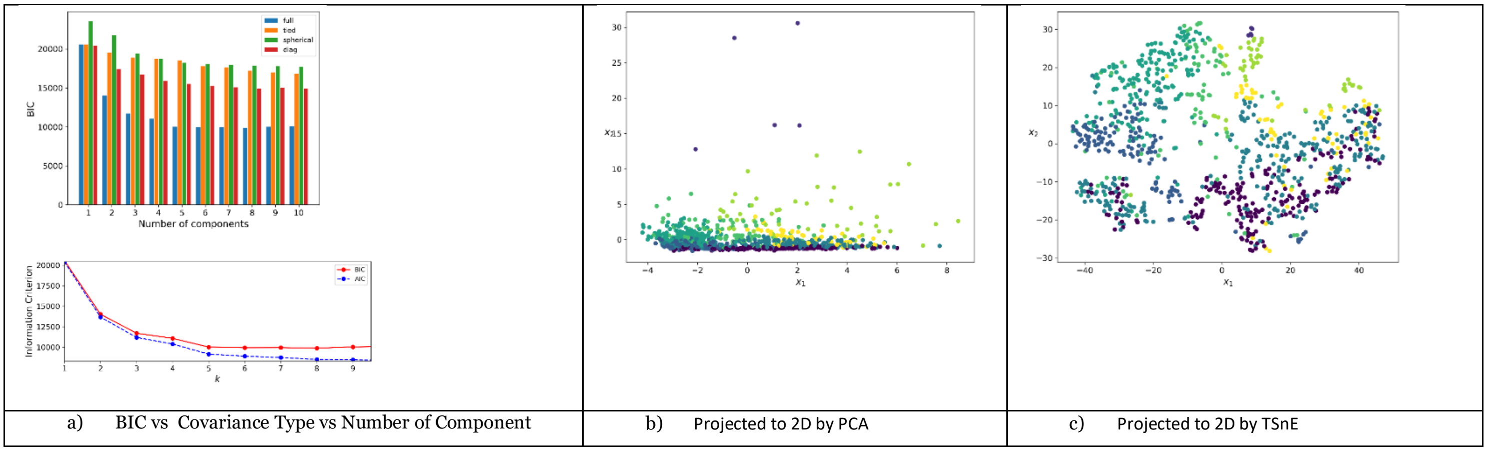
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 18 - Dataset#2 – PCA – KMeans

c) Gaussian Mixture

Based on Figure 19.a&b, BIC and AIC are minimum at k=8 with a 'full' covariance type as oppose to k=9 for the full dataset. Similarly, the 2D visualization is not capable of presenting any clear clustering structure.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 19 Dataset#2 – PCA – Gaussian Mixture

1.2.2.B- ICA

Based on the plot of the kurtosis value in Figure 20, ICA projection using 17 component results in the maximum kurtosis.

