Assignment 3:
Unsupervised Learning and Dimensionality Reduction

Vipul Koti
vkoti7@gatech.edu

1 INTRODUCTION

This report summarizes the analysis done on Unsupervised learning and Dimensionality Reduction on the two datasets from Assignment 1. The report is divided in three sections, In section 1, We talk about clustering algorithms on both the dataset, In section 2, for both datasets, we talk about Dimensionality reduction(DR) and clustering on the reduced datasets and, in the last section, for one of the dataset, we discuss the Neural Network (NN) performance comparison between NN performance on original dataset, reduced dataset and Original Dataset with an additional clustered feature. I have divided the datasets in train and test partition of 70-30. For section 1 and 2 I use train partition only. For all the algorithms in the experiments, I have used Sklearn library in python. The Dataset Recap is as below.

Wine Quality Dataset - This dataset is related to red variants of the Portuguese "Vinho Verde" wine, This dataset is available on Kaggle UCI Machine Learning Repository. There are 1143 rows, 11 continuous features and the outcome is multiclass. Alcohol content, sulfate, citric acid are positively correlated to wine quality. The initial class distribution is as follows, 5 (42.5%) , 6(40.2%), 7(12.0%), 4(3.2%), 8(1.5%) and 3(0.6%). I used Assignment 1 data preprocessing, removed Duplicates and reduced the dataset to 979 rows and output to three classes, 5(44.2%), 6(41.8%) and 7(14.0%), Class 5 and 6 dominates the dataset (Unbalanced). As it is multiclass, I have used f1_micro score as our evaluation metric. This dataset is referred to henceforth as dataset 1.

Heart Failure Prediction Dataset - This dataset is created by combining 5 heart datasets over 11 common features. Cleveland: 303, Hungarian: 294, Switzerland: 123, Long Beach VA: 200, Stalog (Heart) Data Set: 270, Total: 1190 Duplicated: 272 Final dataset: 918 observations. All the 5 datasets originally sourced from UCI Machine Learning Repository.There are 918 rows 12 columns out of which 6 features are continuous(c), 5 our categorical(b) and the outcome is binary. Age(c), sex(b), fasting blood sugar(c), Exercise Angina(b) and old peak(c) are positively correlated to heart disease. The class distribution is as follows, 55.3% positive, 44.7% negative.(Almost Balanced). As false negative are most important for this dataset, recall evaluation is considered. This dataset is referred to henceforth as dataset 2.

How the datasets are interesting and different ? - The two datasets are different and interesting for this assignment, as dataset 1 has all continuous features and dataset 2 has 45% of features as categorical (PCA, ICA, RCA ignores categorical features[1]). The output class has different properties. Dataset 1 is unbalanced and has multiclass output. On the other hand, Dataset 2 is balanced and has binary class output. Dataset 2 is a combination of 5 different heart disease datasets, which also makes it interesting.

2 PART 1: CLUSTERING

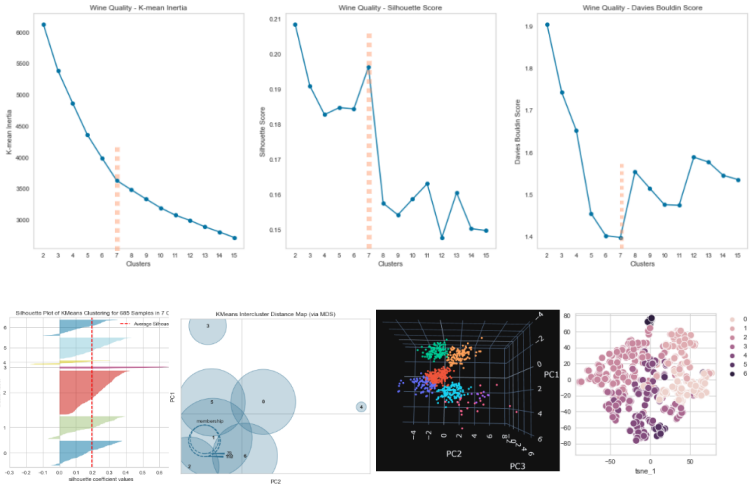
For this section, two algorithms were analyzed, K-means and Expectation maximization. K means is a hard clustering (no overlap, element either belong to cluster or not) algorithm, it sorts the data into K different clusters around nearest centroids. Each point is assigned to one center based on the distance matrix, and centers are reevaluated iteratively to minimize error. Gaussian Mixture Model, Expectation Maximization on the other hand is a soft clustering algorithm (cluster may overlap) where we fit a set of K Gaussian's to the data such that the likelihood of data given distributions is maximized.

Common Methodology - For **K-Means, K evaluation -** I used elbow Method on the scree plot to find the initial value of K.(Mean square distance between the sample and center of the cluster it belongs, lower value, means clusters are tightly formed). I also looked at cluster K value, which maximizes silhouette score (separation between different clusters) and minimizes Davis Bouldin(DB) Score (average similarity measure of each cluster with its most similar cluster[2]). Later, I used Mean shift clustering (it aims to discover "blobs" in a smooth density of samples) to check the suggested number of unique cluster centers. Furthermore, I used Silhouette Visualizer to check the density of the clusters for the value of K and I used inter-cluster distance maps to check inter cluster distances. After this analysis and choosing the value of K, For clusters visualization and validation, I plotted the clusters in 3 dimension using PCA. I validated my clustering by comparing metrics like homogeneity and accuracy score. Also, I used Dummy Random forest classifier with 80-20 train/test split, analyzed accuracy and overfitting, classifier's good performance suggests good clustering.

