Unsupervised Learning and Dimensionality Reduction

Jacob Seo
CS7641
jseo302@gatech.edu

Abstract

This assignment discusses unsupervised learning and its dimensionality reduction. It contains two clustering algorithms that are K-means and Expectation Maximization, three dimensionality reduction algorithms that are Principal Component Analysis, Independent Component Analysis, and Randomized Projections. It also includes the Neural Network implementation with the clusterings and the reductions.

I. Dataset Descriptions

A. Heart Disease Dataset

This is the same dataset that I used for the previous two assignments. There are 303 instances along with 14 features/attributes. The y-variables are classified as 1 or 0 (positive and negative for heart disease). Continuous values are encoded/scaled.

B. Potable Water Dataset

This dataset is also from the previous two assignments. There are 3276 instances with 10 features/attributes. Its "dimensions" have lots of continuous values. Its dependent variable also consists of 0 and 1 for its classification being positive and negative potability. Please note that this is imbalance data such that there is a lot more non-potable water than potable water.

Before going into clustering, I cleaned data of NaN values and changed the categorical data into binary forms. I also standardized/normalized data with continuous values to keep the data more consistent.

II. Clustering Algorithms

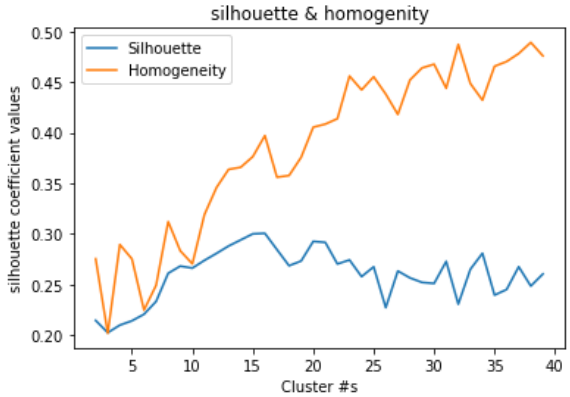
A. Heart Disease Dataset

Overall, there are not many instances for this dataset (303 instances). So I am expecting to see less accuracy with more clustering.

1. K-means Clustering

To get the brief picture of the distributions, I used the sum of squared errors to see how much the clusters needed for optimizing the errors. Graph below shows that from the clusters 16, the degree of its decrease gets smaller.

To see more of its detail, I also used the silhouette and homogeneity coefficient values. Here, you can notice that the cluster number of 16 gets its highest coefficient values of 0.3 with its high spike of homogeneity.

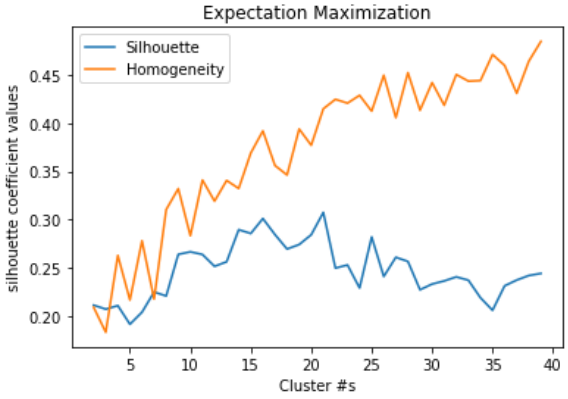


Overall the values of k=16 would seem to be a good cluster according to this graph. However, such a value did not yield good accuracy (which is lower than 0.2). Silhouette and homogeneity describes the shapes of its cluster distribution but since the data is quite small for heart disease, its distribution did not help much to get a better

accuracy score. Due to this reason, the clusters with k=2 give out the highest accuracy of 0.79.

2. Expectation Maximization

For the Expectation Maximization algorithm, it considers the variance of Gaussian Mixtures which would 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 expected, both trends of silhouette and homogeneity lines are going in a similar direction to that of k-means. Here the k value of 21 has the peak silhouette coefficients.

