Unsupervised Learning Analysis

Maseerah Mahboob Khatoon

NEU ID: 002778147

Northeastern university, ON

In this report, I have analyzed and explored the performance of unsupervised learning techniques. There are 2 clustering algorithms and 3 dimension reduction techniques that I have implemented on 2 datasets. The comparison of the implementations on both datasets is stated along with a brief exploration of the algorithm. Relevant graphs and outcomes are attached along as well.

Dataset 1:

For the first dataset I have used is the Heart disease dataset. It consists of 303 instances with 14 features. The "target" field refers to the presence of heart disease in the patient which is classified as 0 or 1 for the presence of a heart disease.

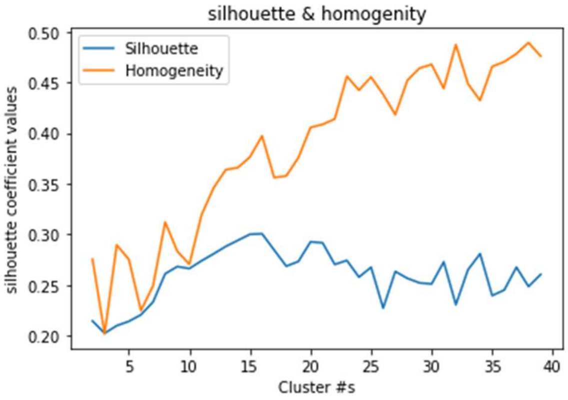
Dataset 2:

For the second analysis, I am using the Potable water dataset. This dataset includes 3276 instances with 10 attributes. The dependent variable in this also is 0 or 1 referring to the positive or negative potability of water. The data is cleaned and categorized into binary forms before clustering. I have also normalized the data with continuous values to keep the data consistent.

Clustering Algorithms on Dataset 1

1. k-means Clustering

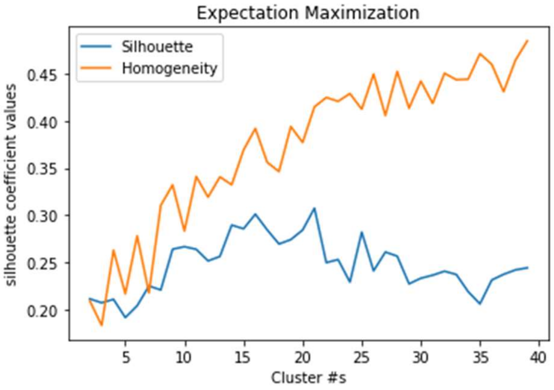
K-means is a centroid based clustering algorithm where the goal is to identify k number of groups in dataset. To get a picture of the distributions, I used silhouette and homogeneity coefficient values. The graph below shows k=16 gives a high spike in homogeneity with coefficient values as 0.3.



Overall k=16 would seem to be a good cluster. However, such a value did not yield good accuracy. Since the heart disease data is small, the distribution did not help much to get better accuracy. Therefore, k=2 results in highest accuracy of 0.79.

2. Expectation Maximization

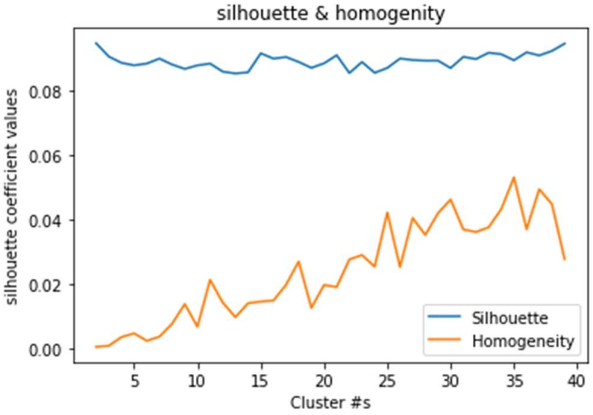
This algorithm considers the variance of Gaussian Mixtures that will describe the cluster shapes; however, due to the heart disease dataset being quite small, I expect the cluster results would be similar to that of K-means. As seen in the graph, both trends of silhouette and homogeneity are going in almost a same direction as that of k-means. Here the k=21 has the peak silhouette coefficients.



