Assignment 3: Unsupervised Learning CS 7641

Connor Beveridge (#902451314)

1. Introduction – Clustering and Dimensionality Reduction (DR)

In this first analysis, I compared 2 datasets representative of different conceptual difficulties to humans. In this analysis, I will compare the same 2 datasets that fairly represent the opposite ends of this spectrum of human understanding however I will be analyzing how these datasets are handled by unsupervised method. I believe the data types (discrete vs continuous) will have a large effect.

2. Data Introduction: Two Spectrum-Spanning Datasets

a. Tic-Tac-Toe Endgame (TTT) (find in "tic-tac-toe.csv")

The tic-tac-toe dataset is categorical and has an obvious true model of a disjunction of conjunctions (3 x's could be here or there or there etc.). The target values are + or – indicating victory for x. Each feature is the state of one of the 9 squares with values of x, o, or b for blank. It will be interesting to know if clustering will create useful groups for this dataset, given that the positive and negative targets are very mixed in Euclidean space.

b. Heart Disease (HD) (find in "heart.csv")

The heart-disease dataset is largely composed of continuous data from physiological measurements but also has a couple nominal and ordinal features. Each instance is a person, their characteristics such as age and physiological measurements such as cholesterol level. The instances have classification of either having the presence of heart disease or not. The target concept here is a combination of many probabilistic factors and is not intuitively understandable to a person like tic-tac-toe is.

c. An interesting Comparison

Both datasets have a similar number of features and training instances. Both sets are balanced. It will be interesting to see if clustering is able to group the complex heart disease task compared to the simple logic rules of the tic-tac-toe dataset. It will also be interesting to see how the groups and components created by these unsupervised methods will map to some cognitively useful structure.

d. Preprocessing

The tic-tac-toe data was encoded numerically for use with one-hot encoding which was done with the *pandas* package. This encoding was selected to prevent any-kind of ordinal attribution to the data. The heart dataset was already all numeric, however one feature had multiple options which were nominal in quality and so one hot encoded.

3. K-means Clustering

KMeans from sklearn.cluster was used. Similarity for both datasets are simple Euclidean distances between instances. This will be easy to visualize and makes sense when none or not much domain knowledge can be brought in on similarity. The silhouette score was used to determine the best k. I believe this scoring method is best because It looks not only at distances between clusters, but the density of the clusters, thus providing information on which number of clusters provides individualistic, compact groupings. This will be good for these datasets as it will value cluster quantities that do not overlap much, have instances match the clusters well, and have instances that poorly match other clusters. I believe this will provide a better chance of their being some useful meaning to the clusters..

Hyperparameter tuning for parameters other than k and training size for the K-means algorithm was initiated, however fiddling with hyperparameters by hand to start revealed no significant change in silhouette scores with a wide variety of hyperparameters tried. I initially scaled the data but it ended up making the silhouette score worse, thus indicating the features with large numeric values which put more weight in the distance measure are more important overall in determining if a person has heart disease.

a. Tic-tac-toe (TTT)

A range of potential k-values was tested and k=14 was found to be the best, but not by much. All silhouette scores for k's between 2 and 20 ranged between .07 and .106. These are very poor scores. Pondering on this, it does make sense as there is a moderately even distribution of instances among the endgame side of the feature space. Most states are grouped away from the 'blank' side of the total feature space because no end-game states exist were many squares are blank. Pondering on these results further it also seems k-means alone would be very bad at creating clusters that represent the correct target classifications as many board states which are very similar, only differing in a couple or few Os and Xs, can be different classifications, however it does seem that some information can be gleaned as certain board states and their neighbors would be more likely to lead to a win for x. This distribution is hard to visualize in 2D and 3D space as many of the points have the same coordinates for many feature combinations. As the silhouette score was so low, I tried to find better groupings that could have more meaning by using a smaller number of instances, thus preventing overfitting. This is shown in Figure 1, also for heart disease. A grid search was done to find the best number of instances and the best k to use for that instance count. K=6 with only 25 instances was found to be the best with a .166 silhouette score, which is still low. One improvement that may help find better clusters is to use the total number of different squares as the distance instead of the multidimensional Euclidian distance as it will provide an overall greater distance between different instances. For instance, 2 states can be at most 9 squares different. With this improvement the distance would be 9, however with Euclidian this is only the square root of 9.