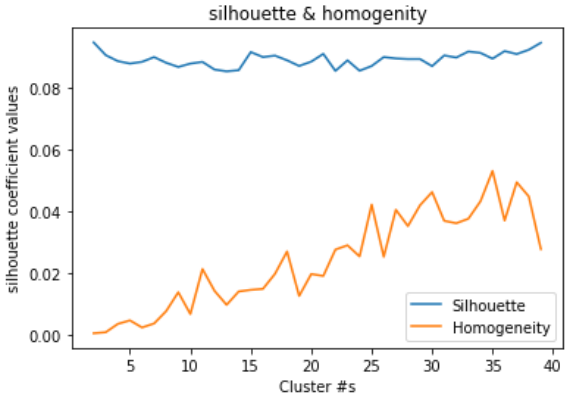
B. Potable Water Dataset

Potable water dataset contains comparatively larger data datasets such that running two different clustering algorithms would show different results. Since there are low correlations between the features, I expect to see the data are widely distributed.

1. K-means Clustering

The SSE graph shows that there is a sudden drop up to 20 clusters and its rate of decrease in the SSE sum is slowing down; however, the graph does not show the clear elbow point so I included the silhouette and homogeneity coefficients.

Overall, the silhouette and homogeneity coefficients are very low, less than 0.1. This means that the potable water dataset has the widely distributed data clusters and the observations of the same class label are not in the same clusters. This may be due to low correlation within the data features.

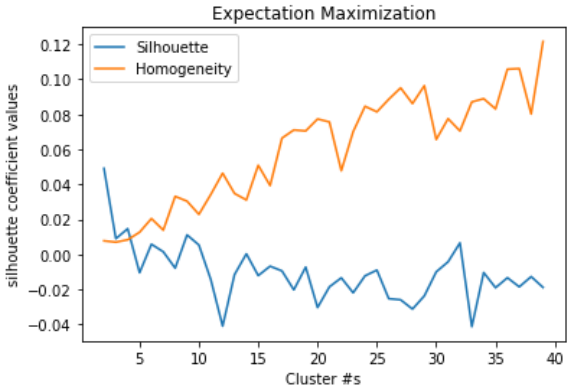


The silhouette coefficient is highest when the number of clusters are either k=3 or k=40. There is no need to go for the 40 different clusters for the data size of 3276 so the cluster of 3 would be the best option. The accuracy score is 0.51 which is not high but this may be due to the data's randomness or process of data balancing with its data rescaling.

2. Expectation Maximization

The EM yields quite similar values to the k-means such that the clusters of 3 would yield the best silhouette coefficients. Interestingly, with the gaussian variance, the homogeneity goes beyond the silhouette values; however,

since the overall coefficient values are still low, there would be any noticeable difference.



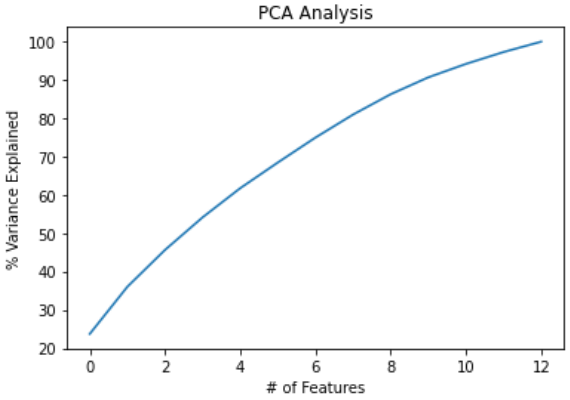
Since both datasets are quite small, clusterings for both datasets have overall low silhouette values which means that the shape of clusters are widely distributed and the same class labels may not be in the same cluster.

III. Dimensionality Reductions

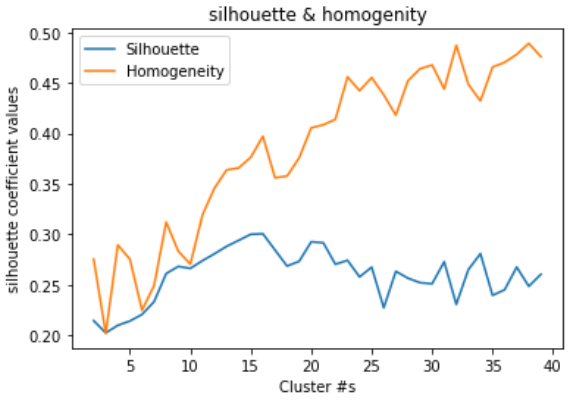
A. Heart Disease Dataset

The Heart Disease Dataset contains 14 features. It does not have that many dimensions but still I expect some reductions can improve its biases and overall score.