Clustering Algorithms on Dataset 2

1. K-means clustering

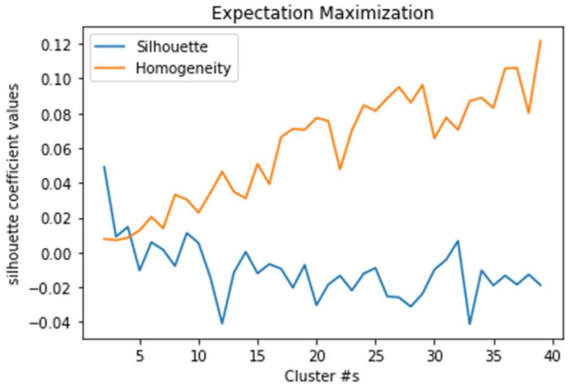
The potable water dataset has large data so its expected that there are low correlation features across the data. Overall, the silhouette and homogeneity coefficients are seen to be very low, less than 0.1. This means that this dataset has the widely distributed data clusters and the same class label observations are not in the same clusters. This could be due to low correlation within the data features.



As per the graph, the silhouette coefficient is highest when k=3 or k=40. The efficient number of clusters for the large data would be 3 clusters rather than 40 different clusters. The accuracy gained here is 0.51 which is not high but this maybe because of randomness of the data.

2. Expectation Maximization

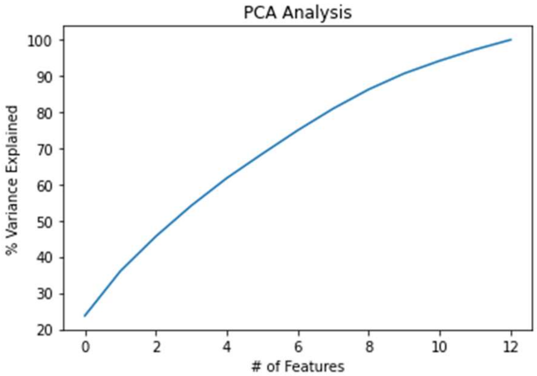
The number of clusters giving the highest silhouette coefficient (that is 0.05) is 3 similar to k-means technique. But here homogenity goes beyond the silhouette values due to the gaussian variance. However as the coefficient values are very low, the difference is barely noticeable.



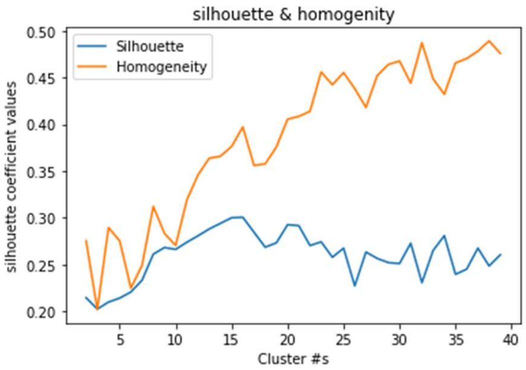
Dimensionality Reductions on Dataset 1

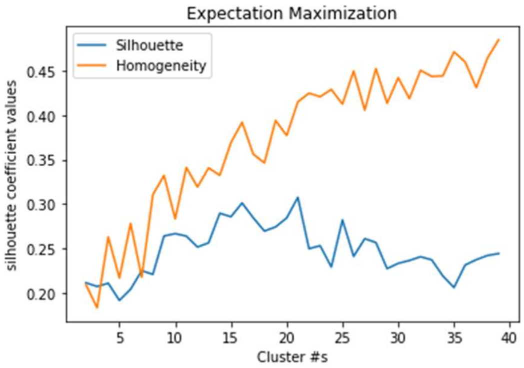
1. PCA

For reduction using PCA, the eigenvalue of variance % is used such that the graph below describes its variance %. Even with a few attributes, the graph has a slight curve shape such that there can be seen a diminishing return on the component value of 12.