For **Expectation Maximization (EM), K evaluation -** To choose the value of K for EM, I used Akaike Information Criterion (AIC, estimates the relative amount of information lost by a given model[3]), the Bayesian Information Criterion (BIC, penalizes the complexity of the model, free parameters are penalized more strongly than AIC), and the silhouette score (As described in K-means section, it scores clusters on being compact and well separated). Also plotted Log-likelihood, which is monotonically non-decreasing for EM and validated its higher value. For clusters visualization and validation, I used the same plots and metric as K-means.

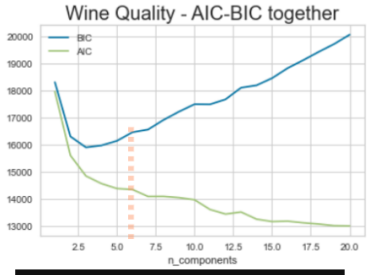
Dataset 1, Wine Quality Dataset-

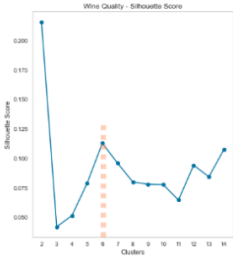
K-Means, K evaluation -

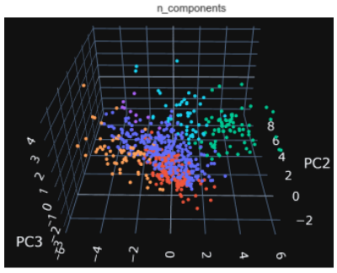


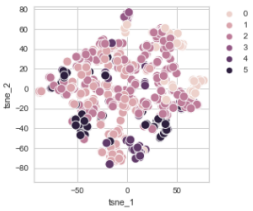
As we see from the plots on the left, we see a clear elbow at K=7, Also we see confirmation from the Silhouette score plot (0.195 peak) and Davis Bouldin plot(1.4 bottom) as well. This means the clusters at k = 7 are tightly formed, they are well separated with the least possible similarity. Mean shift cluster also suggest 7 clusters, i.e., it found 7 density blobs. Silhouette coefficient plot, we see 5 clusters with good density and 2 with small density. Also, from inter-cluster distance we see 5 dense plots are well separated, Both analysis can be confirmed, looking at PCA component 3 and T-SNE visualizations. The cluster distribution is as follows 32.85, 18.69,16.35, 16.06, 11.39, 3.07 and 1.61. Dummy classifier suggests

overfitting, this is because the test-train split might have different class labels due to limited data, so ignoring Classifier analysis. The homogeneity score is 0.08 and accuracy of 0.11, which suggests the class labels doesn't relate well with the clusters.





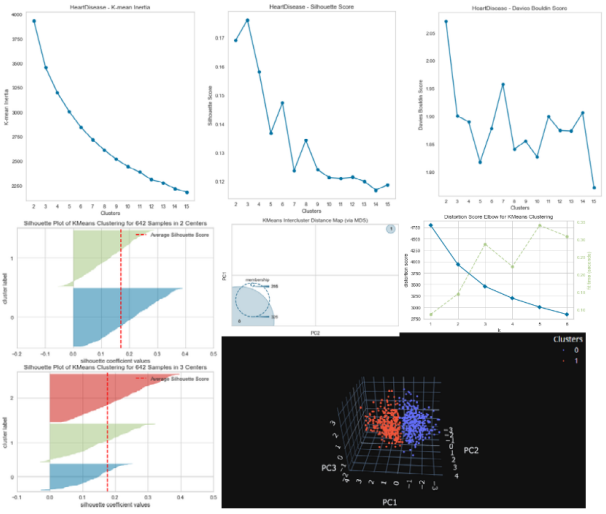




Expectation Maximization (EM), K evaluation - From the plots on the left, we see the AIC-BIC score dips around 3 to 6 but after 6 AIC score drops more but BIC increases, As BIC penalizes more for complex model, for this data, large number of clusters are more complex, hence penalized. From the silhouette score curve, we see an interesting drop at component=3, which I further analyzed with PCA plot, I could see the green cluster is all over the place, which causes this drop as the cluster is not separable, so low silhouette score. We see good silhouette score at component=6. The Dummy classifier analysis was rejected for the same reason as K-means. From the visualizations, PCA 3D plot showed clear cluster separations, but clusters are closely packed, which matches with silhouette peak 0.10 for k=6. Cluster distribution is 44.23, 26.57, 11.39, 8.91, 7.15 1.75. The homogeneity score is 0.067 and accuracy of 0.182. As expected, Accuracy score is better, as EM does soft clusters and doesn't have hard boundaries. Homogeneity score on the other hand is bad, it is possibly due to our preprocessing as actual data had 6 classes, we reduced it to 3 classes. **For Dataset-1 EM cluster performs better on accuracy metric.**