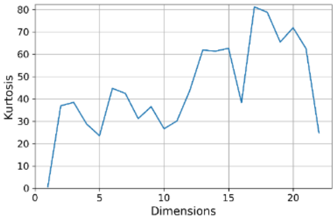
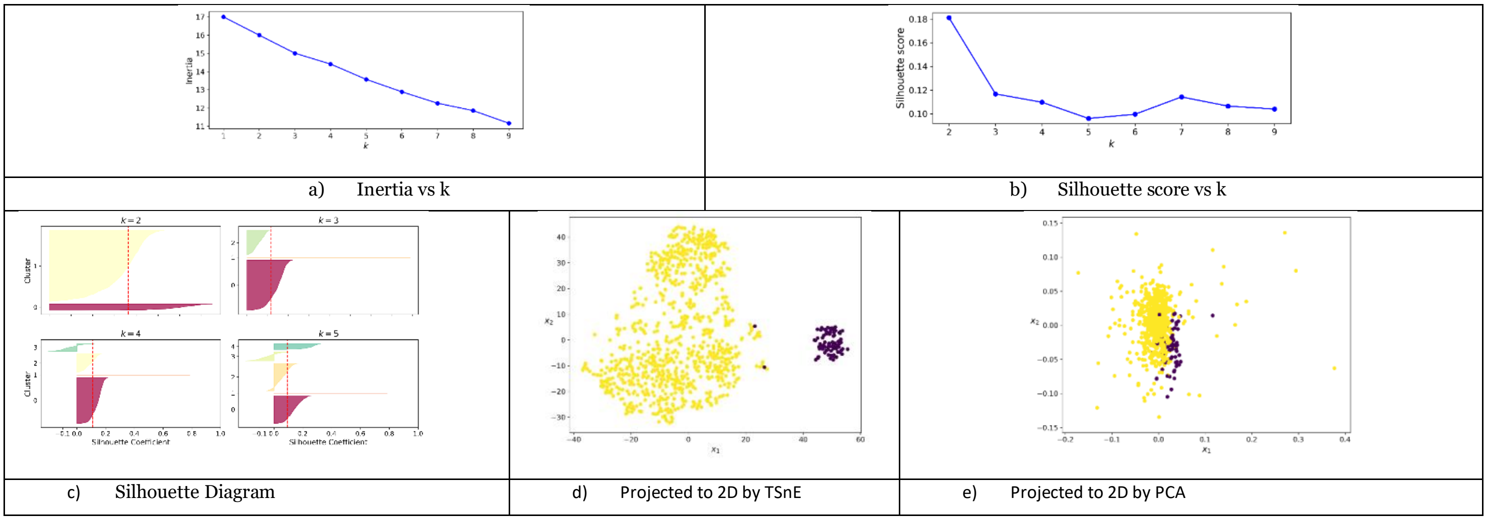


Figure 20 Dataset#2 – ICA - Kurtosis

a) K-Means

Based on the result in Figure 21.a-c there is not clear elbow in inertia plot but the silhouette score for k=2 will result in the max score based on these result we consider k=2 as the optimum number of clusters. Clustering present interesting result in Figure 21d-e, data is clearly separated in two separate clusters. For the current clustering we achieve an Adjusted Mutual Info of 0.008, adjusted Rand index of -0.004, Homogeneity of 0.006, completeness 0.014 and V-measure 0.009. While this clustering does not agree with the true labels, it present a clear grouping that might indicate additional information about the patient's retina images.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

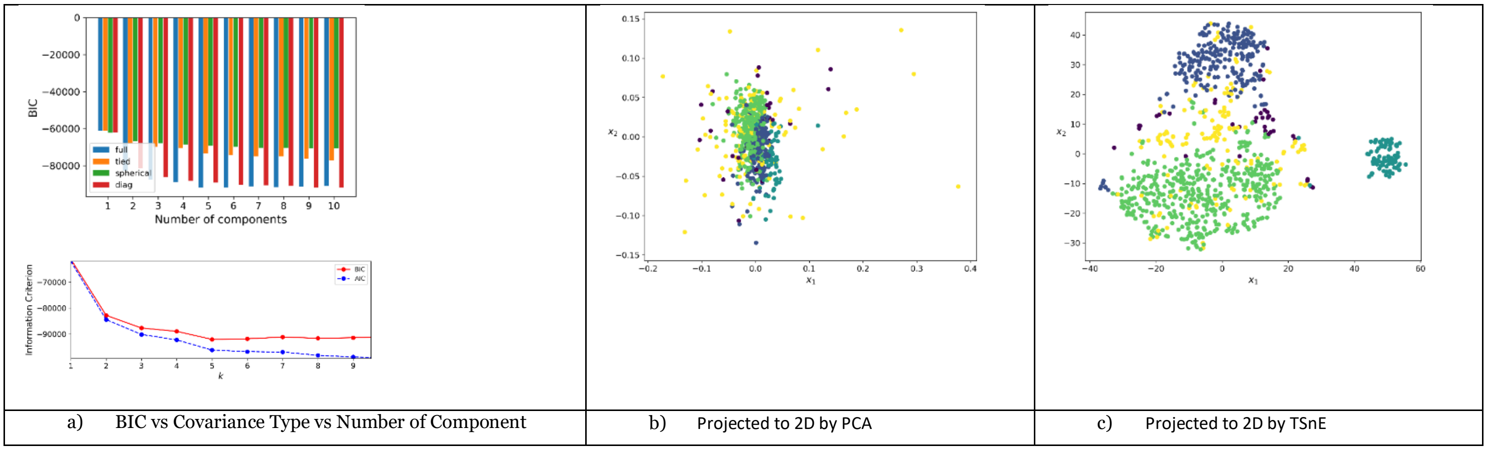
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 21 - Dataset#2 - ICA - KMeans

