Unsupervised Learning and Dimensionality Reduction

CS 7641: Assignment 3

Joseph Waugh, (Gatech ID: 903563084)

Abstract:

This paper will cover unsupervised learning and dimensionality reduction methods used for class prediction in two different datasets via clustering algorithms and neural networks. Specifically, this report will include results on K-Means Clustering and Expectation Maximization clustering on each of the datasets. Then, an analysis of Principle Component Analysis, Independent Component Analysis, Randomized Projection, and Linear Discriminant Analysis were applied on the dataset. An analysis of these same models was performed, based on the results from the original dataset transforming the input data, and then finally an application of these methods on a neural network.

Description and Importance of the Datasets:

The chosen datasets were selected from Kaggle. The specific datasets used in this report include the following:

1. Bank Marketing Dataset (Link):
    a. This data contains a list of demographic, socio-economic, and loan specific details resulting from a direct marketing campaign via phone calls from a Portuguese banking institution. The target variable here, is the decision in whether the client has subscribed to a product offered by the bank because of the campaign.
2. Heart Dataset (Link):
    a. This dataset contains a list of medical test results (i.e., electrocardiogram results, etc.), and demographic variables (i.e., age, sex, etc.) that may potentially play a role in heart disease. The target variable here is a binary variable establishing a patient as lower likelihood of heart attacks or a higher chance of a heart attack.

With regards to importance of the bank marketing dataset, the past 12 months have seen massive growth in the housing department, thus resulting in a recent increase in mortgage rates to slow the effects on inflation on the current state of the market. Home loan products are a competitive commodity among many large banks based on the increase in the interest rates from the Federal Reserve, where offering lower variable mortgage rates can allow the primary banks to win market share. Various machine learning methods can be applied to this type of dataset, in order to classify specific types of customers that determine which specific types of customers respond well to the bank marketing campaign based on whether or not the individual has signed up for one of the bank products, including housing loans.

As for the importance of the heart dataset, it has always been the case that heart disease is the leading cause of death for all major categories of demographics (i.e., sex, race, etc.). Therefore, it is of the utmost importance that determining key factors related to heart disease are identified to reverse this trend. Identifying these key factors can then be crucial in educating others about which specific tests and/or demographic characteristics are highly correlated with heart disease, which then can be used to address how to target how to reduce heart disease among these groups in addition to others.

Both datasets offer a unique distribution of categorical and numerical variables that allow for straight-forward usage in unsupervised learning algorithms. The datasets are moderately sized as well, which allows for a train/test methodology to be used to determine clusters of a given set of datapoints.

Data Pre-Processing:

To extract the dataset, the Kaggle API was used with both datasets. Once the information was extracted as saved to a CSV file, a pre-processing pipeline was used to process categorical and numerical columns efficiently. Specifically, a MinMaxScaler() was applied to numerical values in order to standardize the data, whereas a vectorization algorithm was applied in order to create numerical classes from categorical variables. The specified datasets were then spilt into a 80:20 ratio, with 80% of data belonging to the training dataset and 20% to the testing dataset.

Methodology – Clustering Algorithms & Dimensionality Reduction:

The clustering algorithms applied in this experiment aim to understand the data without the presence of class labels. By taking the predictor variables, a clustering algorithm can try to classify a given record based on these variables, which can then be used to compare against the true class labels (or create new labels altogether). Specifically, this experiment focuses on two clustering models: K-Means clustering, and Expectation Maximization clustering.

K-Means clustering begins with *n* randomly assigned center nodes (identified as centroids), which then change based on the additional datapoints that are added to each cluster that require a new "central position" of the cluster to be identified. This process of adding additional data points and updating centroids is iterated through until the cluster centers remain consistent after several iterations.

The Expectation Maximization (EM) algorithm utilizes the context of the training dataset to determine the joint probability of the data, and then apply maximum likelihood estimation to determine the associated class variables by "maximizing the likelihood" that datapoints belong to a specific cluster. This algorithm slightly differs from the K-Means clustering algorithm, where maximizing distances between is the goal to separate clusters, whereas for EM the goal remains to maximize the likelihood that a given datapoint belongs to a particular cluster.

The different dimensionality reduction algorithms applied to this dataset aim to reduce the set of predictor variables to avoid issues of multicollinearity, which is defined as the presence of multiple predictor variables that are highly correlated to each other. In the domain of machine learning, multicollinearity can result in poor test accuracy, given the predictor variables may be incorrectly weighted due to the variables having strong connections to other predictor variables. Thus, this technique helps to reduce overfitting among the models.

Principal Component Analysis is the first dimensionality reduction algorithm that reduces the set of variables into a smaller list. PCA has been referred to as the optimal method for dimensionality reduction in terms of accuracy; however, there are other algorithms that work well against some of the shortfalls of PCA. For example, Randomized Projection is a model that performs similarly, but with significantly less time required to process. In addition, this method requires less memory given that PCA requires all the data for a random projection. Independent Component Analysis (ICA) aims to maximize the space of each component based on a normalized Kurtosis score. Linear discriminant analysis looks for classifications but also utilizes the target label, and thus is a supervised learning algorithm.