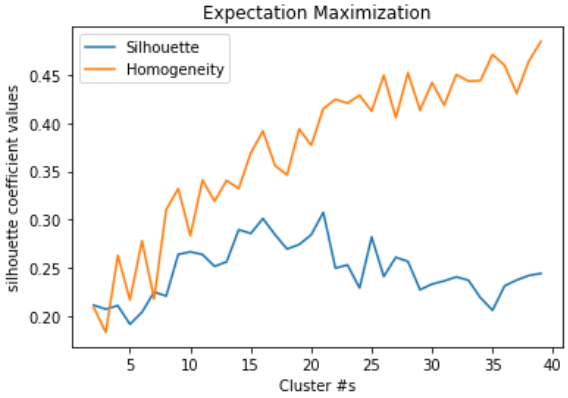
1. PCA



For PCA, I used the eigenvalue of variance % such that the graph above describes its variance %. Even though there are less attributes, the graph has a slight curve shape such that there is diminishing return on the component value of 12.



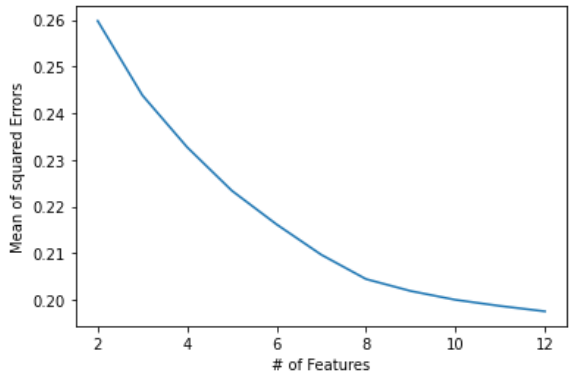
The implementation of the k-means clustering on the reduced heart disease dataset looks very much the same with the one without the reduction. Its best silhouette coefficient value is around 0.30 with the 16 clusters.



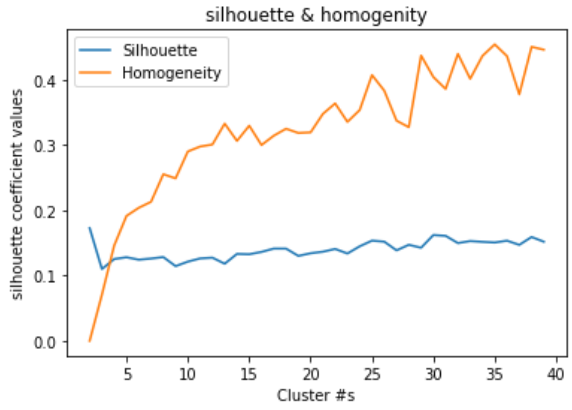
The Expectation Maximization algorithm also yields similar results for this dataset. Its best coefficient value is at 0.3 with 21 clusters.

Overall the PCA did not make much difference for its clustering for Heart Disease Dataset due to the eigenvalue variance being almost linear and the number of attributes and features are relatively small.

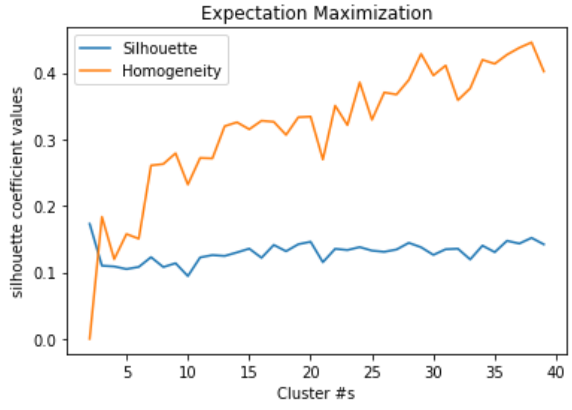
2. ICA



For ICA, I used mean squared errors to determine its number of features (reductions). The graph above shows that it has a nonlinear curve. From the 11 features it has comparatively lower slopes so I choose the ICA's n_component of 11 for this dataset.