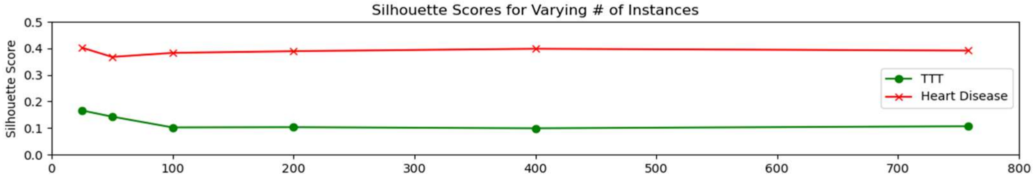

Figure 1 | Best Silhouette Scores for varying number of instances used

b. Heart Disease (HD)

A range of potential k-values was tested and k=2 was found to be the best by a relatively large amount. This is what I was hoping to see as there are 2 classifications and so these two clusters may have correctly labeled, to a degree, the instances in an unsupervised way. The silhouette score for k=2 was .391, indicating light separation of the clusters. For higher k's up to 20, the score is always around .27. As seen in figure 1, using less instances didn't help the silhouette score. Testing these clusters as labels reveals an accuracy of 58%. This implies something useful is happening but the persons' with and without heart disease do not separate in Euclidean space based on these features. To get a visualization on the clusters, they are compared to the ground truth clusters as shown in figure 1. There are many dimensional combinations that could be visualized, however the features with large numeric values as mentioned earlier were chosen, however the cluster labels are based on all features, not just these 3. These features are also things I would personally think would be indicators of heart disease based on my limited medical knowledge. As you can see, the ground truth data is rather mixed up regarding the labels, however some slight separation is visible. Clustering showed some similarity regardless of this mixing. When considering these clusters as the labels themselves, these clusters seem to group lower cholesterol, higher blood sugar individuals as those with heart disease which is somewhat seen in the ground truth. Improvements to the similarity/distance measure for this dataset should produce better clusters. This could be done by bringing in domain knowledge from a heart doctor.

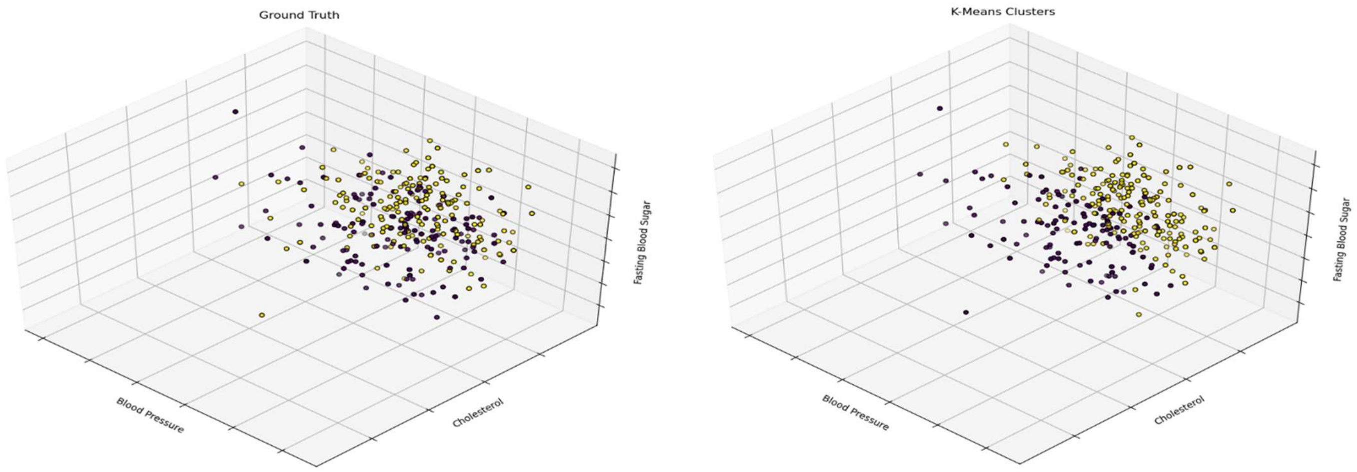

Figure 2 – k-means clustering viewed on 3 dimensions. Yellow indicates heart disease, purple without

4. Expectation Maximization (EM)

Sklearn's Gaussian Mixture Model, GuassianMixture from sklearn.mixture, was used to implement EM. For the same reason, distance in Euclidian space was used. I predict EM will do better than k-means on the heart disease data because it will not be biased towards certain cluster structures for certain covariances. Again, for the same reasons, silhouette score is used to utilize the density and separation of clusters in picking the best number of components. A grid search was done to find the best number of components and covariance combination. Covariance type determines the best boundary shape and expanding this shape from the circular shape of k-means might provide more accurate clustering with respect to some underlying meaning.

a. Tic-tac-toe

A range of potential k-values was tested and k=15 was found to be the best, but not by much. All silhouette scores for k's between 2 and 15 ranged between .07 and .108. These again are very poor scores, indicating the instances are fairly evenly distributed in the space. Similar reasonings for this are given in the previous section. The same improvement as mentioned in k-means would also apply to EM.

b. Heart Disease

A range of potential k-values was tested and k=2 was found to be the best by a relatively large amount. Similar to k-means, the silhouette score for k=2 with spherical covariance was .38, indicating mild separation of the clusters. Surprisingly, the score for the 'full' covariance which allows for a variety of cluster structures only scored .07. Using less instances did not help the silhouette score.