Upon implementing the two clustering algorithms on the reduced heart disease dataset yields the below graphs.

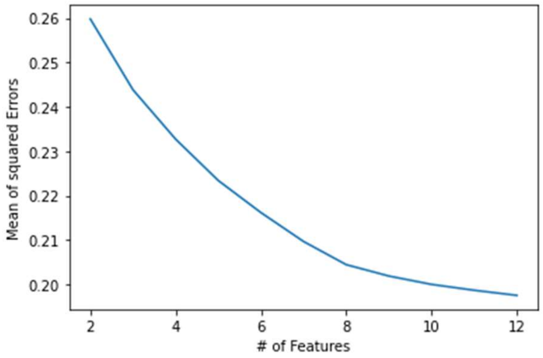




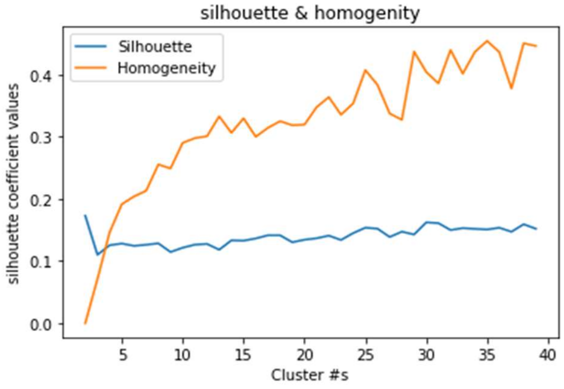
For k-means clustering, the best silhouette coefficient is 0.30 with k=16. This is very much similar to the result obtained prior to reduction. Similarly for the EM clustering the best coefficient is 0.3 at 21 clusters. It can be noted from these that PCA did not make much of difference for clustering the Heart disease dataset. This is due to the attributes/features of the dataset being relatively small.

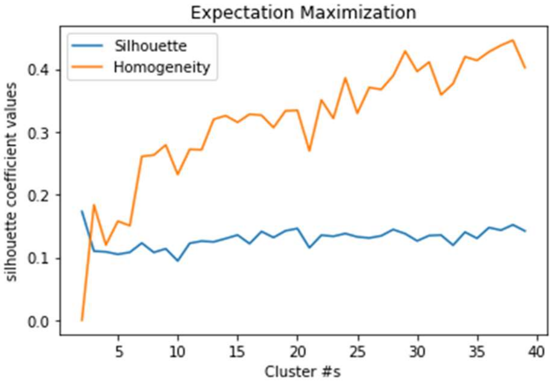
2. ICA

In ICA reduction algorithm, I have used mean squared errors to find the number of features. The below graph shows a non linear curve. At 11 features comparatively lower slopes are observed, hence I take ICA's n_component=11 for this dataset.



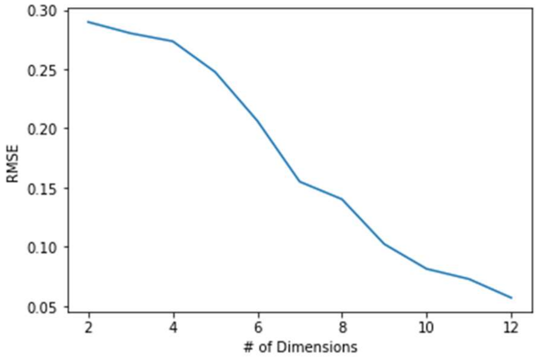
With the dimension reduction, comparatively lower silhouette coefficients are observed than the original data. The EM clustering also depicts lower silhouette scores which is lower than 0.2. It can be concluded that reducing the features for this dataset using ICA has not been helpful.