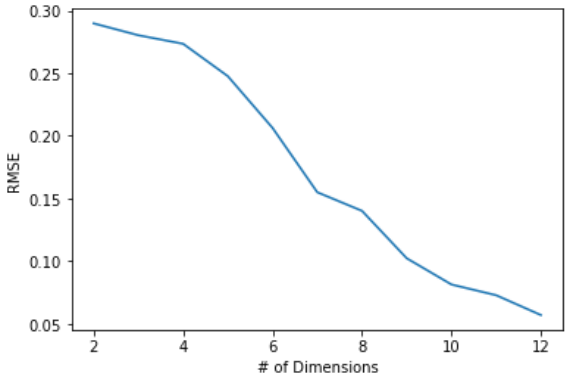
With the reduced dimensionality, it shows the comparatively lower silhouette scores than the original data which is lower than 0.2.



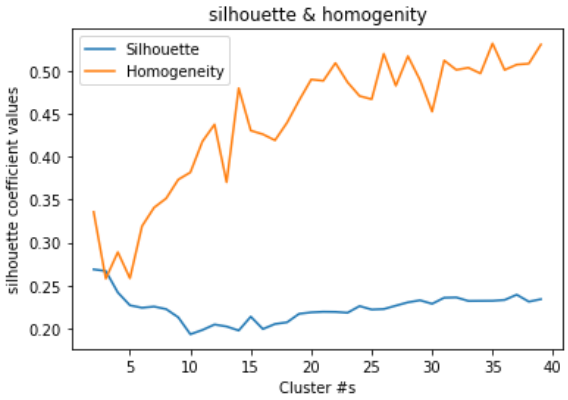
Similar to the k-means clustering, the EM algorithm also has the lower silhouette coefficient score of it being lower than 0.2.

Reducing the dimension for this dataset via ICA has not been helpful.

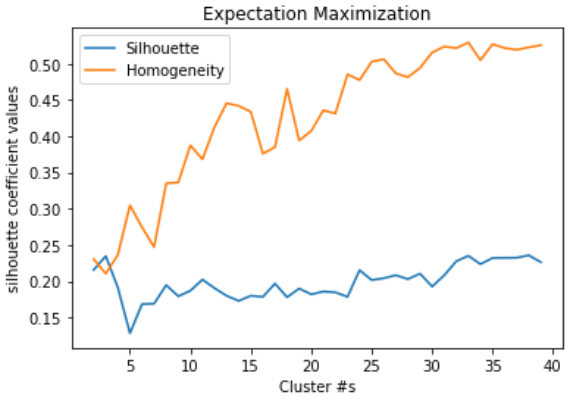
3. Randomized Project



For the Randomized Project, I used the reconstruction error. The graph above is quite linear. I chose the dimension of 7 since the slope of the line becomes less steep from that point.



The result came quite similar to that of the original such that the maximum silhouette is close to the 0.3 for k-means clustering.



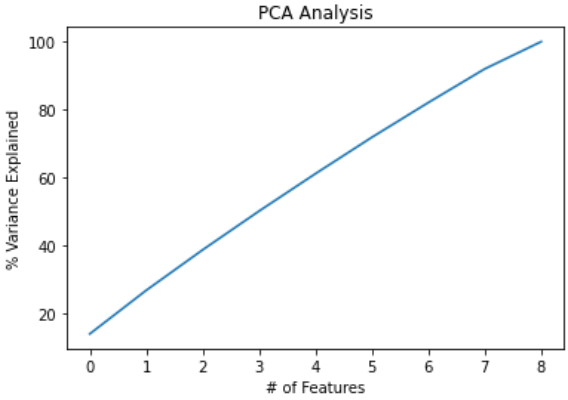
EM's coefficients are also closer to that of the original such that the maximum silhouette value is close to 0.3.