Testing these clusters as labels with these parameters (also spherical boundaries) reveals an accuracy of 60% compared to the true labels, which is only slightly better than k-means. I attribute this slight improvement to the fact that while k-means retains instances in one cluster or the other, EM can keep track of probabilities of instances belonging to certain clusters, and thus model overlapping clusters better.

Out of curiosity, I tested the accuracy of the non-spherically restricted 'full' covariance with respect to the true labels, and discovered an accuracy of 67%, even though it had the worst silhouette score. This seems to indicate that for moderately mixed labels, silhouette score is not appropriate because it aims for highly separate, condensed clusters instead of best fitting the overlapping nature of the clusters with ovoid shapes that can varying dramatically in diameter among dimensions. It does indeed provide a score near 0 for overlapping clusters, so maximizing it will not provide appropriate clustering when the ground truth clusters are indeed overlapping.

To get a visualization on the EM version of the clusters with ‘full’ covariance, they are compared in **figure 3** to the ground truth clusters shown in **figure 2**. You can see how much more ovoid shaped clusters can better represent the intricacies of oddly shaped, overlapping clusters of the true distribution.

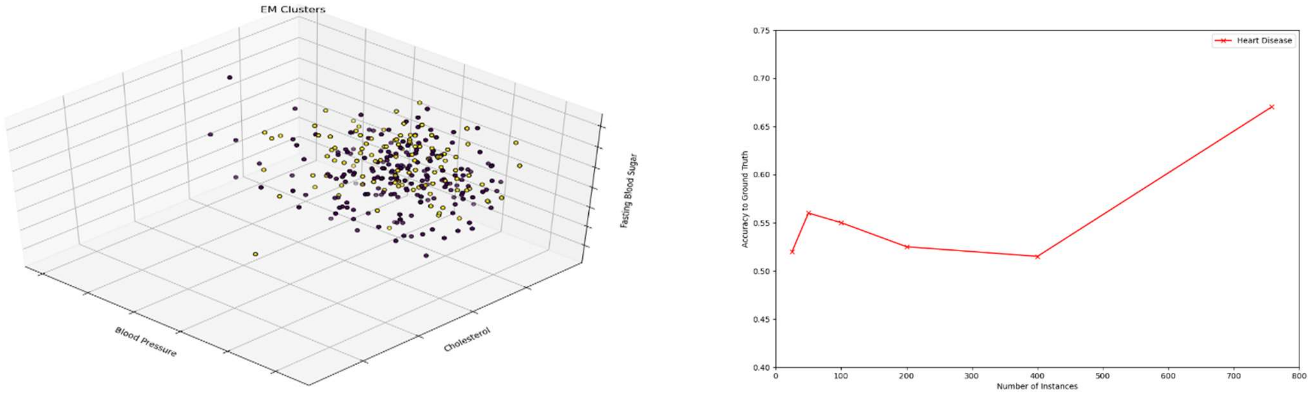

Figure 3 | EM clustering and Accuracy to Ground Truth with number of instances used

Reasoning for the features selected here and the improvements that could be made are the same as previous with k-means, with the addition that a better cluster number scoring method should be selected for such a cluster overlapping dataset.

5. Principle Component Analysis (PCA)

Dimensionality reduction techniques in the following 4 sections reduce the dimensionality of the datasets to speed up and improve fitting with other algorithms. PCA is first used to reduce the dimensionality to 3, finding the best fitting principle components of the original number of dimensions by capturing correlations which preserve variances of the original dimensions. Sklearn’s PCA from its decomposition package is used. Percent of total eigenvalues, eigenvalue or explained variance ratio are reported instead of eigenvalues for better conceptualization. Instances are scaled around the origin with variance 1 in order to capture correlations among dimensions.

a. Tic-tac-toe

PCA reduction is done on the TTT data and the results can be seen in the first graph of **Figure 4** where the ground truth labels are colored differently. The top 3 components based on eigenvalues are chosen. At first glance, It appears that PCA did not leave the ground truth labels very separated. However, upon closer inspection, it does seem the purple labels (victory for O), are aligned somewhat down the middle, so may be more separate in a higher dimensional space. Because there are so many initial dimensions due to one hot-encoding (27), boiling them down to 3 dimensions may loose to much information. Derived from eigenvalues, explained variance can tell you how much information was retained in the new dimensions as a percent of the total variance. For this TTT data, 8% of information was retained for each of these top 3 components (where each dimension was responsible for ~4% before), and the rest drop slowly down to 1% at the 16th component. 92% of the information (via variance) can be retained with 14/27 components.