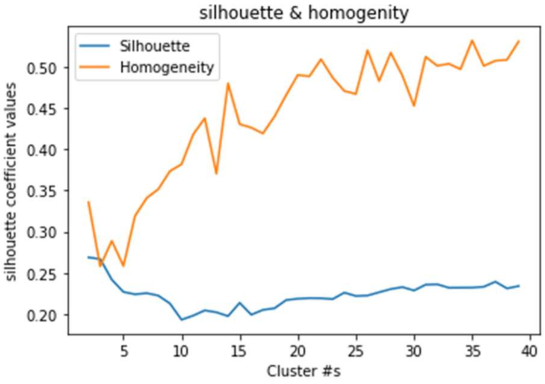


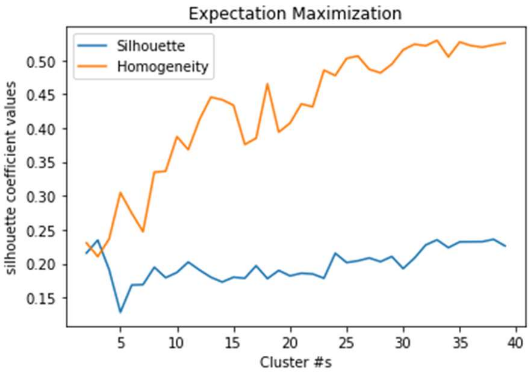


3. Randomized Project

In randomized project, I used the reconstruction error. The graph turns out to be kind of linear, I opted for dimension of 7 as the slope of the line seems to be less steep from this point.





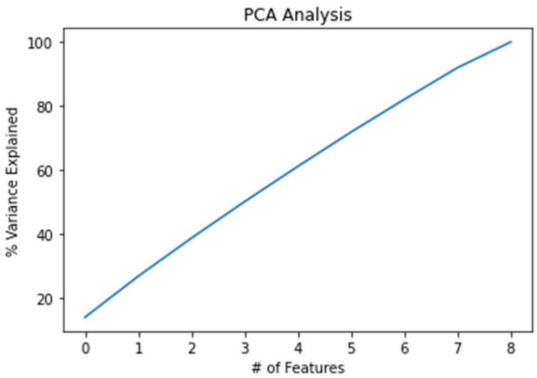


The output results for k-means clustering is quite similar to the original data such that the highest coefficient is close to 0.3. The best coefficient for EM clustering also are close to the original outcome which is 0.3 for the maximum silhouette. Both the clustering algorithms show that clusters of 3 will give the best cluster value for the heart disease dataset.

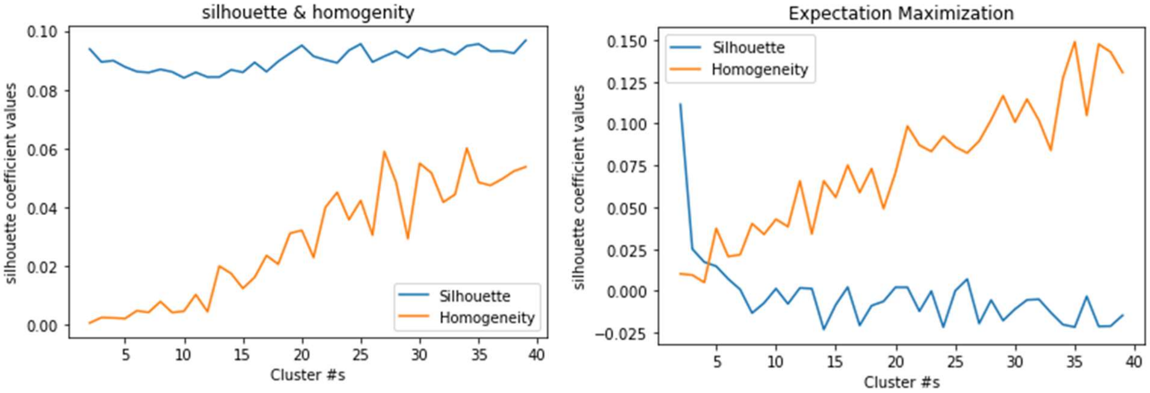
Dimensionality Reductions on Dataset 1

1. PCA

The eigenvalues in PCA Analysis shows a linear graph which is not helpful to reduce the dimensions, however for analyzing, I chose the dimension of 7 as a slight curve is seen in the slope.