c) Gaussian Mixture

The minimum value BIC is at k=5 while for AIC the k=9. The simpler model was selected. Based on the clustering visualization in Figure 22 we can see interesting result for case of t-SNE 2D space with 3 distinct clusters (Navy, light green, and teal) that we were able to observe in the full dataset.



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 22 - Dataset#2 - ICA - Gaussian Mixture

1.2.2.C - RP

Based on the reconstruction error threshold of %10, 17 was selected as minimum number of dimensions. Figure 23.

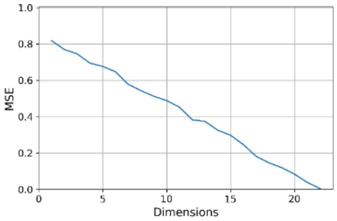
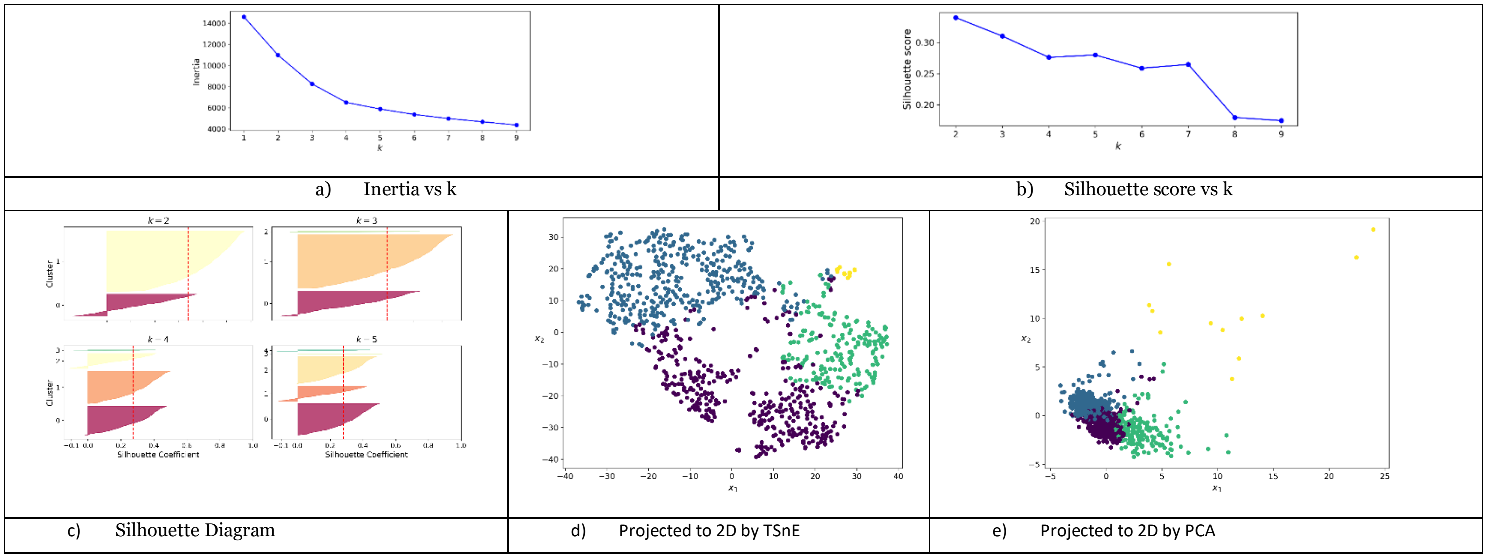