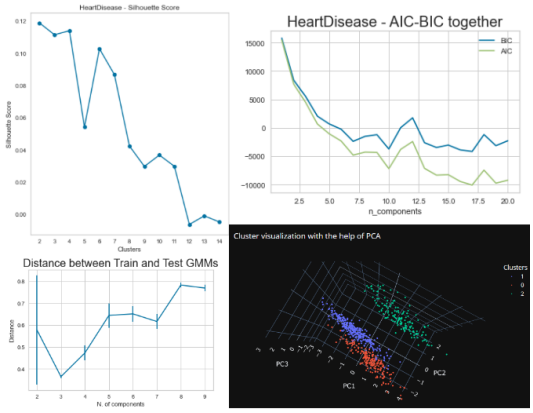
Dataset 2, Heart Disease Dataset-

This data set has 45% of features as categorical, K-Means algorithm doesn't work well on categorical features as they don't have inherent distance/similarity measure between the categorical samples. K-modes clustering algorithm works better on categorical dataset and k-prototype algorithm[4], which is a combination of k-means and k-modes, works better for mixed datasets. For EM, The Gaussian Mixture model also models Gaussian distribution to the data, it doesn't work well with Binary data. For this report, we constrained our analysis to K-means algorithm and Gaussian Mixture model on this dataset as well.



K-Means, K evaluation - Figure on the right, the K means inertia plot is smooth, but we see a slope change at 3 which suggests, the clusters are tightly formed, k can be 2 or 3. Silhouette plot suggest k as 2 & 3 again at k=2 its 0.175, k=2 its 0.17.Davis Bouldin score varies from 1.8 to 2, it has a drop at k=3 and k=5. But as the variation is low, it can be ignored. Mean shift cluster Elbow visualizer didn't work on this dataset, and it suggested one denisty blob as it ignore categorical values. It suggests, the non-categorical doesn't have density blobs. To choose between k=2,3 I checked Silhouette Visualizer, two clusters have similar density at k=2, and it is better than k=3. From the PCA 3 visualization, we can see, the data has smooth overall density, the clusters have equal den-

sity(51% & 49%) and are closer to each other. The dummy classifier, train, and test accuracy of 0.98 and 0.93, suggests well-defined clusters. Also, the homogeneity score is 0.25 and Accuracy score is 0.21 is better than Dataset 1, as it is linearly separable dataset and has binary output as suggested clusters.



Expectation Maximization (EM), K evaluation - For the silhouette score, we can see good value at cluster value 2,3,4. The highest silhouette score value of 0.12 suggests, the data points are very close. AIC-BIC score, 0's out around 3 and goes negative further. Therefore, 3 is suggested as per AIC BIC plot. Also, the GMM distances (Jensen-Shannon distance capturing similarity[5]) between the Gaussian distribution for randomly chosen test/train data is least at 3, suggests test/train GMM's are similar for component=3. Test/Train score for Dummy classifier is 1, suggests easily separable clusters.
The PCA visualization of cluster, agree with our analysis for k=3. The distribution is 46.57, 29.75 & 23.68. Also, the homogeneity score is 0.15 and Accuracy score of 0.24, better as EM soft clusters the data. **For Dataset 2, EM cluster performs better on accuracy and K-Means cluster does better on homogeneity**

Conclusion- For Dataset 1, the original data had 6 classes, but we preprocessed down to 3 classes, K-means 7 clusters and EM 6 clusters aligns approximately with 6 original classes, when I removed preprocessing (keeping it original 6 classes) and evaluated the Homogeneity score, it improved to 0.13, but it is still bad, i.e., the clusters line up naturally instead of lining up with the labels. For Dataset 2, even though the data has categorical features, K-means 2 clusters and EM 3 clusters has better accuracy score and homogeneity score. It lines up better with the actual labels. From Assignment 1 and 2 analysis, we know Dataset 1 has a lot of noise, which may be effecting the clusters. Also, EM performs better than K-means on accuracy for both the datasets, this is because EM is not constrained to spherical shape clusters, and it does soft clustering. For improving the K-means algorithm, we can try different seed, which will do different initialization and prevent K-means algorithm to get stuck. Also in the next section, we will discuss Independent Component Analysis(ICA) and clustering on ICA transformed components, which should also improve clustering (both EM and K-Means).

3 PART 2 & 3: DIMENSIONALITY REDUCTION & CLUSTERING-

According to "curse of dimensionality", if we add more features to the data, the amount of data that we need for training the model grows exponentially. Dimensionality reduction is the process of reducing the data features to a set of principal components (Usually less than original features). There are two approaches, feature selection or feature transformation, In feature selection, we select important features from original feature set and in feature transformation, we transform the features into lower dimensional space. For this project we analyzed, 3 feature transformation methods (principal component analysis (PCA), independent component analysis (ICA) and random projection (RP)) and 1 feature selection method Extremely Randomized Trees Classifier (ETC, uses Wrapping Technique as we select features after training). PCA find direction of maximum variance, it transforms the data into components which maximizes variance and these components are