Clustering Algorithms on Both Datasets:

The first performed experiment involved running the clustering algorithms on each dataset, and then describing the results of these algorithms. The following results were achieved on the Bank Marketing dataset:

	Clusters	Time (sec)	Homogeneity Score	Completeness Score	V-Measure Score	Adjusted Random Score	Adjusted Mutual Information Score	Silhouette Score	Accuracy Score
Expectation Maximization (EM)	2	0.64	0.02	0.01	0.02	0.04	0.02	0.28	0.35
	3	0.82	0.02	0.01	0.01	0.04	0.01	0.24	0.19
	4	1.03	0.03	0.01	0.01	0.02	0.01	0.26	0.38
	5	1.36	0.03	0.01	0.01	0.01	0.01	0.22	0.27
	6	1.80	0.03	0.01	0.01	0.03	0.01	0.24	0.10
	7	2.18	0.04	0.01	0.01	0.02	0.01	0.16	0.22
	8	2.17	0.04	0.01	0.01	0.01	0.01	0.14	0.18
	9	2.75	0.04	0.01	0.01	0.01	0.01	0.21	0.13
	10	3.28	0.04	0.01	0.01	0.01	0.01	0.11	0.08
K-Means Clustering	2	0.58	0.11	0.06	0.08	0.16	0.08	0.14	0.25
	3	0.95	0.11	0.05	0.07	0.18	0.07	0.18	0.70
	4	1.86	0.14	0.05	0.07	0.11	0.07	0.05	0.57
	5	2.41	0.15	0.04	0.07	0.07	0.07	0.07	0.43
	6	3.56	0.13	0.03	0.05	0.09	0.05	0.00	0.53
	7	6.07	0.15	0.04	0.06	0.09	0.06	0.00	0.05
	8	10.37	0.15	0.03	0.06	0.07	0.06	-0.12	0.26
	9	5.08	0.17	0.04	0.06	0.07	0.06	-0.08	0.28
	10	8.46	0.17	0.03	0.06	0.04	0.05	-0.16	0.28

Figure 1: *Clustering Algorithm Results – Bank Marketing Dataset*

To calculate the optimal k clusters in the K-Means clustering algorithm, an elbow method was applied to determine the best tradeoff between increasing the number of clusters and the associated sum of squared distance. Based on that methodology, the optimal number of clusters for the Bank Marketing dataset was 4. The results for k=4 indicate a high silhouette score, which is indicative of clusters that are separated as far as possible (values near 0 indicate overlapping clusters; -1 represents incorrect cluster classifications based on dissimilar datapoints). The overall time increase appears to be linear based on the number of clusters resulting in a standard increase in the amount of computational time required to process an output.

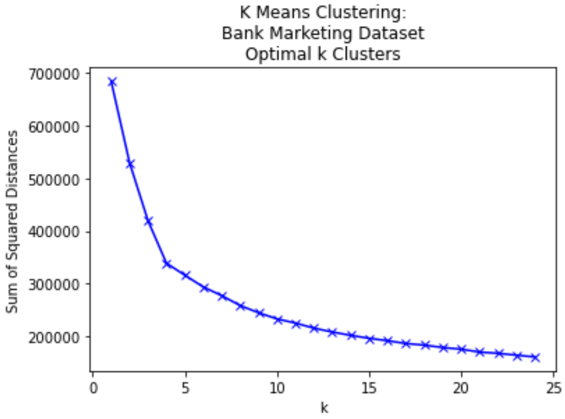


Figure 2: *Sum of Squared Differences vs. K – Bank Marketing Dataset*

Regarding the Expectation Maximization outputs, the same number of clusters were tested to determine the optimal k clusters for predicting this dataset. The results showed that k=3 gave the optimal result compared to other values. Specifically, the Completeness Score compared to the Homogeneity score appears to be negatively correlated, which can be attributed to the fact that all datapoints need to belong to the same class within each cluster for the completeness score, whereas homogeneity allows for this to occur since the score is purely based on datapoints only belonging to a single cluster.

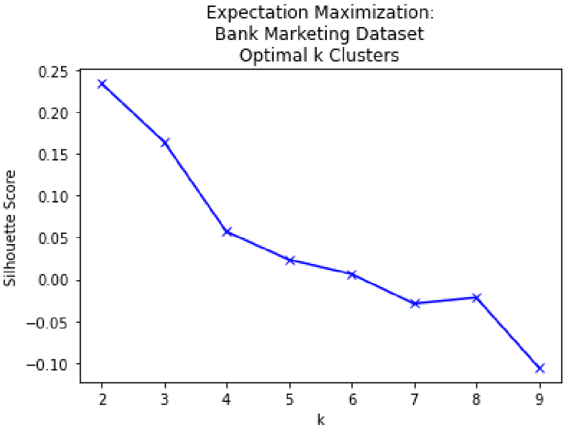


Figure 3: *Silhouette Score vs. K – Bank Marketing Dataset*

Next, the same experiment was replicated for the heart dataset to classify the datapoints using K-Means Clustering and Expectation Maximization. The results of that algorithm are shown next:

	Clusters	Time (sec)	Homogeneity Score	Completeness Score	V-Measure Score	Adjusted Random Score	Adjusted Mutual Information Score	Silhouette Score	Accuracy Score
Expectation Maximization (EM)	2	0.02	0.16	0.17	0.17	0.20	0.16	0.24	0.27
	3	0.02	0.16	0.12	0.13	0.16	0.13	0.27	0.43
	4	0.03	0.18	0.10	0.13	0.12	0.12	0.26	0.24
	5	0.02	0.21	0.10	0.13	0.12	0.13	0.30	0.22
	6	0.05	0.24	0.10	0.14	0.10	0.14	0.29	0.22
	7	0.04	0.28	0.10	0.15	0.12	0.14	0.23	0.08
	8	0.04	0.31	0.11	0.16	0.14	0.15	0.20	0.07
	9	0.09	0.29	0.10	0.15	0.13	0.14	0.20	0.06
	10	0.07	0.31	0.10	0.16	0.12	0.14	0.19	0.06
K-Means Clustering	2	0.06	0.16	0.17	0.17	0.20	0.16	0.25	0.73
	3	0.09	0.21	0.14	0.17	0.15	0.16	0.23	0.39
	4	0.11	0.19	0.10	0.13	0.12	0.13	0.26	0.43
	5	0.12	0.20	0.09	0.13	0.11	0.12	0.30	0.08
	6	0.11	0.27	0.11	0.15	0.14	0.15	0.22	0.08
	7	0.11	0.26	0.10	0.14	0.10	0.13	0.20	0.17
	8	0.16	0.28	0.10	0.15	0.09	0.14	0.21	0.29
	9	0.17	0.27	0.09	0.14	0.08	0.13	0.21	0.14
	10	0.16	0.36	0.11	0.17	0.12	0.16	0.20	0.06

Figure 4: *Clustering Algorithm Results – Heart Dataset*

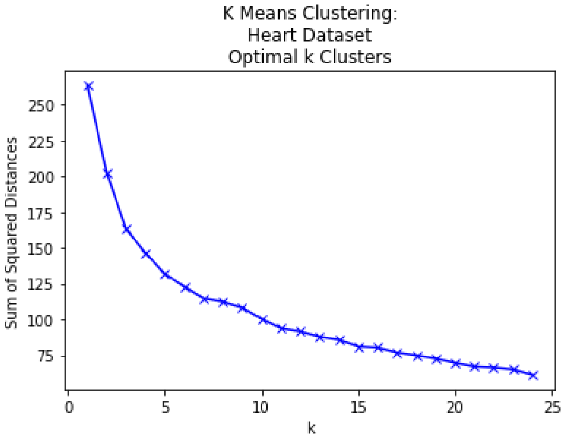


Figure 5: *Sum of Squared Differences vs. K – Heart Dataset*

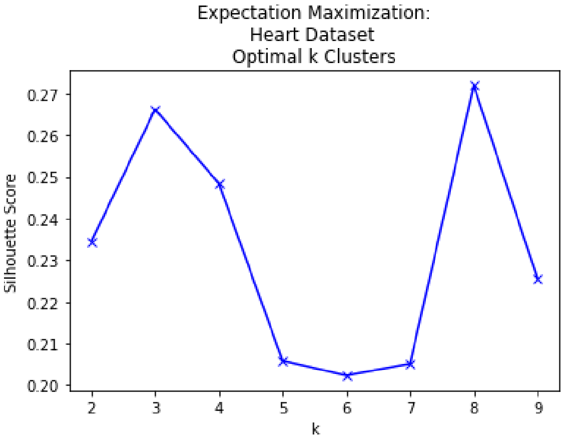


Figure 6: *Silhouette Score vs. K - Heart Dataset*

The same elbow method was applied to this dataset to calculate the optimal k clusters via K-Means clustering. Based on this method, the optimal number of clusters here was 3, based on a high silhouette score and overall accuracy score based on the true class labels. The time required for each number of clusters appears to trend in a linear direction; however, the results aren't 100% linear given that some lower K values have a longer runtime vs. higher K values.

For the Expectation Maximization outputs, again the same number of clusters were applied to determine the optimal K clusters for predicting the dataset. The results showed the optimal number of clusters here was 2, based on the maximized adjusted random score and 2nd highest silhouette score. As the number of clusters start to increase, the overall accuracy of the results decreases significantly, thus requiring an optimal value that is lower due to overfitting.