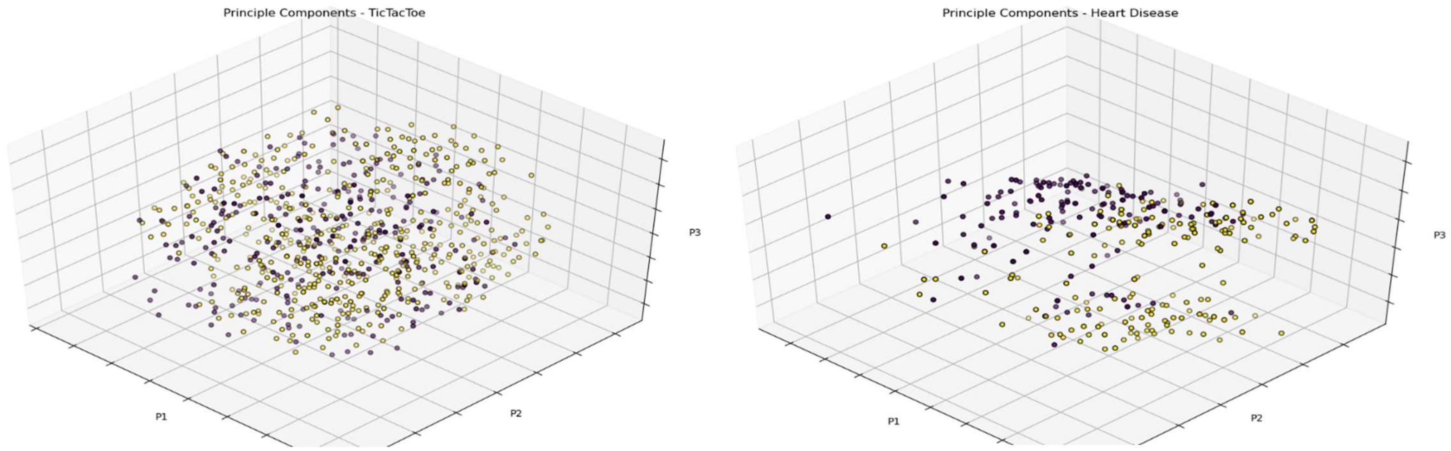

Figure 4 | Principle Components Transformed Instances and their ground truth labels (by color)

b. Heart Disease

PCA reduction is done on the heart data and the results for the top 3 components can be seen in the second graph of **Figure 4** where the ground truth labels are colored differently. Using eigenvalues, it is found that the first principle component (P1) held 20%of the variance/information, the second (P2), held 10%, and the third (P3) holds 9%. While the instances in just 3 new dimensions only retained ~40% of the information, the results are impressive. The instances are much more “clusterable” and are separated greatly by label as compared to the ground truth 3D clustering in **figure 2**. One drawback is that it takes 11/16 features to retain 91% of the information. However, it appears based on how well these 3D clusters have been made, that it could be beneficial to lose certain information as it could be noise or outliers.

6. Independent Component Analysis (ICA)

ICA will try and discover the values for some probabilistically independent variables when these variables are responsible independently or together for the values of the observable features of each dataset. It is very hard to imagine what any number of hidden variables could be for the TTT dataset as the states of each square are more influenced by the states of other squares not some external variable. Maybe the independent components could important original features not transformed. For the heart data I can however easily imagine what the hidden components may be and this type of analysis seems to fit well the true target of this dataset when predicting labels which is a combination of certain factors which probabilistically result in these feature values. Because the features (observations) consist of physiological measurements and things like age, variables that directly result in these observables could be things like exercise amount, diet quality in numeric form, age, sex, race, stress levels, mental health level, etc. As mentioned, in the real world these variables probabilistic result in certain observations which are the features here. For instance, low exercise would increase the probability of higher resting blood pressure. These observables thus ultimately determine the presence of heart disease as well, but these labels are disregarded in the ICA. Features are again centered around the origin. Kurtosis is used to find the best number of hidden variables by maximizing independence and reducing gaussianity. The number of features with the highest mean kurtosis for all features is determined to be optimal.

a. Tic-tac-toe

ICA analysis with 3 components Is shown in the first imagine in **figure 5** with ground truth labels colored. I am not surprised that these components do not differentiate the labels well at all because it seems as this problem just does not have independent external variables influencing the specific, observation of an X, O, or B at each square. Increasing or decreasing the number of hidden variables did not significantly improve kurtosis, however, using 3 components gave the best kurtosis (-.96), but barely.

b. Heart Disease

This dataset as I thought does seem to have some real independent components. The second image in **figure 5** shows the data mapped to the computed values of these new dimensions with ground truth labels shown with different colors in 3 new components for best visualization. Even though this distribution had a negative mean kurtosis, it can be seen in the figure that it separates the labels moderately well. By decomposing the features into the variables which cause them, it appears to have grouped the labels closer together as these unknown variables likely are responsible in some way for heart disease. You can see that as IC1 increases, instances are much more likely to be negative (no HD). Thus, this component could for instance represent amount of exercise per day. Testing for several components with a better mean kurtosis reveals 8 components yields a mean kurtosis of .56.