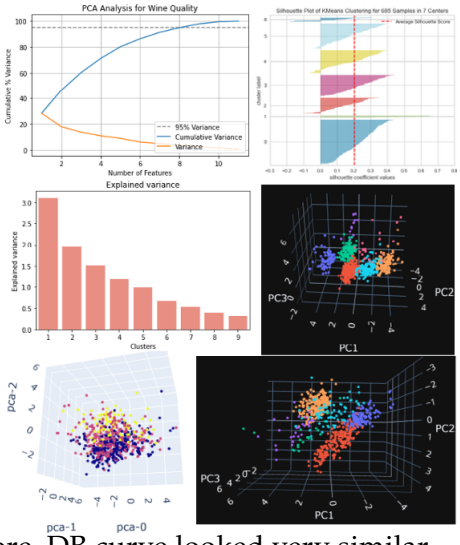
orthogonal to each other. Each PCA component has an Eigen value associated, which decreases monotonically, and the components with small Eigen value can be ignored, reducing the dimensionality. ICA approaches the problem as if the original features are the output of unobserved independent sources, it tries to maximize independence by finding the linear transformation of a feature space such that each of the individual new features are mutually independent, keeping the mutual information between original feature and new feature high, so we don't lose any information during transformation. RP is similar to PCA as it transforms the data in lower dimension, the difference is generated components are random and not constrained to be orthogonal as PCA, which makes it more efficient on the run time. Lastly we analyzed, ETC, it is an ensemble decision tree learning technique which gives us best feature selection based on the information Gain from the features.

Common Methodology- For **Dimensionality Reduction (DR) Analysis,** both the datasets, the 4 algorithms(PCA, ICA, RP & ETC) are applied. First we choose the best value of n (methods explained with Dataset 1 explanation), n is the number of components we chose from the algorithm, then to validate if chosen n is correct, we compute the reconstruction error, we also find the MSE for each reconstructed feature, once we select appropriate n, we visualize first two/three components of the reduced feature space and also run Dummy NN on the reduced feature set to confirm that we don't lose any information during transformation.

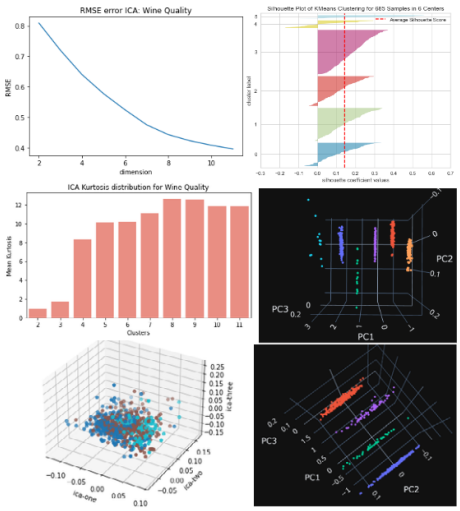
For **Cluster Analysis on Reduced Feature space,** For both the datasets, the same clustering methods from part 1 are applied to the 4 Reduced Feature space for each dataset, best k clusters is determined for new Reduced Feature space, also its compared to baseline clusters discussed in part 1.

Note- All images for this section, Left 3 images belongs to DR and right 3 images is for k-means (top 2) and EM(3) on the reduced feature space in order. **Clusters Metrics Comparison is discussed later.**

Dataset 1, Wine Quality dataset-

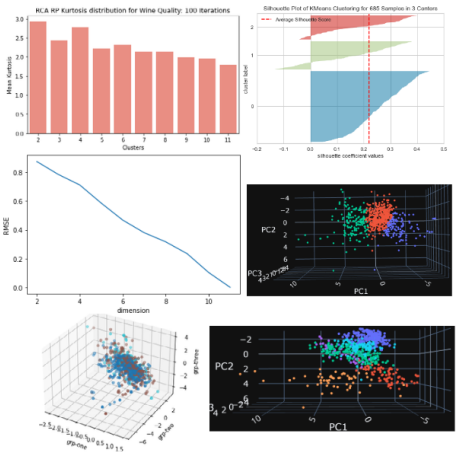


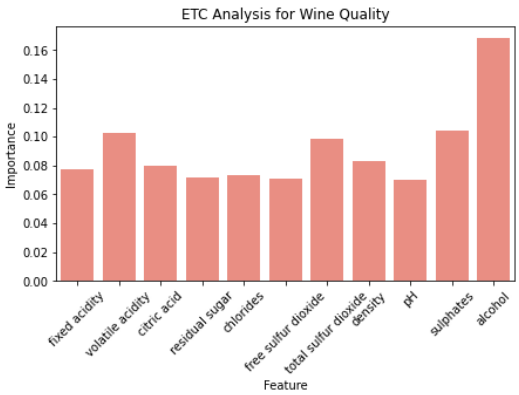
Principal Component Analysis (PCA)- To determine the PCA component length, I plotted, cumulative explained variance and explained variance of each PCA component, as explained above, PCA maximizes the variance, each PCA component has Eigen value/ explained variance, which decreases monotonically (bar plot on the right). Features which have least explained variance can be discarded. We consider 0.95 as our cumulative variance threshold, the components after that we discarded (as they offer only 0.05 explained variance). For this Dataset, we choose 9 components (original features 11) with reconstruction error of 0.021. The training f1_micro score for the reduced dataset is 74.01% in comparison to 72.26% on original features (n =10, perfect 74.01%), suggests no info loss on DR. When I plotted, PCA 3 components (0.50+ explained variance) and see classes segregating. Applying K-means to this reduced feature space, the Scree plot, Silhouette score, DB curve looked very similar to the original dataset and all suggested 7 clusters (same as original dataset). All 7 clusters on PCA 3 plot are well-defined, 5 clusters have good defined density and all the clusters have defined boundaries. EM also resulted in same cluster size 6 as original dataset, the elliptical clusters are well separated.