Both clustering algorithms indicate the clusters of 3 would give out the best cluster values for the heart disease dataset.

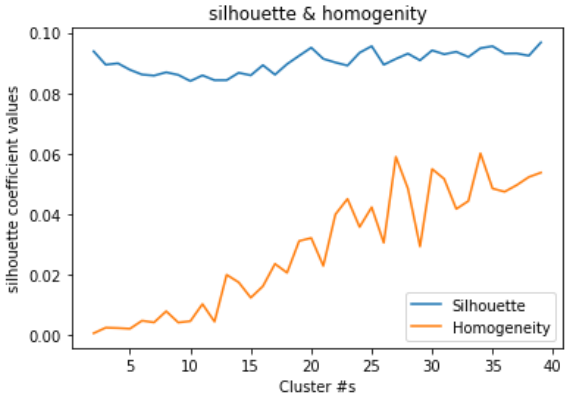
B. Potable Water Dataset

For the potable water dataset, there are only 10 dimensions/features. Since it is rebalanced and scaled, the reduction may be less impactful for this dataset as well.

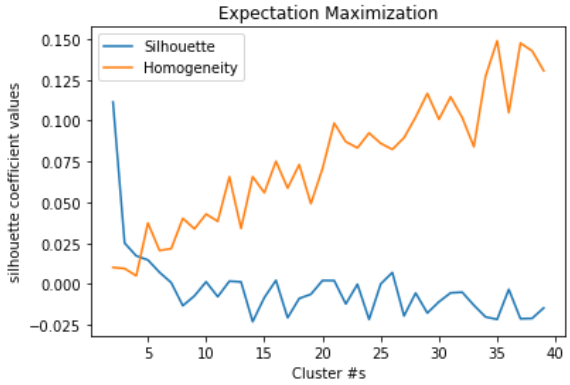
1. PCA



The eigenvalue for this dataset shows almost the linear relationship between the variance and the number of features such that it is not helpful to reduce its dimension; however, for the purpose of testing out, I picked the dimension of 7 since it has a slight curve in slope.

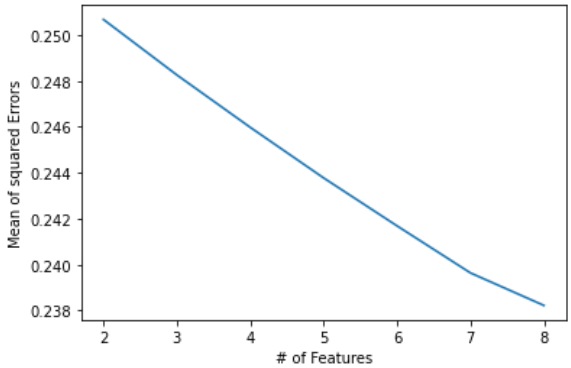


The implementation of clustering shows very similar results to that of original clustering. The silhouette coefficient is less than 0.1 and best cluster numbers would be 3.

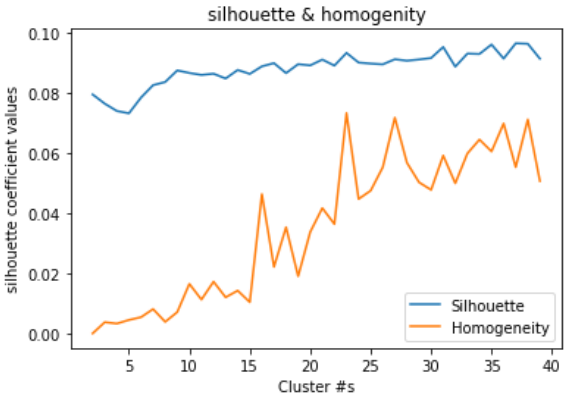


The EM value is slightly different but the trend of silhouette lines is very similar. It drastically drops up to the point where it has 7 clusters. But the cluster of 3 giving out the best silhouette values are still the same.