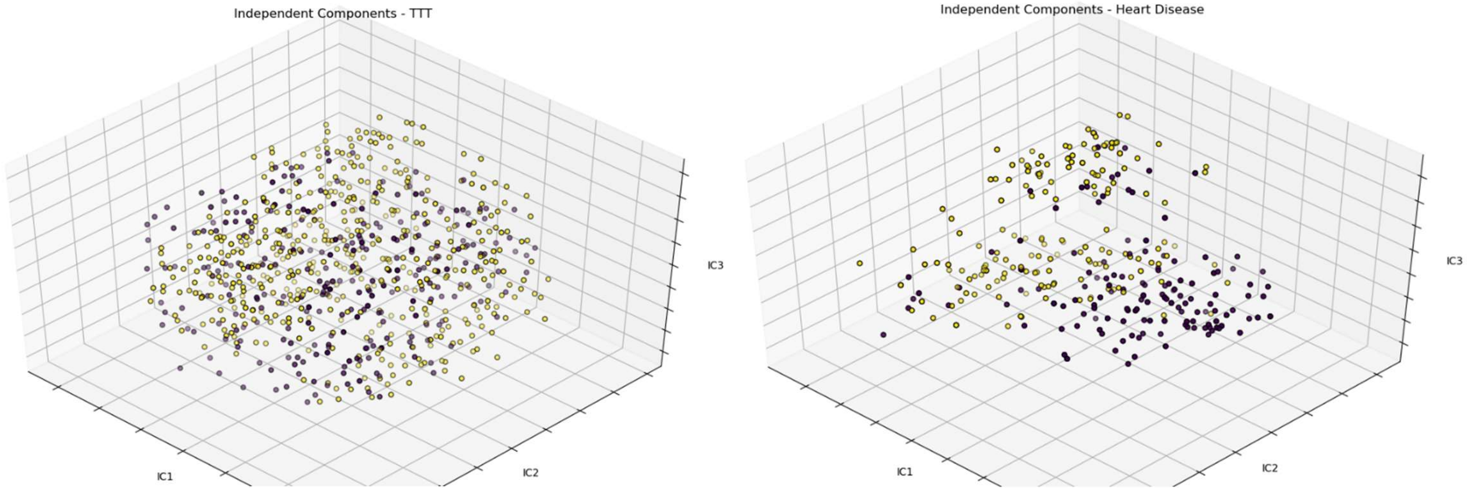

Figure 5 | 3-component ICA with colors representing labels (yellow=positive)

7. Randomized Projections (RP)

Reduction with RP may sacrifice accuracy, but it is much more computationally efficient. Reconstruction error in the form of mean squared error (MSE) is used to determine the best number of components. From 2 to the original number of components was tested to find the best MSE.

a. Tic Tac Toe

On a first attempt with 2 to 27 components, MSE Decreased in a linear fashion from **.86 to .~0** when the same number of dimensions are used and so all information is conserved. Running the projection algorithm 10 times with 2 to 27 components. revealed that MSE only varied slightly with a standard deviation range (STD) from **.01 to .02** and a similar mean of **.92 to .008**. Due to the linear decrease in error towards the original number of dimensions, a threshold in error must be so dimensionality reduction can be done to some degree. For this analysis, 20% of the original variance of 1 will be used for MSE threshold, which is .2. This resulted in a reduction to 21 dimensions from an original 27.

b. Heart Disease

On a first attempt with 2 to 16 components, MSE Decreased linearly from .86 to to .~0, indicating that Running the projection algorithm several times revealed that MSE only varied slightly with a **.02-.04** STD range for each and a linearly decreasing mean range from **.875 to .02**. Again a mse of 20% of original variance threshold is used to pick a reduction to 13 dimensions of the original 16.

8. Feature Agglomeration (FA)

The unsupervised method FA is like clustering, but instead of clustering the similar samples together it clusters the similar features together. Thus, the resulting dimensions can capture meaning related to how certain features related to each other. For instance, the heart disease features cholesterol, blood sugar, and blood pressure could be combined into a single dimension that captures some sort of physiological indication. FeatureAgglomeration from sklearn.cluster is used for implementation. It will merge clusters so that in the new cluster. To try and produce a feature space in which the instances are more clustered together, silhouette score is used to determine the optimal number of clusters to end up with. Hyperparameter search is done on distance measure (variance or mean distance) with 3 components (for visualization) generated and with a range of components. Kurtosis is again used as a goodness metric in the hyperparameter search to again minimize conditional dependence to capture when combing more feature clusters would do no good due to lack of dependence between them.

a. Tic Tac Toe

With 3 components, **variance** of instances was found to be the best distance measure between features with a mean kurtosis of -.07. The projection onto this new space can be seen in **figure 6** with respect to the 2 labels identified by color. The results have a hypnotizing structure but may lack meaning, instead being due to how the one-hot-encoded features are scaled and then agglomerated. The gridsearch with other numbers of components revealed the 2 component version is best by kurtosis of .07.

b. Heart Disease

With 3 components, **average distance** of new feature clusters was found to be the best distance measure between features with a mean kurtosis of 2.27. The projection onto this new space can be seen in the second figure of **figure 6** and does separate the labels fairly well. The grid search with other numbers of components revealed this 3-component version is best by kurtosis.