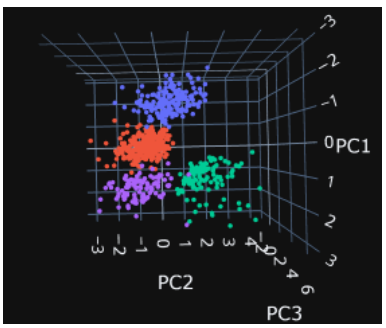


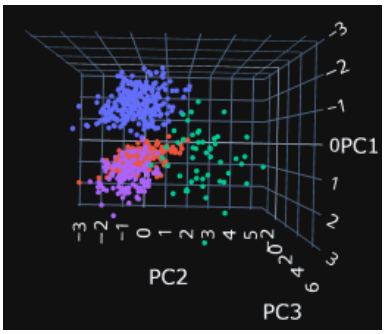
Independent Component Analysis (ICA)-
For ICA, it assumes the features are non-Gaussian which combines as per center limit theorem to a Gaussian output, so we used mean kurtosis to check for maximum non-Gaussian/independent features and RMSE for reconstruction to check on information loss, means high mutual information preserved at low RMSE. We choose 8 ICA component to be best (Kurtosis plot), with reconstruction MSE as 0.05. Applying K-means and EM on ICA reduced features, we see 6 clusters for K-means and 5 clusters for EM, Interestingly from PCA 3 plot for ICA clusters, we can see clusters are far apart on parallel planes for both K-means & EM. It is because of the ICA independent components. **(Plotly 3D Clusters rotated to show planes)**

Random Component Analysis (RCA)-
For RCA, I used GaussianRandomProjection, I plotted RMSE to component curve, to get the reduced component size, I ran the Random projection for 10, 100 and 500 iterations, I don't see much change, this is because the dataset doesn't have the high number of features, so on this small feature set RP doesn't show advantages with higher iterations. I choose RP with 10 components, it has 0.103 reconstruction error which is higher than other Algorithms, On the RCA reduced feature set, K-means got 3 clusters and EM got 6 clusters, RCA clusters are closer to each other in comparison to ICA and PCA.









Extra Tree Classifier (ETC)- Interestingly, the ETC algorithm chooses a set of only 4 features from 11 original features. ETC is a wrapping method, I have plotted feature_importances_ as a bar plot as we can see on left. 4 features, alcohol, sulfates, total sulfur dioxide and volatile acidity are chosen as selected features for Wine Quality Dataset. Bottom two PCA 3 component shows K-means (left, 4 clusters) and EM (right, 4 clusters).
Dataset 1, Data Transformation (PCA, ICA & RCA), Observations- PCA components density distribution is bell-shaped (Gaussian), whereas the ICA has sharper peaks as the components are independent, and the RCA component distribution is less smooth. As we discussed in previous section, ICA improves clustering,

we can see that from PCA 3 plots both EM and K-means clusters are aligned on parallel planes.

Wine Quality, Clustering Metric Comparison -

Below are two tables, listing various Metrics to compare the chosen best K-means Clustering and best EM clustering on the original and reduced Dataset 1.

Wine Quality K-Means Metrics

	init	time	inertia	homog	compl	v-mea	ARI	AMI	silhouette
0	Original	0.120000	3626	0.136000	0.081000	0.101000	0.074000	0.095000	0.196000
1	PCA-based	0.070000	4701	0.127000	0.096000	0.109000	0.084000	0.106000	0.189000
2	ICA-based	0.090000	4	0.128000	0.082000	0.100000	0.075000	0.094000	0.145000
3	RP-based	0.070000	4631	0.038000	0.039000	0.038000	0.057000	0.035000	0.219000
4	ETC-based	0.080000	1332	0.179000	0.138000	0.156000	0.119000	0.153000	0.262000