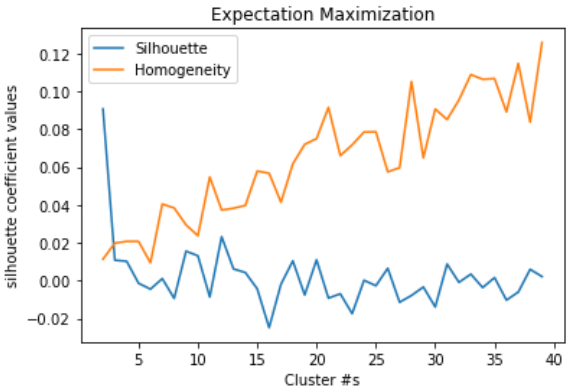
2. ICA



Similar to the PCA values, the ICA's Mean of squared error line is linear. While it may be better to keep all the dimensions, I saw that there is a slight curve of line at the 7 features and picked used that value to run the clustering.

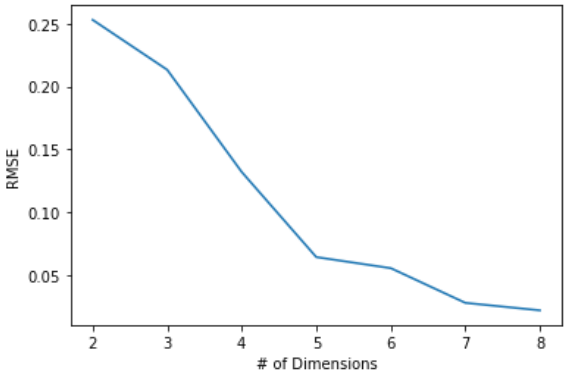


There is a slight difference in the silhouette value; however, its max silhouette value does not go beyond 0.1. The cluster of 3 would still be better since there is not that much difference between 0.08 and 0.095.

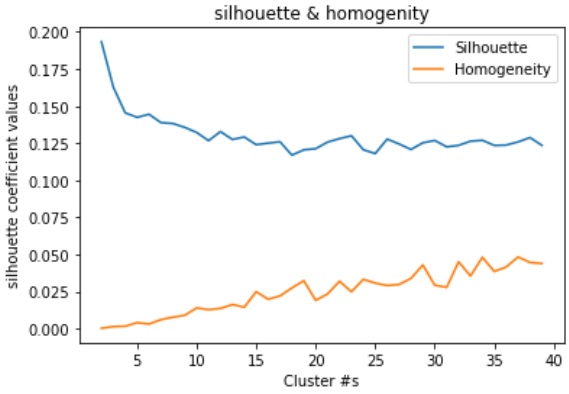


EM also has a similar trend such that it drops drastically to the point where the cluster number hits 7. The best number of clusters would be 3.

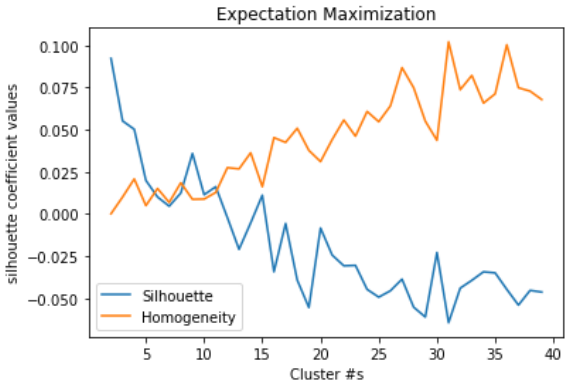
3. Randomized Project



I also used the reconstruction error for this dataset. The graph above is also linear. I chose the dimension of 7 for the slope of the line becoming less steep from that point.



There are some differences from the original silhouette graph. Overall the silhouette coefficient has increased such that its coefficient almost hits 0.2. This means that reducing its dimensionality according to its RP has put the data distributions in comparatively "closer" clusters. Nonetheless, the number of clusters of 3 giving the highest coefficient still stays the same.