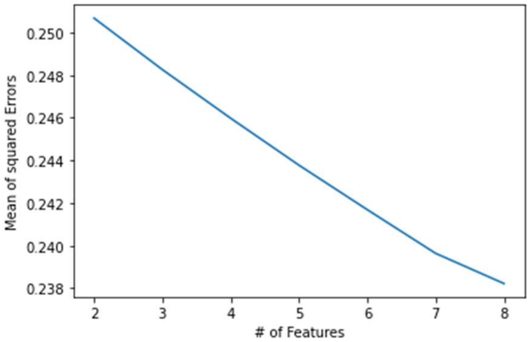
The implementation of clustering shows same results as that of the original data. As in k-means clustering, the silhouette coefficient is less than 0.1 and the highest cluster score would be 3.

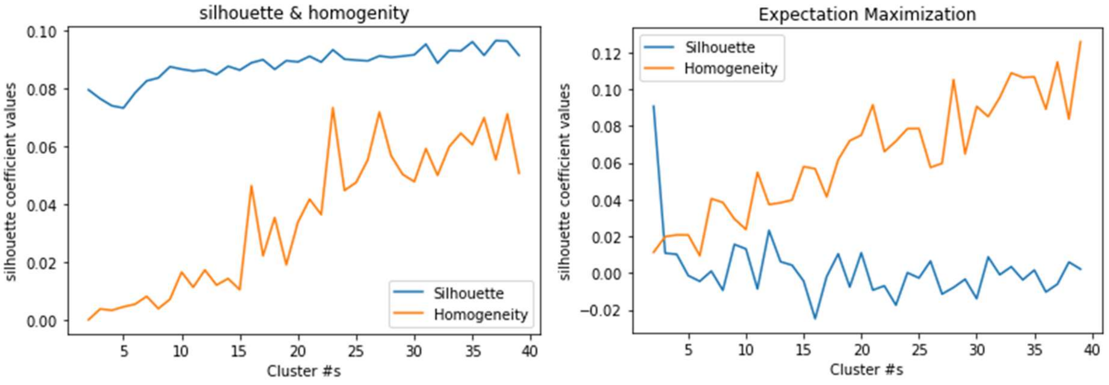


The EM value in Expected maximization clustering is slightly different but the silhouette line is quite similar. A drastic drop can be seen at k=7 but the best silhouette coefficient is at k=3 same as the original data.

2. ICA

The ICA's Mean of squared error line is linear just like we obtained for PCA. It might be better to keep all the dimensions but I picked value 7 from the features as there is a slight curve of line at that point and used that value to run the clustering.