Wine Quality EM Metrics

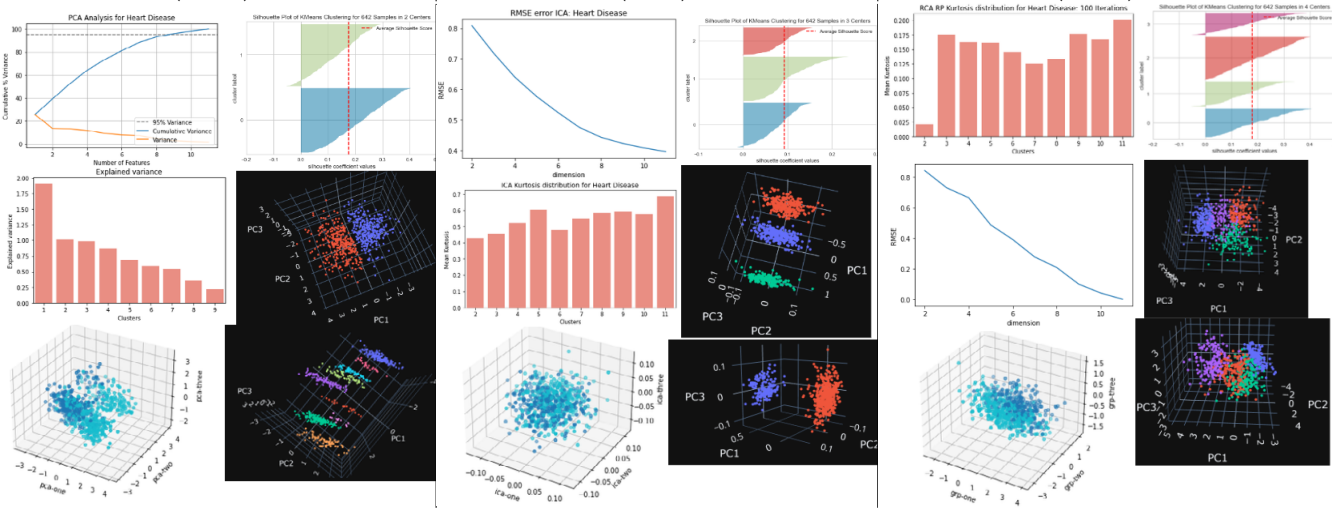
	init	time	aic	bic	homog	compl	v-meas	ARI	AMI
0	Original	0.120000	14260	16375	0.118000	0.075000	0.092000	0.068000	0.087000
1	PCA-based	0.090000	13800	15290	0.119000	0.076000	0.093000	0.064000	0.087000
2	ICA-based	0.100000	-22922	-21907	0.130000	0.093000	0.109000	0.113000	0.104000
3	RP-based	0.120000	7877	9666	0.121000	0.072000	0.090000	0.068000	0.085000
4	ETC-based	0.050000	6757	7024	0.193000	0.152000	0.170000	0.119000	0.167000

Original clusters 7(K-means) and 6(EM), the clusters changes after DR. For PCA, 7(K-means) and 6(EM). ICA, 6(K-means) and 5(EM), RCA, 3(K-means) and 6(EM). ETC, 4(K-means) and 4(EM). We got different clustering as the features are projected to different reduced space except PCA and ICA, so the cluster's alignment changes, as we see the reduced dataset has equal or fewer clusters, after DR, the outliers might be ignored and removed resulting in fewer clusters. From the tables, we can see ETC-based Clusters, beats both for K-means and EM Metric, this is because the ETC based dataset has just 4 best features, we also see the clustering time has direct correlation with Feature count, so we can say the clustering is faster after DR as dimensions are reduced. When we compare feature transformation DR algorithms, we can see RCA performs worse than the Original dataset, as RCA doesn't have smooth and symmetric density, which effects the clustering negatively. Also, our dataset has the small number of features, so we are not running RCA to its advantage. On K-means clustering we don't see expected metric improvement for ICA, but for EM, we can see the metric improved for ICA, as the components are independent and have high kurtosis peaks, clustering works better.

Dataset 2, Heart Disease Dataset-

We applied same Dimensionality Reduction algorithms on Dataset 2 which has 45% features as categorical. Same as clustering section, we used label encoding and also tried one hot encoding.

(PCA) (ICA) (RCA)



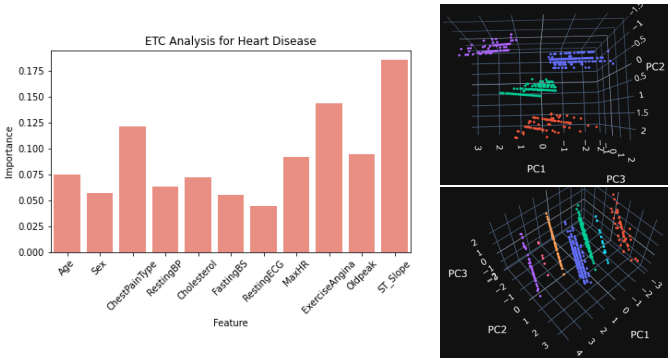
Principal Component Analysis (PCA)- From the PCA component's cumulative explained variance plot, I choose 9 components (original features 11) with reconstruction error of 0.171 (Reconstruction error for 10

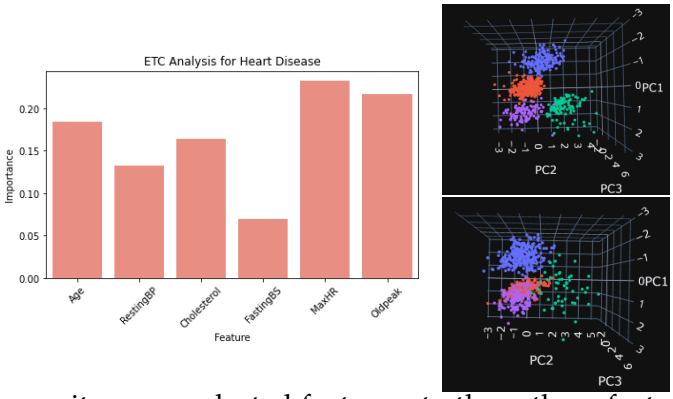
components is 0.017) The drop in the explained variance on subsequent PCA components (bar plot) is not smooth as continuous feature's dataset 1. On the PCA reduced Dataset 2, I applied K-means and EM, K-means got 2 clusters and EM resulted in 7 clusters. All the clusters for both K-means and EM are well separated, as in the figure (Left) above.