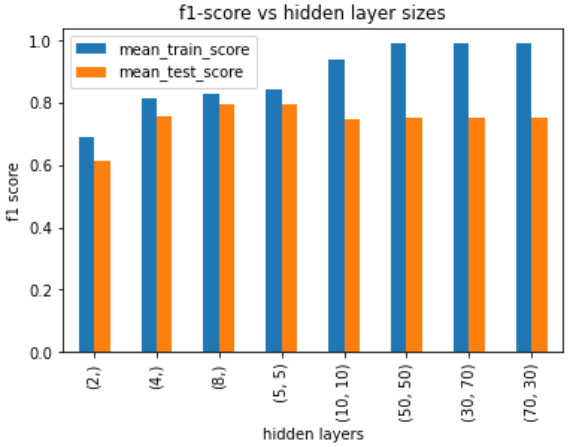
Similarly for the EM, the coefficient is slightly improved; however the silhouette line trends are still the same with the original one.

Overall, dimensionality reduction has very less impact on both datasets. This is due to its limited number of attributes and features. The data observation of same class labels may be in different clusters. The correlations between each feature may be extremely low and randomized.

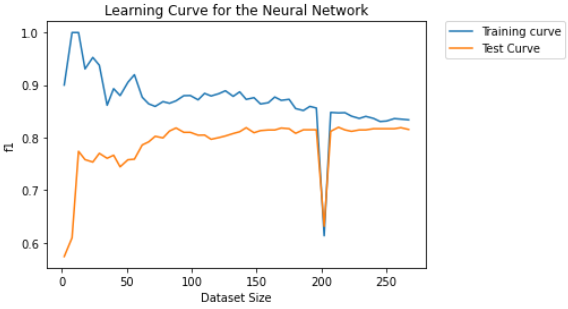
IV. Neural Networks with Dimensionality Reduction (Heart Disease)

For the Neural Network, I used the assignment 1's Heart Disease data. Even though it has the least amount of data, 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 as well as its f1-score.

A. PCA

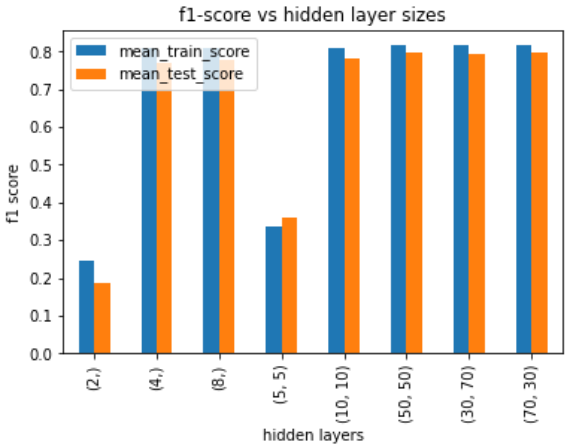


With the new dimension of the PCA, I got the different hidden layers parameters. The optimal hidden layer has become the one hidden layer with 8 nodes.

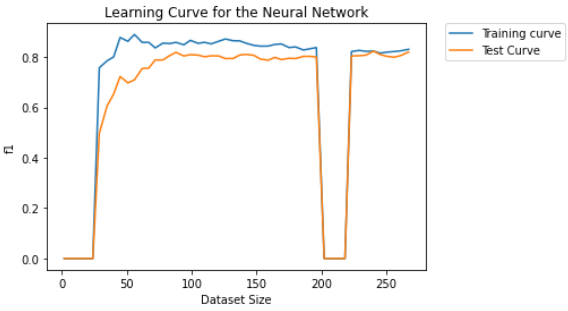


With the new parameter, the f1-score has improved very slightly such that the overall test curve has shifted upwards.

B. ICA

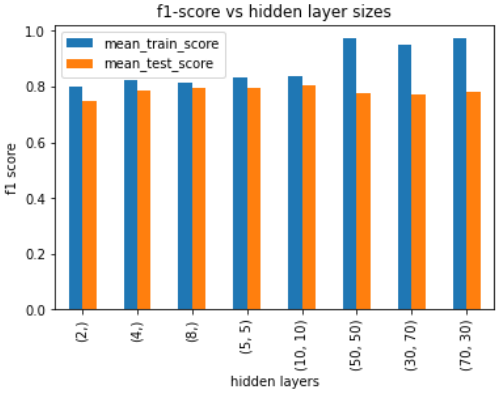


ICA also gave out the different distributions of neural network parameters. According to its gridsearch, two hidden layers with 50 nodes gives out the best optimal f1-score.