The V-Measure scores appear to be much higher for the Heart dataset. This score combines both the homogeneity and completeness scores into a single metric, which would work in terms of being able to predict an optimal K value based on both metrics without bias. The Rand Index (Adjusted Random Score) works to account for variance in the data based on the various clusters of the samples. The scores for both datasets appear to show that a larger number of clusters have a lower score compared to those with high values of K. This can be attributed to the underfitting of data with too few clusters to accurately cluster the data. Adjusted Mutual Information was also used to show the optimal mutual information based on the clusters. This metric is similar to the adjusted random score metric; however, this score works well when the dataset contains an uneven distribution of target classes in the dataset. Given that the Bank Marketing dataset and Heart dataset are both balanced in terms of true class variables, this was purely for observation to see if the same trends observed from the optimal K value can be determined using the Adjusted Random score and Adjusted Mutual Information score.

Dimensionality Reduction Algorithms on Both Datasets:

The first applied dimensionality reduction algorithm was Principal Component Analysis (PCA). This algorithm again aims to reduce the number of variables, especially those that are correlated to reduce multicollinearity. It's often the case that the first few components via PCA can explain most of the variance in the dataset. This is evidence when plotting the explained variance against the number of components, that the curve appears to rise sharply until it starts to form a log curve that shows each additional component only explains an increased marginal amount of the variance.

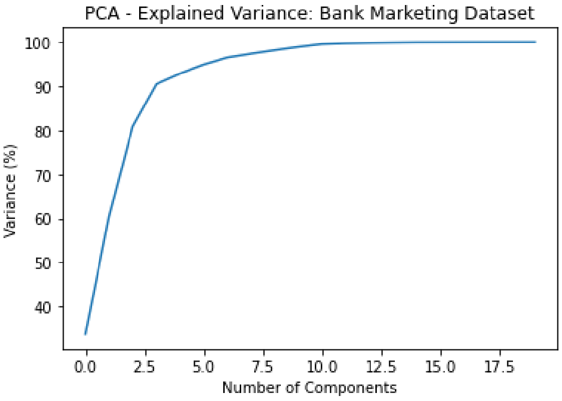


Figure 7: *PCA Explained Variance - Bank Marketing*

With the Bank Marketing dataset and a baseline of 80% explained variance, we can use 3 principal components to explain most of the data in the visualization. This visualization is shown right, and it appears to show that the customers who subscribed to the bank products can largely be predicted based on negative scores of PC1, negative scores of PC2, and mostly any score of PC3; however, customers who aren't subscribed have values with higher PC1 and PC2 values, and sporadic PC3 values that don't have a particular trend.

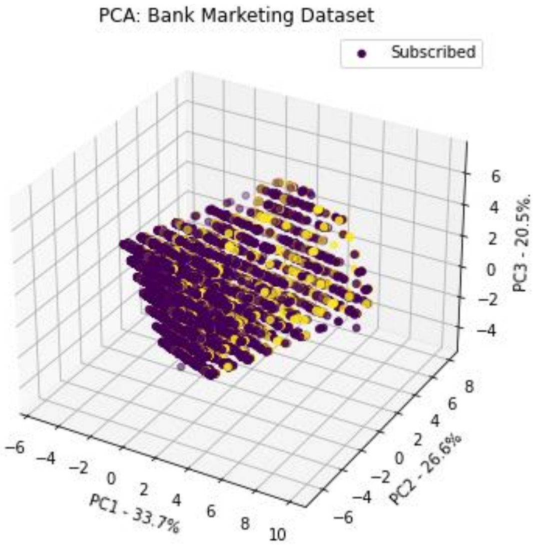


Figure 8: *PCA Visualized - Bank Marketing Dataset*

As for the Heart dataset, this dataset required fewer principal components to account for 100% of the variance in the dataset, with 12 components required vs. the 18 required for the Bank Marketing dataset. With that being said, the first three clusters are used this time (despite not making the 80% threshold, but for visualization purposes), and is responsible for explaining 58.6% of the data. This visualization is shown below as well. The variance calculations made in each of these visualizations are based on the explained_variance_ratio function, which works to represent the variance explained based on each eigenvector.