Independent Component Analysis (ICA)- Using the Reconstruction Mean square error plot, I choose 10 ICA component to be best, with reconstruction MSE as 0.012. From Kurtosis plot we see mostly negative kurtosis, it might be due to categorical features. Applying K-means and EM on these ICA reduced features, we see 3 clusters for K-means and 2 clusters for EM. Same as Dataset 1, for Dataset 2, the clusters for ICA are on parallel planes, as we see in the figure (Center). In Dataset 1, we saw similar parallel cluster planes.

Random Component Analysis (RCA)- By using Reconstruction Mean Square error to RCA component plot, I choose RP with 10 components, it has 0.041 reconstruction error. I ran the Random projection for 10, 100 and 500 iterations, I don't see much change, this is because the dataset doesn't have the high number of features, so on this small feature set RP doesn't show advantages with higher iterations. On the RCA reduced feature set, both K-means and EM, resulted in 4 clusters. The homogeneity score falls for both K-means and EM, as we also see from PCA 3 component plots the clusters shows overlaps.

Extra Tree Classifier (ETC)- There are interesting observations, when we use ETC with the dataset containing Categorical features. It selected 4 features (Original 11), out of these four features, 3 features are categorical. When I did K-means, EM clustering on these 4 features(3 categorical) below figure on left, we can see the 4 clusters with discrete layout. As it's not valid clusters because there is no inherent distance/similarity measure between categorical features.





So I ran the ETC on the dataset without categorical features, it gave 3 selected features, to these three features I added the 4 categorical features and plotted the clusters on the right side.

Heart Disease, Clustering Metric Comparison -

For Dataset 2, From the below table metric comparison, we can see the time is not proportional to the number of features, as it was for Dataset 1.

Heart Disease K-Means Metrics

	init	time	inertia	homog	compl	v-meas	ARI	AMI	silhouette
0	Original	0.070000	3935	0.250000	0.248000	0.249000	0.317000	0.248000	0.169000
1	PCA-based	0.060000	3744	0.264000	0.262000	0.263000	0.331000	0.262000	0.177000
2	ICA-based	0.070000	8	0.280000	0.179000	0.218000	0.246000	0.217000	0.100000
3	RP-based	0.080000	2981	0.087000	0.044000	0.058000	0.047000	0.056000	0.172000
4	ETC-based	0.060000	616	0.222000	0.116000	0.152000	0.182000	0.150000	0.332000

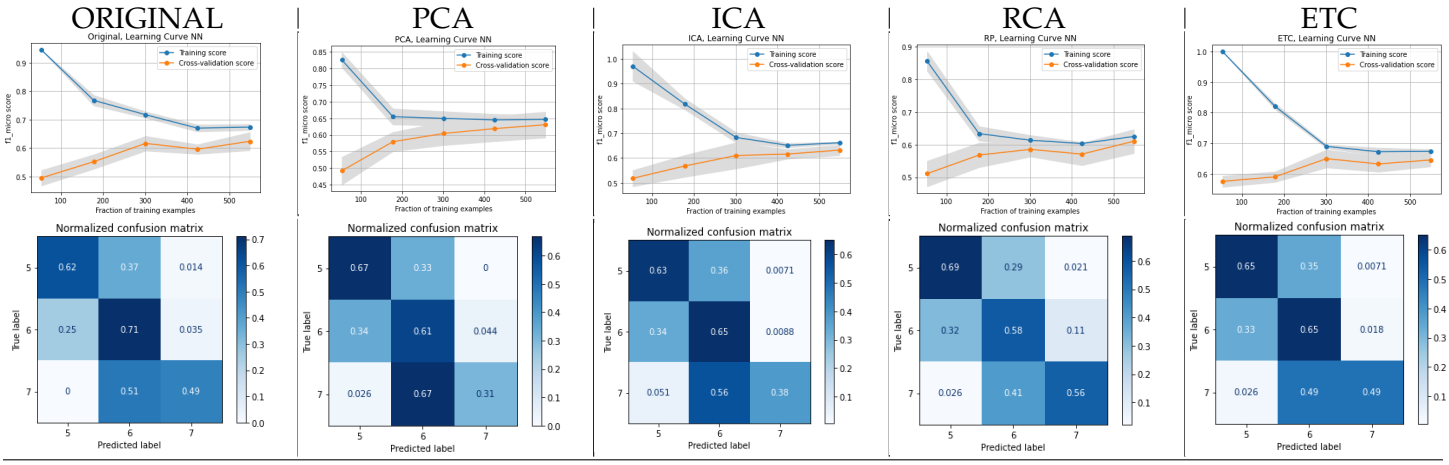
Heart Disease EM Metrics

	init	time	aic	bic	homog	compl	v-meas	ARI	AMI
0	Original	0.040000	801	1841	0.220000	0.143000	0.173000	0.189000	0.172000
1	PCA-based	0.110000	4320	6280	0.403000	0.152000	0.221000	0.181000	0.217000
2	ICA-based	0.060000	-26783	-25903	0.362000	0.235000	0.285000	0.327000	0.284000
3	RP-based	0.060000	614	1788	0.245000	0.140000	0.178000	0.197000	0.176000
4	ETC-based	0.100000	-9420	-8956	0.373000	0.163000	0.227000	0.272000	0.223000