Figure 23 - Dataset#2 - RP

a) K-Means

Based on Figure 24.a&b, there is an elbow at K=4 in inertia plot but the silhouette score for k=2 will result in the max score. Based on the silhouette diagram Figure 24.c it appears k=4 will result in better clusters compare to k=2. The clustering result in 2D present interesting and clear clustering, K-means find 4 clusters compare to the full dataset and PCA reduced data that only found 2 clusters.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

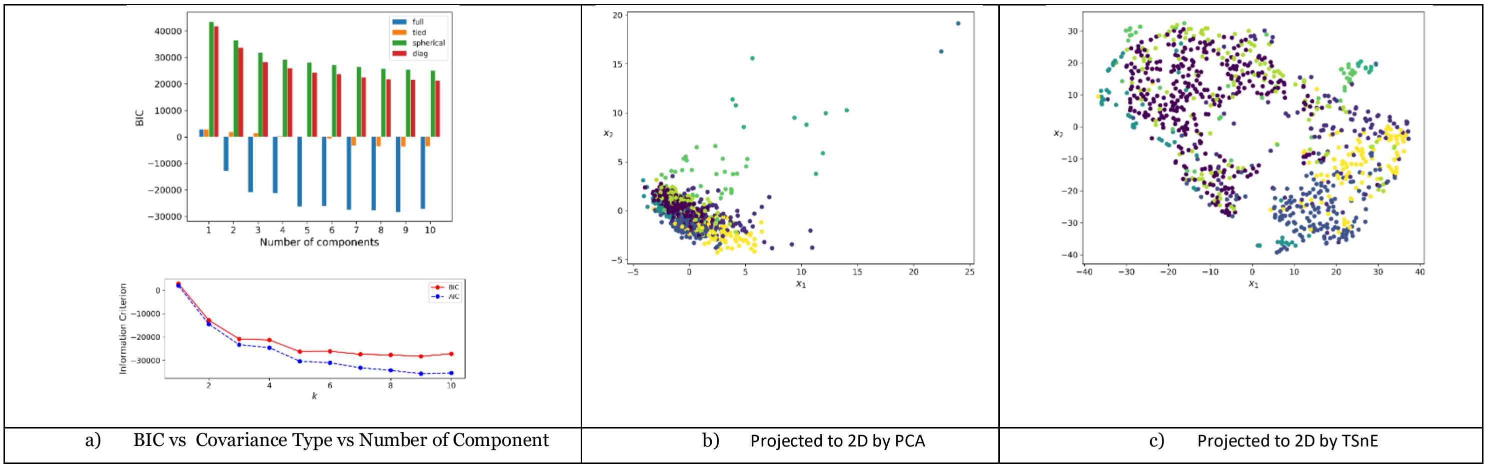
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 24 - Dataset#2 - RP - KMeans

b) Gaussian Mixture

The minimum value BIC is at k=9 while for AIC the k=11. The simpler model was selected with k=9. Clustering visualization does not appear to provide sufficient information in contrast with the result from the PCA method. Figure 25



a) BIC vs Covariance Type vs Number of Component

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 25 - Dataset#2 - RP - Gaussian Mixture

1.2.2.D -LLE

The dimension (n_component=17) that resulted in the minimum value of the reconstruction error was selected for to reduce the dimensionality of the original dataset.