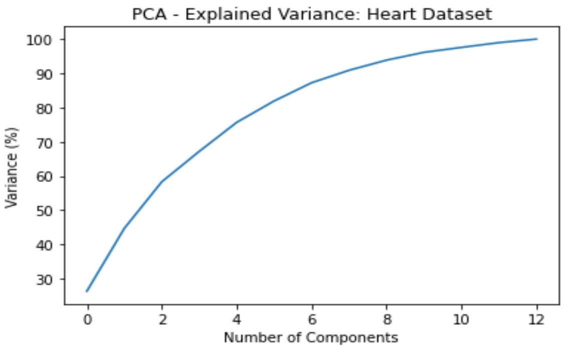


Figure 9: *PCA - Explained Variance: Heart Dataset*

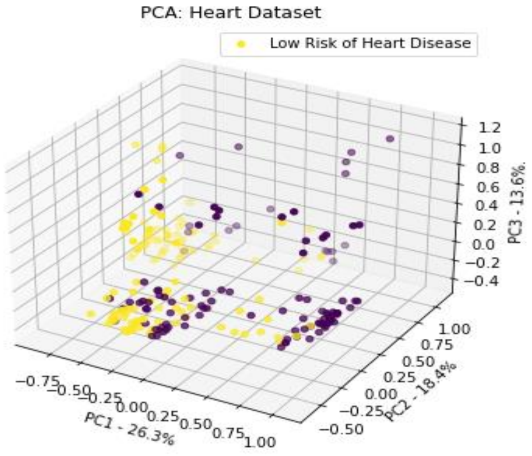


Figure 10: *PCA - Explained Variance: Heart Dataset*

For the independent component analysis, given that this function aims to maximize kurtosis (measures the extent in which the scores from the cluster fall within the tails or peak of a frequency distribution), the number of independent components here appears to be 20 for the bank marketing dataset. As for the heart dataset, the number of independent components appeared to be 11.

Randomized Projection was the next dimensionality reduction algorithm that was included in this analysis. Specifically, the dimensionality is reduced based on a similar method to PCA, but instead focusing on faster computational results with a slight trade-off on accuracy out the clustering ability. Specifically, a GaussianRandomProjection was utilized with 30 components initialized, and a low EPS score to allow the data to be mapped into a higher dimensional space. The output comparing two randomized projections is shown below. For the heart dataset, it appears that the two projections can separate the classes of the bank marketing dataset moderately well, with the subscribed individuals (purple) for bank products encompassing the top range of the data, whereas the non-subscribers are mostly on the bottom of the range. As for the heart dataset, there wasn't a clear differentiation among the two classes. This suggests that the model's dimensionality reduction doesn't perform well with this specific dataset, which can be attributed to the size of the data being too small, or the data isn't representative enough to make an accurate prediction on how to separate the classes.

The Linear Discriminant Analysis was the selected choice of an additional method of dimensionality reduction, given the possibility to retain the class information but still maximize separation for clustering. Given the binary nature of each class output for both datasets, a single component was selected.

Clustering Algorithms on Dimensionality Reduced Dataset – Bank Marketing:

The same methodology applied in the first experiment was applied here, with the only difference being the usage of the dimensionality reduced datasets instead of the raw data and the focus on the Bank Marketing Dataset. The results of these models are shown below:

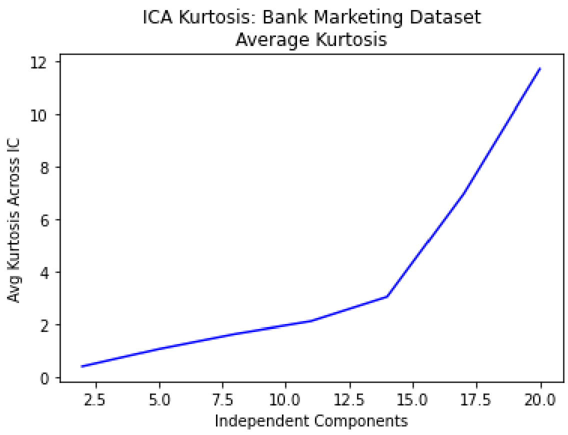


Figure 11: *ICA Kurtosis: Bank Marketing Dataset*

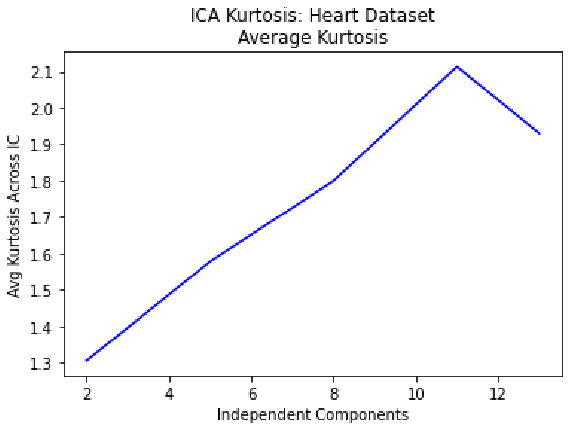


Figure 12: *ICA Kurtosis: Heart Dataset*

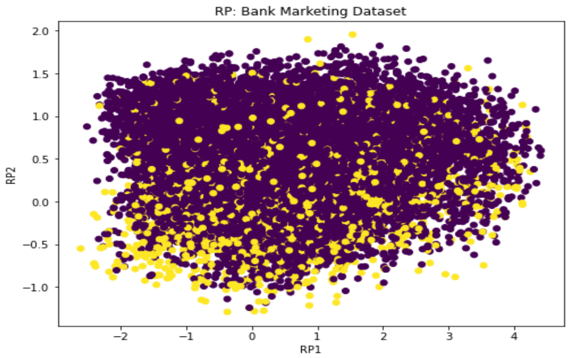


Figure 13: *Randomized Projection - Bank Marketing Dataset*

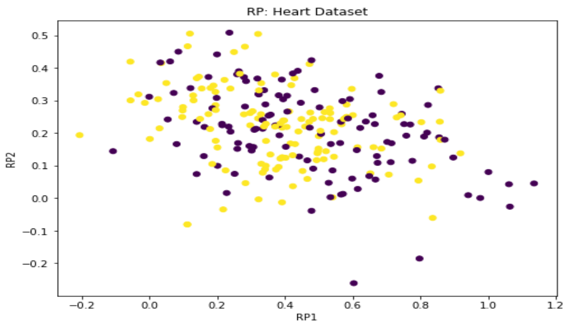
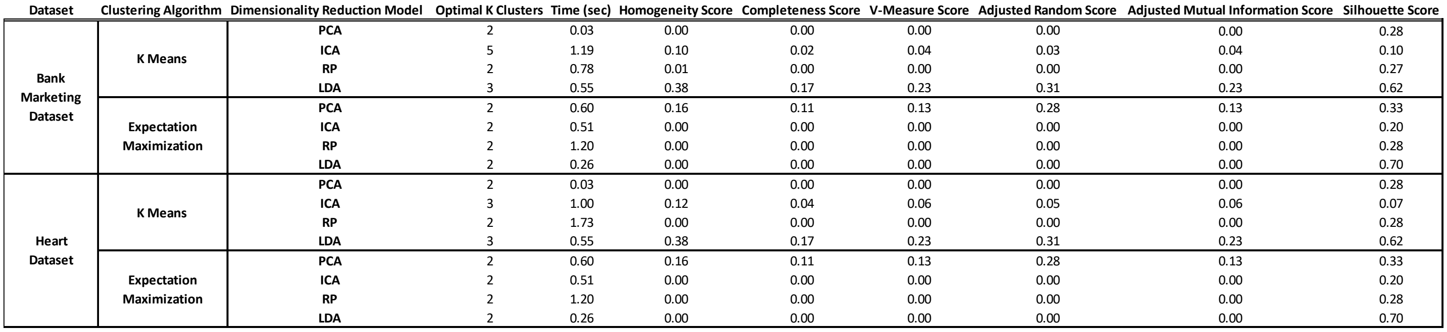


Figure 14: *Randomized Projection - Heart Dataset*

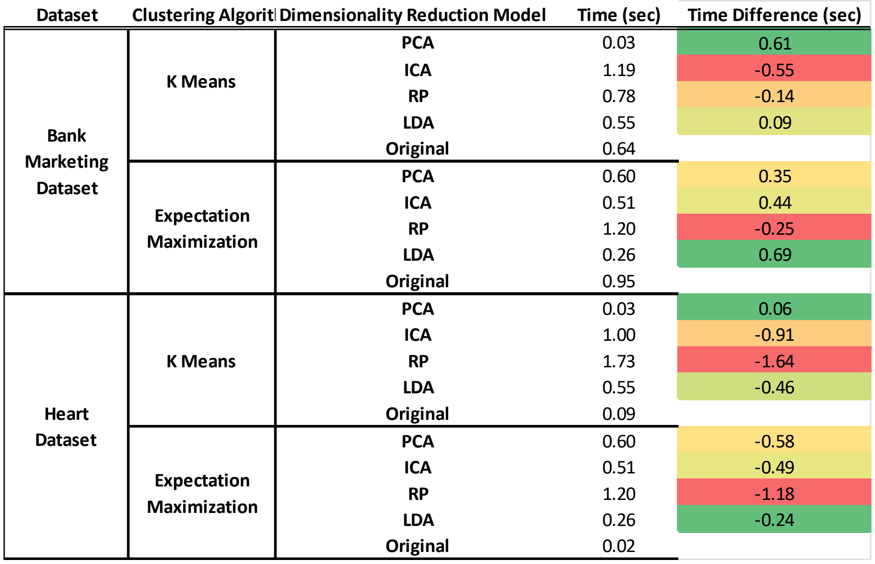


Dataset	Clustering Algorithm	Dimensionality Reduction Model	Optimal K Clusters	Time (sec)	Homogeneity Score	Completeness Score	V-Measure Score	Adjusted Random Score	Adjusted Mutual Information Score	Silhouette Score
Bank Marketing Dataset	K Means	PCA	2	0.03	0.00	0.00	0.00	0.00	0.00	0.28
		ICA	5	1.19	0.10	0.02	0.04	0.03	0.04	0.10
		RP	2	0.78	0.01	0.00	0.00	0.00	0.00	0.27
		LDA	3	0.55	0.38	0.17	0.23	0.31	0.23	0.62
	Expectation Maximization	PCA	2	0.60	0.16	0.11	0.13	0.28	0.13	0.33
		ICA	2	0.51	0.00	0.00	0.00	0.00	0.00	0.20
		RP	2	1.20	0.00	0.00	0.00	0.00	0.00	0.28
		LDA	2	0.26	0.00	0.00	0.00	0.00	0.00	0.70
Heart Dataset	K Means	PCA	2	0.03	0.00	0.00	0.00	0.00	0.00	0.28
		ICA	3	1.00	0.12	0.04	0.06	0.05	0.06	0.07
		RP	2	1.73	0.00	0.00	0.00	0.00	0.00	0.28
		LDA	3	0.55	0.38	0.17	0.23	0.31	0.23	0.62
	Expectation Maximization	PCA	2	0.60	0.16	0.11	0.13	0.28	0.13	0.33
		ICA	2	0.51	0.00	0.00	0.00	0.00	0.00	0.20
		RP	2	1.20	0.00	0.00	0.00	0.00	0.00	0.28
		LDA	2	0.26	0.00	0.00	0.00	0.00	0.00	0.70