The reason may be the categorical features not having inherit distance/similarity measure. For K-means, RCA reduced dataset again performs the worst in comparison to other datasets (Asymmetric distribution).Whereas PCA and ICA perform equivalent to Original Dataset for K-means. But again, for EM ICA performs the best (independent components, high kurtosis). **Detailed analysis, already covered in Dataset 1 section**

4 PART 4 & 5: NEURAL NETWORK PERFORMANCES

Neural Network Performance Comparison on Reduced Dataset - For this section, I have used my Dataset 1, with all the continuous features, The Dataset originally has 11 features, PCA reduces it to 9 components, ICA to 8, RCA to 10 and ETC (Random Forest classifier) reduced it to 4 features only.



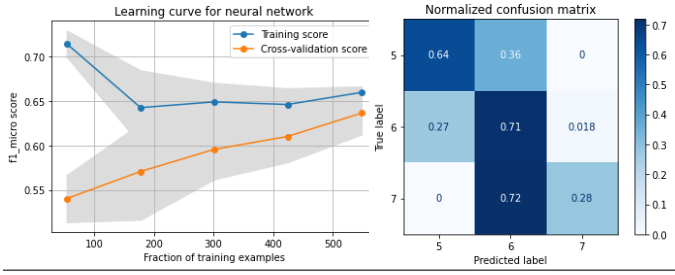
Dataset	Original	PCA	ICA	RCA	ETC	RCA-11
F1_Score	0.623	0.630	0.632	0.610	0.645	0.616
Train Time	149.14	132.40	93.52	124.45	56.37	113.50
Hidden Layers	4	2	18	2	24	10
Learning Rate	0.1	0.01	0.1	0.01	0.001	0.001

As we can see from the table, the train time correlates with the number of features. ETC performs the best when we compare the train time, as ETC has only 4 selected features it trains faster than other dataset with 8 plus features. On the f1_score comparison, almost all the Neural Networks performs equivalent to the Original Dataset, that means even though the dimensionality is reduced the information is still preserved. RCA performs the worst, We also saw in previous section, the reconstruction error on RCA is high even after selecting 11 components it doesn't preserve complete data information, One reason can be the data is small, it has only 11 features and around 1000 samples, RCA is at disadvantage.

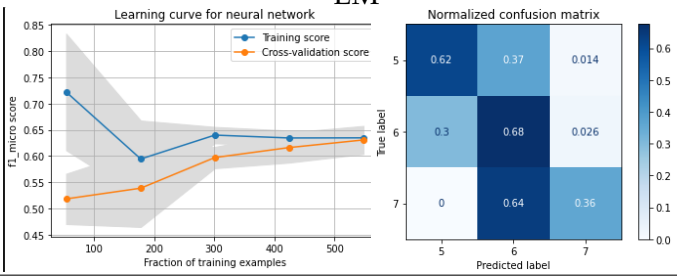
Conclusion, ETC performs the best on this dataset when we compare train time and f1_micro score. Overall, DR decreases the processing time of algorithms with lowest information loss and without losing accuracy.

Neural Network Performance Comparison on Dataset with clustered Feature - For this section, We took the clustered labels from K-means / EM Clustering and added them to the original dataset to create two new datasets for Neural Network performance comparison. In section 1, we choose 7 clusters for K-means and 6 clusters for EM, For both the clusters, we got almost all the data in well-defined clusters.

K-Means



EM



Dataset	Original	K-means Clustered	EM clustered
Score	0.623	0.636	0.630
Train Time	118.30	120.48	127.48
Hidden Layers	4	2	2
Learning Rate	0.1	0.1	0.001

The original dataset has 6 classes, but we cleaned the dataset for Assignment 1 and reduced the classes to 3. From the table above, we see train time increase as we have added a new feature, which contributes a little to the train time. But as this new feature has 7 and 6 labels respectively, it doesn't provide good information for 3 class out we are confined too. So this new feature from both K-means and EM doesn't provide information as the clusters doesn't align to considered classes, that's why the f1_score has little to no improvement.

5 FUTURE WORK

I want to do DR analysis on a larger dataset, with at least 100 features, and see the impact of DR and RCA advantages. Also for clustering, problem like image segregation are my interest to be explored in the future. I would also like to continue exploring on clustering with categorical data, where I would like to explore K-modes and K-prototype algorithm and other techniques. ICA's positive effects on clustering also caught my eyes and interest.

REFERENCES

[1] (n.d.). Retrieved November 3, 2022, from https://www.researchgate.net/publication/226400727_EM_Cluster_Analysis_for_Categorical_Data

[2] sklearn.metrics.davies_bouldin_score. (n.d.). Scikit-learn. Retrieved November 3, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[3] Wikipedia contributors. (2022, September 15). Akaike information criterion. Wikipedia. https://en.wikipedia.org/wiki/Akaike_information_criterion

[4] Into the world of clustering algorithms: k-means, k-modes and k-prototypes. (2016, October 26). AMVA4NewPhysics. https://amva4newphysics.wordpress.com/2016/10/26/into-the-world-of-clustering-algorithms-k-means-k-modes-and-k-prototypes/