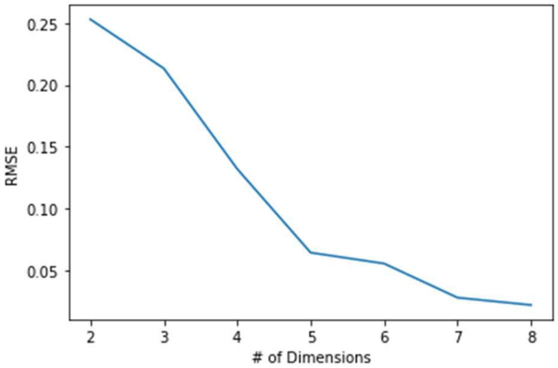




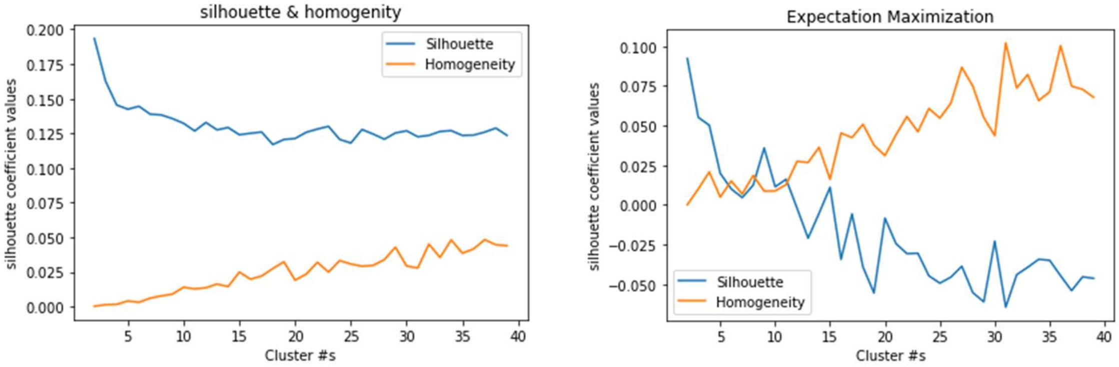
In k-means clustering, there is a slight difference in silhouette value but it does not go beyond 0.1. The better outcome is still with k=3 as there is very little difference in the coefficients (0.08 and 0.095). The EM clustering also has similar results where the silhouette value drop at k=7 and is the highest at k=3.

3. Randomized Project

Using reconstruction error, the graph obtained for randomized project is linear. Dimension of 7 is picked as the slope of line gets steeper at this point.



When clustering is implemented on the reduced data, some significant changes are observed. The silhouette coefficient has increased, almost reaching 0.2.



Reducing the dimensions with this technique has put the distributions in closer clusters. Still though with k-means clustering, the number of clusters of 3 gives the highest coefficient. With EM clustering, the coefficients are slightly improved but the silhouette trends are kind of same as the original one.

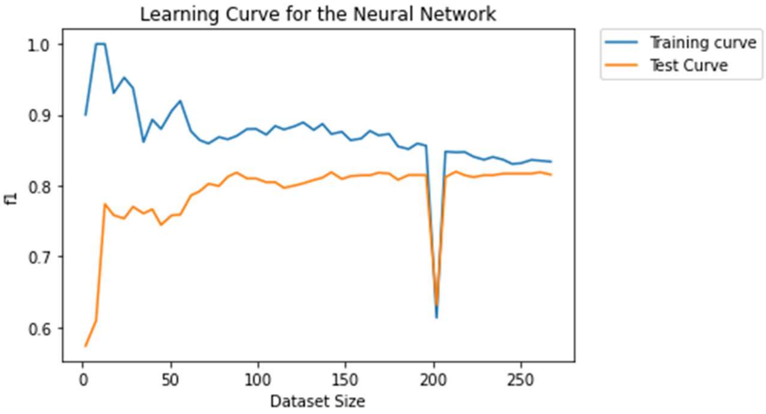
Overall, the dimensionality reduction on both datasets have less impact. This is small number of attributes/features. Data observation of same class labels maybe in different clusters and correlations of feature is randomized and low in such case.

Neural Networks with DimensionalityReduction

For the Neural Network, I used the assignment 2's Titanic dataset. It had the better f1-score and clear bias variance relationship between the training curve and testing curve. I expect to see some improvements on the curve gaps.

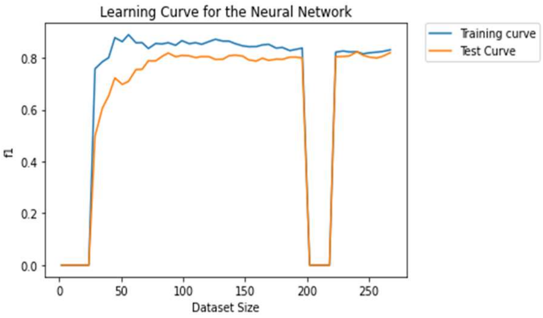
1. PCA

With the new dimension of the PCA, I got the different hidden layers parameters. The optimal hidden layer has become the one hidden layer with 8 nodes. The learning curve that resulted shows a slight shift upwards for the test curve.