Figure 15: *Clustering Algorithm & Dimensionality Reduction Result Comparisons – Bank Marketing Dataset & Heart Dataset*

The results here indicate a few interesting trends. Based on the results in both the bank marketing dataset and the heart dataset, the number of optimal K clusters is higher for Independent Component Analysis models versus any other dimensionality reduction model. This can potentially be related due to the complex nature of the model (i.e., the overall size of the dataset and the number of components used initially can result in a more complex output), and thus result in a higher number of clusters that would be required to accurately separate the data. Conversely, the lower complexity of the other dimension reduction models likely resulted in a lower number of optimal K clusters as a result. For example, LDA's 2-3 K clusters can potentially be attributed to the usage of 1 component (given the binary target variable), which requires only processing of 1 component vs. many more.

From a runtime comparison, there was a slight decrease (though not majorly significant) in the runtimes based on the dimension reduced data vs. using the raw data for clustering. This visualization can be shown below:



Dataset	Clustering Algorit	Dimensionality Reduction Model	Time (sec)	Time Difference (sec)
Bank Marketing Dataset	K Means	PCA	0.03	0.61
		ICA	1.19	-0.55
		RP	0.78	-0.14
		LDA	0.55	0.09
		Original	0.64	
	Expectation Maximization	PCA	0.60	0.35
		ICA	0.51	0.44
		RP	1.20	-0.25
		LDA	0.26	0.69
		Original	0.95	
Heart Dataset	K Means	PCA	0.03	0.06
		ICA	1.00	-0.91
		RP	1.73	-1.64
		LDA	0.55	-0.46
		Original	0.09	
	Expectation Maximization	PCA	0.60	-0.58
		ICA	0.51	-0.49
		RP	1.20	-1.18
		LDA	0.26	-0.24
		Original	0.02	

Figure 16: *Clustering Algorithm & Dimensionality Reduction - Runtime Comparison*

Apply Clustering Algorithm on Neural Network Learning Using Dimensionality Reduced Data:

To generate a neural network, a Multi-Layer Perceptron model was used to test performance using the different clustering algorithms and dimensionality reduced models. Stochastic Gradient Descent was used as the solver in this experiment (using backpropagation to update gradients to maximize the prediction accuracy), with a learning rate of 0.1 initiated, hidden layer size of 4, and momentum value of 0.1 to allow the experiment to escape any local optima. This is a relatively simple application of the MLPClassifier function, but nonetheless it will allow for a fair comparison of the different clustering and dimensionality reduction algorithms in terms of prediction power.

For this experiment in the Bank Marketing Dataset, the non-transformed data applied to PCA and ICA appeared to be accurately predicted in terms of a similar score of training and testing accuracy; however, for Randomized Projection and Linear Discriminant Analysis, the train accuracy was significantly higher which indicates that these models were overfitting from what was expected. This result can be seen right. In looking at these results, the first reaction I see is that the LDA algorithm and the PCA algorithm perform relatively well in terms of training accuracy, with a slight reduction in accuracy for LDA when it comes to test accuracy; however, ICA and RP perform much worse when it comes to test accuracy in comparison. This can potentially be explained by LDA featuring the lowest complexity in terms of the number of components used in the model. Given the low complexity of the dataset in this experiment, this could be the key in understanding the high performance of the model.

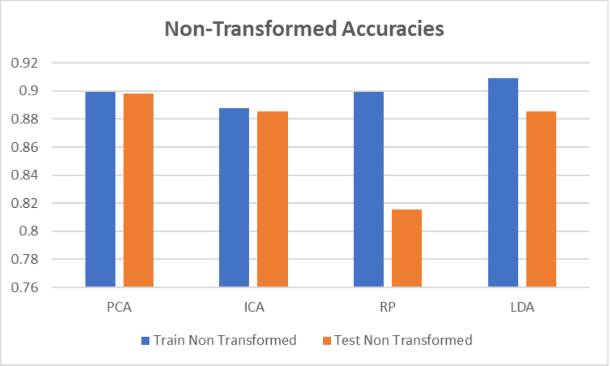


***Figure 17:** Dimensionality Reduced Data Accuracy – Bank Marketing Dataset*

In comparing the same results for the Heart Dataset, the same trend can be seen here with overfitting occurring on almost every algorithm. Similar to what was viewed above, LDA appears to have the highest train and test accuracy, with PCA showing strong performance while ICA and RP show weaker performance as a comparison. The explanation provided above regarding LDA's simplicity and the relative simplicity of the Heart dataset could be the explanation behind these results; however, the weak performance again could potentially be due to overfitting resulting from a non-representative training and testing dataset, or not enough datapoints to accurately represent either of those subsets of data. In addition, the complexity of the models could just simply not correspond well to the lower-dimensional datasets that were tested in this experiment.