9. Clustering Reduced Datasets

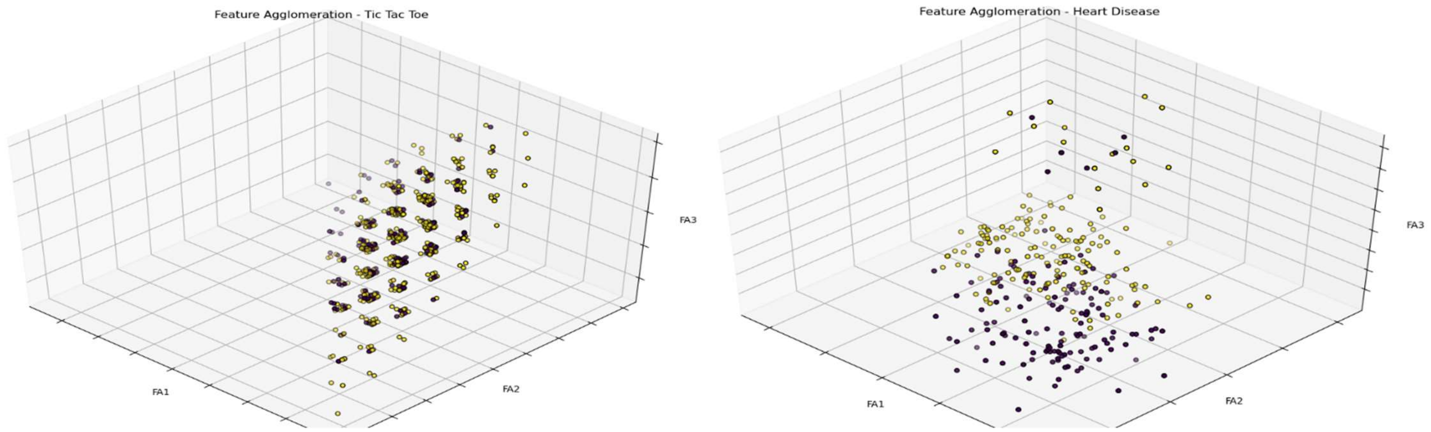

Figure 6 | Data sets reduced to 3D with Feature Agglomeration. The Binary Labels are colored

Given the optimal number of dimensions (based on kurtosis, eigenvalues, and reconstruction error) reported in the previous section for each DR algorithm, both datasets are reduced, and the clustering algorithms are applied to them. Silhouette score is used as the clustering performance method. The 'full' covariance is used for EM to more dynamically structure clusters. The following charts details the clustering results from the 4 reduced datasets with respect to their optimal parameters previously outlined. Best number of components with corresponding silhouette score are listed as well as ground truth accuracy when only 2 clusters are formed.

Tic Tac Toe	Best # Clusters K- Means	Best # Components EM	Ground Truth Accuracy k-Means	Ground Truth Accuracy EM
PCA	3	3	59%	54%
ICA	6	6	59%	61%
RP	2	2	60%	61%
FA	10	9	57%	57%

Table 1 – Tic Tac Toe Clustering with Reduced Data

Heart Disease	Best # Clusters K- Means	Best # Components EM	Ground Truth Accuracy k-Means	Ground Truth Accuracy EM
PCA	4	5	80%	52%
ICA	4	5	63%	59%
RP	3	2	66%	66%
FA	2	2	52%	52%

Table 2 – Heart Disease Clustering with Reduced Data

a. Interesting Results

The most interesting insight to me was the high accuracy to ground truth for PCA with HD data when producing only 2 clusters (80%) You can see in the first graph in **figure 8** how this compares to the ground truth shown in **figure 4**. This is interesting as it is all unsupervised. PCA really captures the essence of the data via correlations, and additionally by reducing it down likely even filters out noise, all leaving the instances less mixed in the projected space for better clustering. PCA also spread the points out more overall in separate directions, creating more than 2 obvious clusters as like before PCA. This is why the optimal number of components was 4 and 5 for k-means and EM respectively instead of 2 as originally found pre-DR. EM's 5-clustering can be seen in the second graph of figure 8. While EM did a poor job with respect to ground truth accuracy finding 2 clusters, it appears to have found 5 useful clusters as seen in the graph which appear to group ground truth labels well (**see figure 4**). They are useful because they can be used to split apart the data better and pickier (with respect to labels), resulting in better accuracy than the 2-cluster grouping with k-means alone via the addition of supervised learning as done in the final section. Although 2-cluster k-means did surprisingly well on this PCA HD data, its spherical boundaries are limited in fitting the non-spherical boundaries of the ground truth distributions (**see figure 4)**. The 2-cluster accuracy results were very poor for k-means and EM in FA even though labels were well separated due to the outliers which were made into their own cluster. While it is easy to see visually the separation of the data with respect to ground truth (**see 2nd graph, figure 6**), this separation is not in the form of clusters in the EM and -means sense. The higher number of clusters chosen for TTT with FA may provide some information on combination of spaces occupied given the structure shown in **figure 6**. The addition of these features to the original ones in supervised learning could provide useful information on relations between squares of the board. This EM 9-cluster result can be seen in the first image of figure 9, revealing interesting groupings.