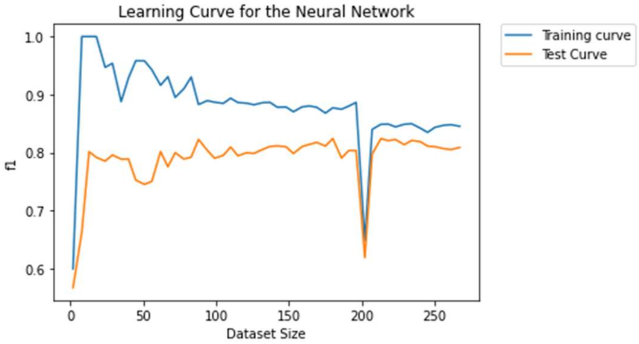
2. ICA

After implementing ICA, different neural networks parameters are yielded. The gaps between training and testing curve is reduced which means that the dimensionality reduction via ICA has increased its variance and lowered its bias of the dataset.



3. **Randomized Project**

After reduction of dimensions, the learning has improved for the data with the best score being 0.81. An optimal f1 score is seen here.

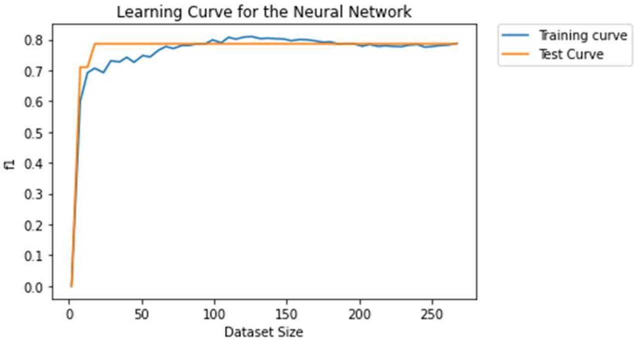


Hence it is observed that dimensionality reduction has reduced the gaps between training and test curves as compared to the original data.

Neural Networks with Clustering Techniques

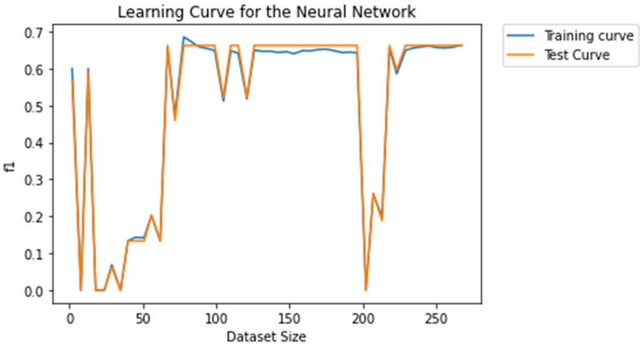
1. K-means

Here, I ran the algorithm with different hidden layers from our NN. It was found that the learning curve for two hidden layers with 100 nodes shows prominently that some reduction of bias has occurred.



2. Expectation Maximization

Upon trying a few combinations of neural networks layering, the EM clustering and hidden layers of (50,50) resulted as shown in the graph below; the training and testing curves almost identical.



Conclusion

After performing dimensionality reduction with different algorithms on the two datasets initially, not much difference was noted in the clustering outcomes. This was due to less data and fewer attributes.

Also, on reducing dimensions of the Titanic dataset, a prominent difference was visible in running the neural network. An improvement is noted as a result of the lower dimensions allowing low spaces to perform.

Hence, I conclude my report with the learning that smaller and randomized data hardly has any improvement upon clustering after dimensionality reduction. It is also clear that neural network model mostly gives a better result upon clustering and running the neural network.