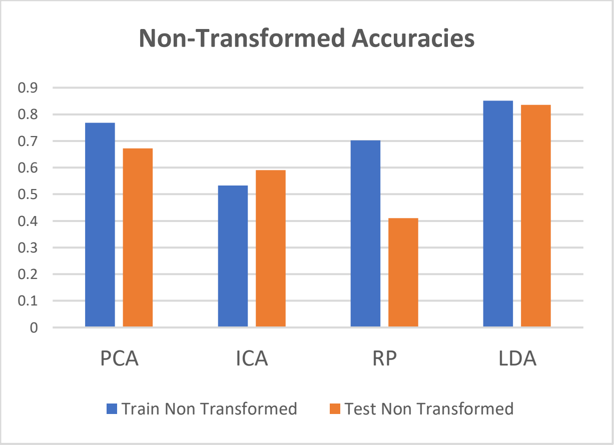


***Figure 18:** Dimensionality Reduced Data Accuracy – Heart Dataset*

Apply Clustering Algorithm on Neural Network Learning Using Newly Projected Data:

To generate a neural network, a Multi-Layer Perceptron model was again used to test performance using the different clustering algorithms and dimensionality reduced models. The same parameters were used as above, with the only difference being the usage of the projected data.

In comparing the accuracies of the Bank Marketing dataset, the initial results showed strong performance for PCA and LDA in both training and testing accuracy. Randomized projection similarly showed strong training data accuracy but failed to produce the same level of test accuracy in this dataset.

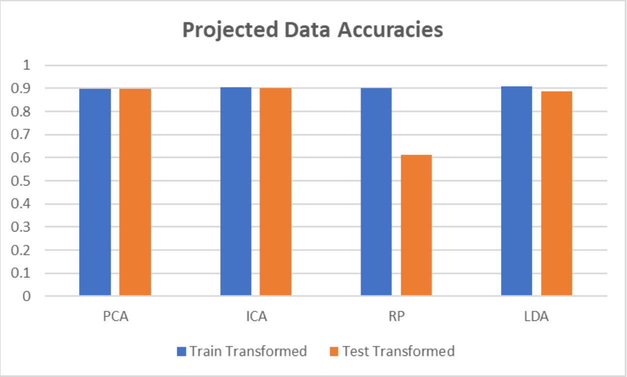


***Figure 19:** Projected Data Accuracy – Marketing Data Dataset*

As for the accuracies with the heart dataset, strong performance from PCA and LDA was observed, whereas weaker performance from RP suggests that overfitting resulted in a higher training accuracy compared to the test accuracy. The ICA model interestingly performed worse in the training accuracy vs. test accuracy; however, the result shows that the model was still not able to accurately predict the class labels as what was required.

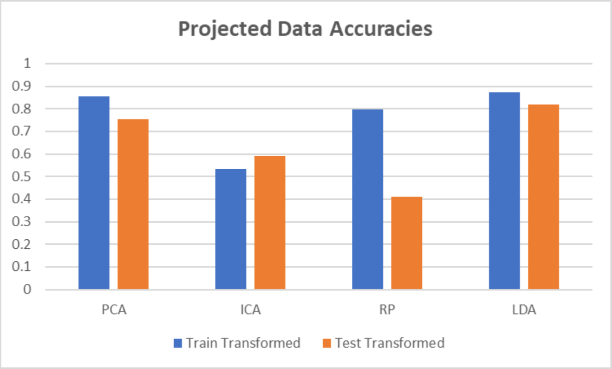


Figure 20: *Projected Data Accuracy – Heart Dataset*

Conclusion:

Overall, throughout the five experiments with the various clustering and dimension reduction algorithms, various trends were observed regarding how the data is clustered and transformed results in changes to runtime, class label accuracy (compared to original data), and overall distinctness in the data for each cluster based on a variety of different metrics. In future learnings, applying larger datasets with increased numbers of clusters may result in a potentially better accuracy score for several of the models (i.e., Independent Component Analysis & Randomized Projection), as the datasets used here featured several instances of overfitting on the training data. Nonetheless, the value of using dimension reduction for reductions in required computational power were assessed with tradeoffs in model accuracy. Additional iterations of these experiments would likely continue to show these trends, and my expose more trends specific to the datasets applied, and the logic applied with each algorithm.

Sources:

- Code Directory: Link