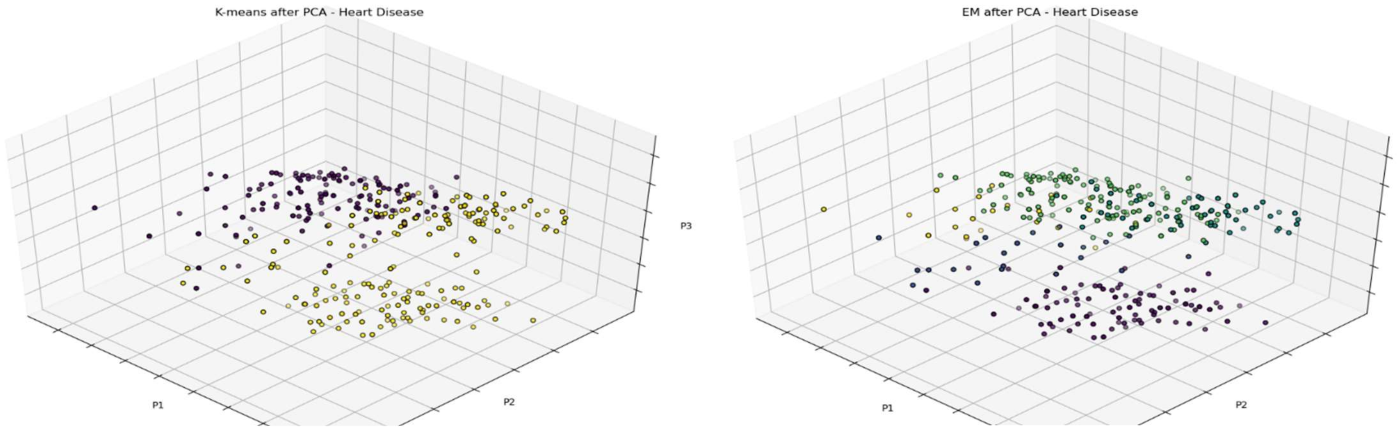

Figure 8 | Heart Disease Clustering with K-means (2-cluster) and EM (5-cluster) after PCA.

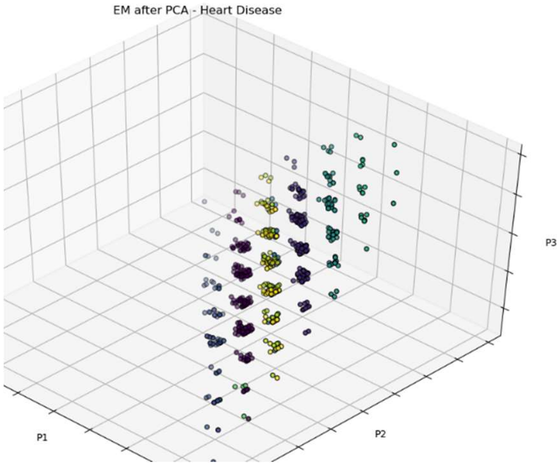

Figure 9 | 'Full' EM after PCA on Tic Tac Toe data. The 9 Clusters are shown by color.

10.Neural Network (NN) DR Application

The dimensionally reduced data from the original heart disease dataset is now applied to supervised learning with the same neural network as I used in assignment 1 with sklearn's MLPClassifier. Before DR, even with 758 instances and optimal hyperparameters and cross validated model, this classifier only achieved **82.5%** accuracy on the HD labels, and took **1.53 seconds** to train and score. For fairness in comparison to the original data, a single hidden layer will be used for all networks. A hyperparameter search was done for all 4 reduced datasets as explored in the previous sections. These optimal hyperparameters were then used to train on each dataset and the cross validation accuracies with respect to the number of iterations can be seen for all 5 datasets (4DR + 1 Original) In **figure 10.** Interestingly, even though the number of dimensions was reduced by a third with PCA, the accuracy on the test data improved by a couple percent to **84%** with a