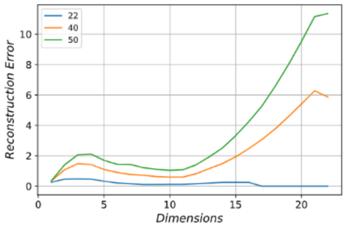
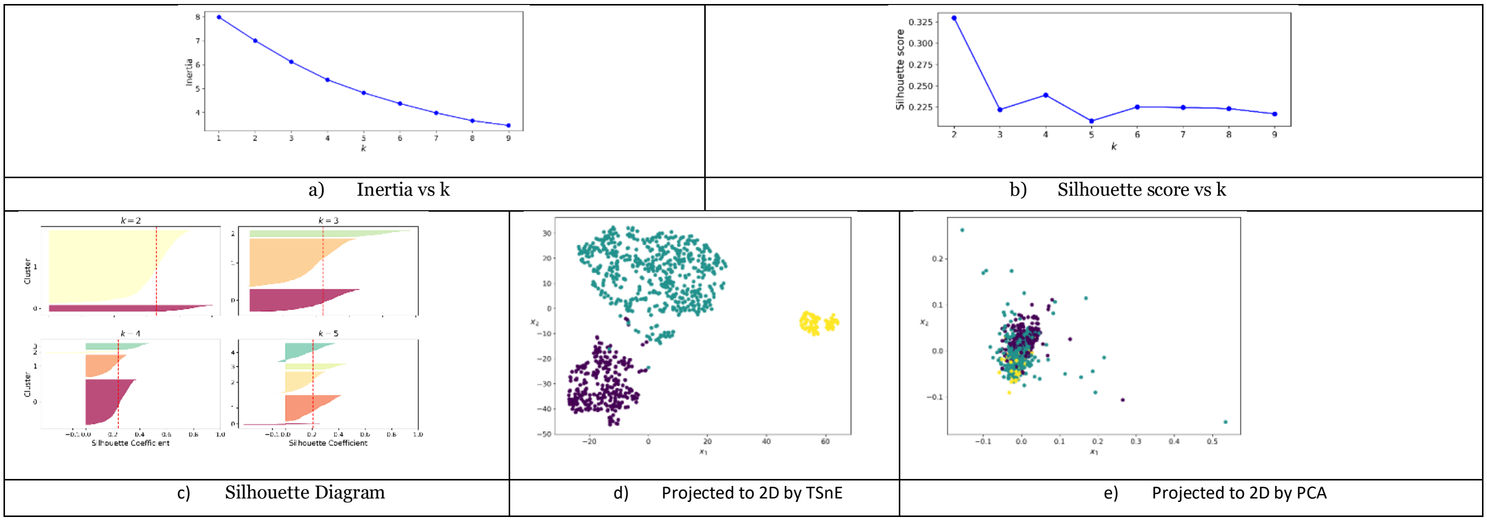


Figure 26 - Dataset#2- LLE - Reconstruction Error

c) K-Means

Based on the silhouette diagram it appears k=2 and k=3 will result in better clusters, but inertia does not present a clear elbow location. Based on this we use k=3 since it results in most of the clusters having similar sizes. The clustering result in Figure 26.d clearly present 3 distinct clusters and result in the best clustering among all DR method and the full dataset.