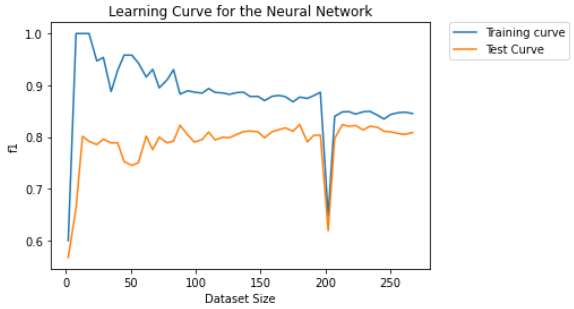


The gaps between the training curve and test curve have been drastically reduced. This could mean that the dimensionality reduction via ICA has increased its variance and lowered its bias of the dataset.

C. Randomized Projections



For the Randomized Projection, the two hidden layers with 10 nodes give out the optimal f1-score.



The f1-score after the reduction has improved such that its best score is 0.81 (rounded).

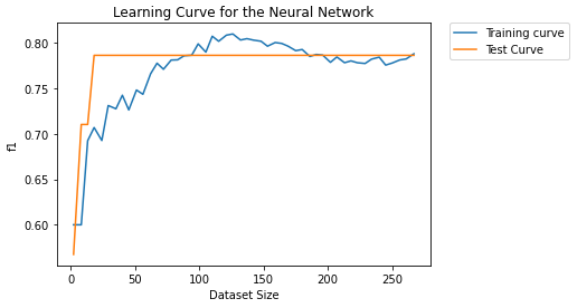
Overall, the dimensionality reduction has improved its f1-score as well as the gap between training curve and test curve.

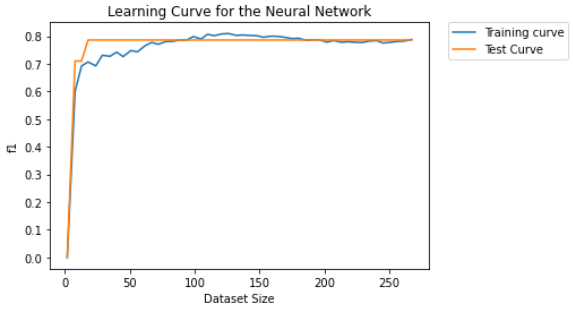
V. Neural Networks with Clusterings (Heart Disease)

For a Neural Network with clusterings, I also used the heat disease dataset for similar reasons. I expect to see the improvements on the balancing of bias and variance trade offs with the clusterings.

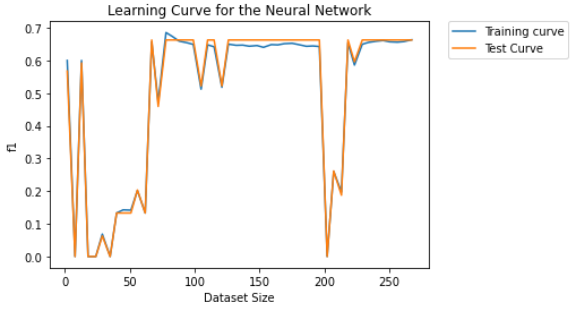
A. K-Means

For K-means, I tried different sets of hidden layers. The first graph is the one with two hidden layers with 60 nodes and the second graph is the one with 100 nodes. The latter graph shows that there has been some reduction of bias.





B. Expectation Maximization



With the EM clustering and the neural hidden layers of (50,50), the testing and training curve align almost identically.

VI. Conclusion and Clock Time

After reducing the dimensions, I noticed there has been a noticeable difference (improvement) in running the Neural Network. This is because lower dimensionality causes the lower space to run.

Overall, there wasn't a groundbreaking difference after reducing its dimensions or applying different algorithms. That is because the two datasets have comparatively smaller attributes and features.