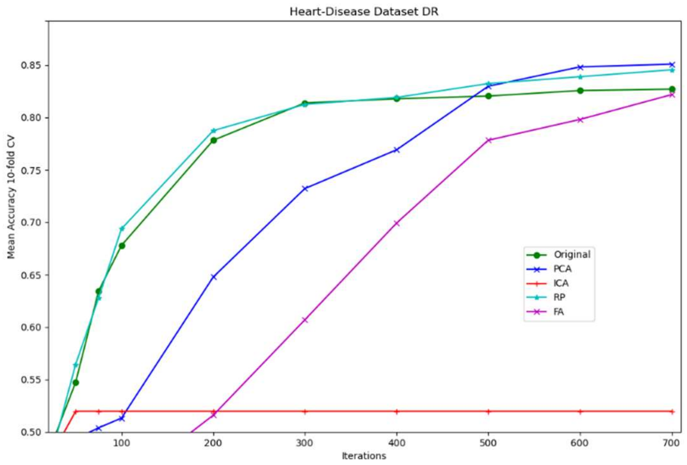

Figure 10 | NN Accuracy vs. Iterations for the 5 datasets

Fit + score time of **2.11s**. By capturing correlations in the data, it likely reduced some of the noise. It took longer to train well than before PCA DR and could afford to be reduced more without losing or maybe even gaining more accuracy. Surprisingly, the ICA test accuracy was only **50%**, equal to chance. Given the clusters seen visually it would seem it would do well, so this seems to indicate there are not good independent components or they are not identified. RP did well at **80% at 1.53s**, however it was only reduced by a few dimensions, thus it could also afford to reduce more at retain accuracy. Most interesting was the fact that FA had a test accuracy of **78%** even though it was reduced from the original 16 to only 3 dimensions! However, it had a training +scoring time of **2.42s**. Overall, here I am really had only 3 dimensions.

It took more time because more convergence time was needed, but with 750 iterations as shown in **figure 10,** times were about exactly the same.

11. Neural Network (NN) DR + Clustering Application

Given the optimal clustering's found on the DR'ed HD data shown in **Table 2**, these clusterings will be applied to neural network training as new features. Because there are a small number of clusters for all 4 DR sets, these new features are added to the original, pre-DR, HD training and test data for each instance. However, PCA seemed to have the best label separation so it will be used to provide all the features and the original will be discarded. The same hyperparameters are used except for hidden unit number which is reswept. The iteration vs 10-fold cross validation accuracy results for k-means can be seen in the first graph of **figure 11** and for EM the second. Note again that the blue line, PCA, was trained with only the 4 and 5 features from the clusters made, and the results are still impressive. I was disappointed that even with the original features, addition of the new DR + Clustering features did not improve accuracy more than it did. Although, it still revealed itself to be a useful means of increasing accuracy and processing time, especially when you have large dimensional datasets. Performance times were tested. Again, original training time was **1.5s.** The highly reduced with clustering PCA data was much faster at **.68s** for EM. FA slightly improved times at **1.25s** for both.

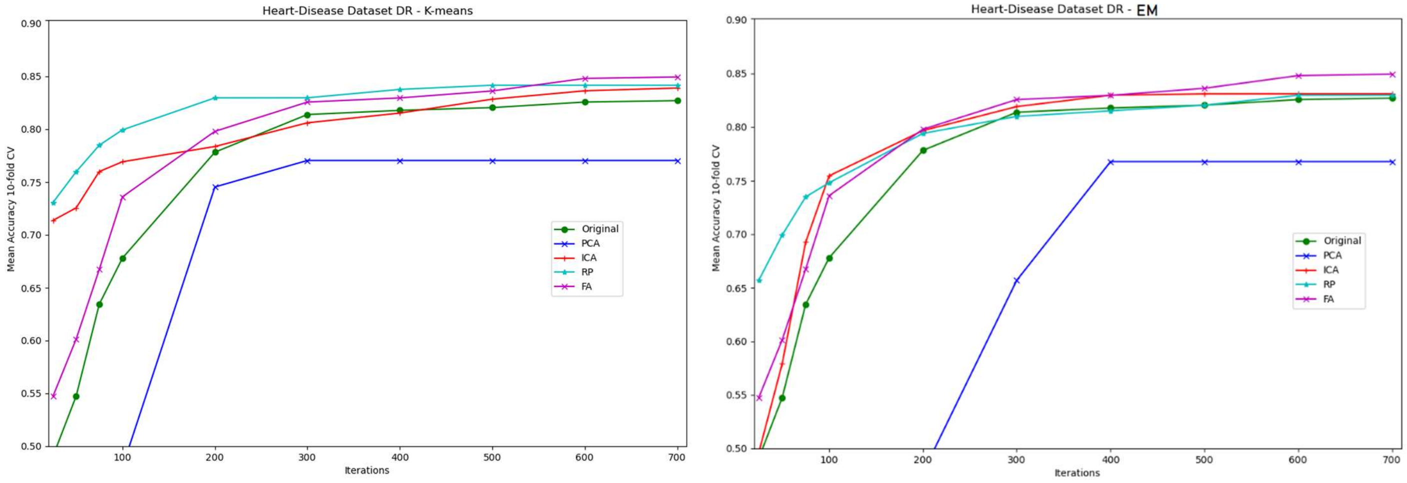

Figure 11 | NN Accuracies with Clustered DR features added. PCA is only Cluster Features.