a) Inertia vs k

b) Silhouette score vs k

c) Silhouette Diagram

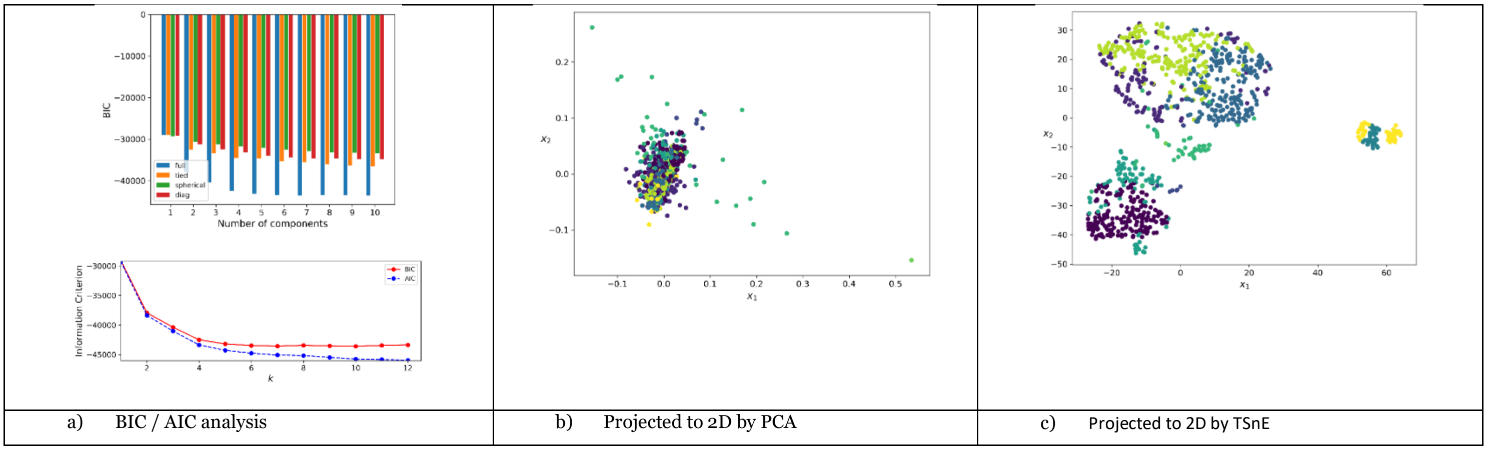
d) Projected to 2D by TSnE

e) Projected to 2D by PCA

Figure 27 - Dataset#2- LLE - K-Means

d) Gaussian Mixture Model

The minimum value BIC is at k=10 while for AIC the k=12. The simpler model was selected with k=10. Clustering visualization shows interesting grouping in the small cluster on the right of Figure 28-d that were not presented in the clustering of other DR datasets.



a) BIC / AIC analysis

b) Projected to 2D by PCA

c) Projected to 2D by TSnE

Figure 28 Dataset#2- LLE - Gaussian Mixture

Part 2 – Neural Network:

2.1 Feature Reduction By DR:

Considering the better clustering outcome in the part 1 for dataset 2, this dataset was selected for this section. A simple 2-layer fully connected neural network with 150 nodes at each layer was selected after extensive hyperparameter and model tuning. Considering the reduced datasets achieved the previous section present new problems the both the structure and hyperparameter of the models were tuned for each DR algorithms. Table 1 shows the confusion matrix for the test set for all models including the original model on the full dataset. Based on the result we can see that reduced datasets with exception of the ICA will result in a simpler structure with a smaller number of nodes and layers. This result is intuitive since the dimensionality has been reduced and the model has a smaller number of parameters to learn. Consequently, train time is reduced for a simpler model with smaller number of dimensions.

The mode trained on the the PCA and LLE reduced datasets does not perform as well as the full model while the models based on the ICA and RP have similar accuracy score on training and testing compare to the full model. ICA performs poorly on the test set while the RP performs very similar to the full model. We need to be cautious about the result of RP due its random nature, for this reason the algorithm must be run multiple times to reach a converging state.

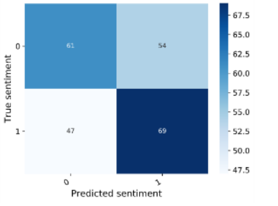
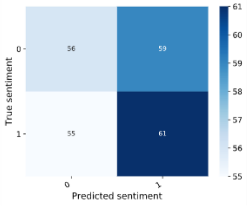
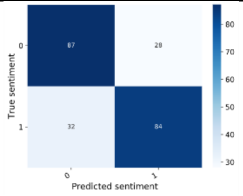
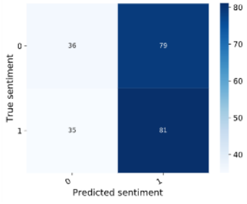
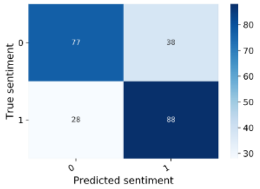
				
PCA	ICA	RP	LLE	Full
1 Layer – 100 Nodes lr = 0.005	3 Layer – 100 Nodes lr = 0.01	2 Layer – 100 Nodes lr = 0.01	2 Layer – 100 Nodes lr = 0.1	2 Layer – 150 Nodes lr = 0.005

Figure 29 – Confusion Matrix – Testing set - NN with Various Dimensionality Reduction Methods

Table 1 Accuracy result for testing set - NN with Various Dimensionality Reduction Methods

	PCA	***ICA***	***RP***	***LLE***	***Full***
Model Parameter	1 Layer – 100 Nodes lr = 0.005	3 Layer – 100 Nodes lr = 0.01	2 Layer – 100 Nodes lr = 0.01	2 Layer – 100 Nodes lr = 0.1	2 Layer – 150 Nodes lr = 0.005
Train Score	0.69	0.72	0.79	0.6	0.79
Validation Score	0.65	0.73	0.71	0.61	0.73
Test Score	0.56	0.51	0.72	0.51	0.71

2.2 Feature Reduction by Clustering:

In this section the original features of the dataset were replaced with result of the clustering algorithms from part 1 on the original dataset For this analysis two option was considered for the features: 1) the class labels predicted by the clustering as the sole features 2) Distance of each instance to center of all clusters as features. For option 1 labels predicted by the clustering were hot encoded before introduction as model inputs. For option 2 the distances were normalized.

Results are presented in Table 2. We can clearly see the inclusion of distances as the features is a better option while the models trained on the labels appear to work only slightly better than chance (K-Mean-Label model classify all instances except 1 as True). The result of the GM-Distance model is comparable with the performance of the model trained on the reduced dataset using PCA. This suggest the clustering algorithms can be used as an alternative DR method.

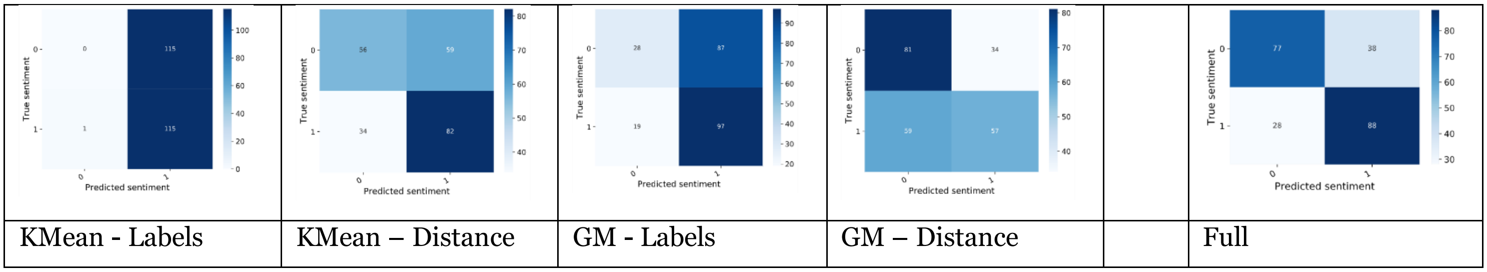


Figure 30 – Confusion Matrix – Testing Set - NN with K-Means and Gaussian Mixture as DR Methods

Table 2 Accuracy Result - Testing Set -NN with K-Means and Gaussian Mixture as DR Methods

Model Name	***K-Mean - Labels***	***K-Mean – Distance***	***GM - Labels***	***GM – Distance***	***Full***
Model Parameter	1 Layer – 50 Nodes lr = 0.001	1 Layer – 50 Nodes lr = 0.001	1 Layer – 50 Nodes lr = 0.005	1 Layer – 50 Nodes lr = 0.005	2 Layer – 150 Nodes lr = 0.005
Train Score	0.53	0.59	0.61	0.63	0.79
Validation Score	0.49	0.58	0.57	0.64	0.73
Model Name	K-Mean - Labels	K-Mean – Distance	GM - Labels	GM – Distance	Full

Reference:

1) https://www.kaggle.com/ronitf/heart-disease-uci

2) https://archive.ics.uci.edu/ml/datasets/Heart+Disease

3) https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

4) Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.

6) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, ISBN: 9781492032649