# Assignment 3 - Unsupervised Learning and Dimensionality Reduction

Michael Lukacsko
*mlukacsko3@gatech.edu*

***Abstract*—This assignment is a multi-layered analysis of unsupervised learning algorithms. Two clustering algorithms (k-means and Expectation Maximization) are implemented along with four dimension reduction algorithms (PCA, ICA, Randomized Projections, and Extra Trees Classifier). Each of the six total algorithms are optimized using two classification datasets. Finally, a neural network algorithm is implemented using the dimensionally reduced dataset as well as the dataset that has had the respective clustering algorithm applied to it. The results of each step are analyzed and discussed.**

***Keywords—clustering, k-means, Expectation Maximization, dimension reduction, PCA, ICA, RP, extra tree classifier, neural network.***

## I. Part 1: Clustering Algorithms

K-means and Expectation Maximization (EM) are both unsupervised machine learning algorithms used for clustering. They differ, however, in their approach to cluster assignment and algorithmic implementation.

K-means is a simple and popular clustering algorithm that aims to partition data points into K distinct clusters, where K is a predetermined number of clusters. The algorithm iteratively assigns data points to the closest centroid (mean) of a cluster and recalculates the centroid until convergence. Expectation Maximization (EM) is a more complex clustering algorithm that is based on a probabilistic model, specifically the Gaussian Mixture Model (GMM). In EM, each data point is assumed to belong to a mixture of Gaussian distributions, and the algorithm aims to estimate the parameters of these distributions.

### A. *Wine Dataset*

#### 1) *K-Means*

The optimal number of clusters determined by implementing and running k-means is 3. Typically known as the 'elbow method', the optimal number of clusters is where an elbow is visible on a plotted line relative to the number of clusters. In the plot below, the left image shows this feature well. Noted by the red arrow, an elbow is present between where the inertia values decrease dramatically with cluster value of 1 to 2 and levels off with between 3 and 25 clusters. As such, the inertia plot on the left determines the optimal number of clusters to be 3. This is supported by the silhouette score plot on the right. The silhouette score, as a metric used to evaluate the quality of clustering, ranges from -1 to 1. As can be seen in the plot, the highest silhouette score is present where the number of clusters is equal to 3.

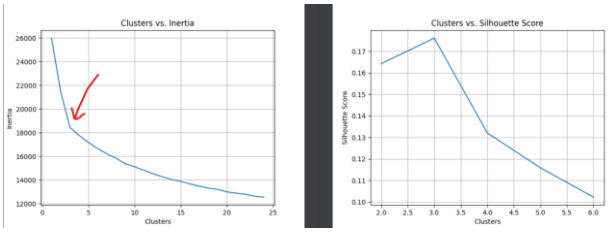


Figure 1 – Clustering Algorithm. K-means inertia and silhouette scores.

Exploring the silhouette score more deeply, the plot in figure two below shows a silhouette analysis using 3 clusters. The image on the left shows the average silhouette score for each cluster is above the overall average silhouette score noted by the vertical red line. This is an indication that the clusters are well separated. On the right, in the plot below, is a visual representation of the similarity of data points within and between clusters.


Figure 2 – K-means silhouette analysis with 3 clusters.

To evaluate how well K-means clustering performed compared to the actual labels or y-values of the wine origin dataset, Adjusted Rand Index (ARI) is used. ARI measures the similarity between the predicted clusters and the true labels. ARI ranges from -1 to 1, with higher values indicating better agreement between the predicted clusters and the true labels. This k-means algorithm, using 3 clusters, returns an ARI of 0.740 which indicates a strong agreement between the predicted clusters and the true labels. Lastly, the distribution of data per cluster is plotted below in figure 3.


Figure 3 – K-means distribution of data, using 3 clusters.

*2) Expectation Maximization(EM)*

The EM algorithm determined the optimal number of clusters to be 3 as well. For this implementation of EM, BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) are used to evaluate performance of EM on the wine origin dataset with 4 different covariance types (spherical, diag, full, and tied). Both plots display an elbow, like the k-means inertia plot, where the number of clusters is 3. After 3 clusters, both AIC and BIC decrease at a reduced rate when compared to 1 and 2 clusters.

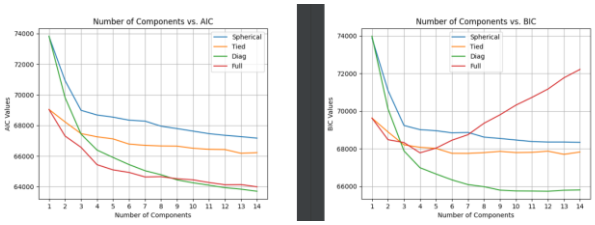


Figure 4 – EM. AIC and BIC scores using varying covariance.

Referencing the silhouette score, again the optimal number of clusters is determined to be 3 and the silhouette score is 0.1616. This can be seen in the plot below. The image on the left shows the highest average silhouette score where the number of clusters is 3, and the image on the right supports the fact that the average silhouette score for each cluster is above the overall average silhouette score, which is again noted by the vertical red line.

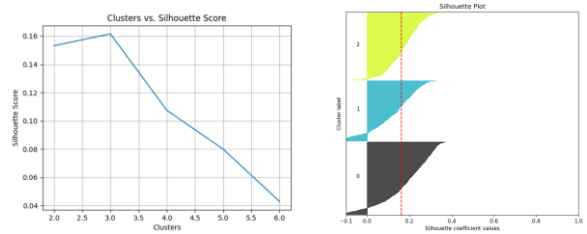


Figure 5 – EM. Silhouette score analysis

Evaluating this EM implementation, Adjusted Rand Index (ARI) is used again and returns a score of 0.625. This score indicates a moderate degree of agreement between the predicted clusters and the true clusters, considering chance agreement. As such, the ARI value of 0.625 suggests that this EM algorithm was able to capture some, but not all, of the underlying structure in the data. Moreover, the fact that is ARI is lower than the ARI score returned by k-means, indicates the distribution of data per cluster is different. This can be seen in figure 6 below.


Figure 6- EM. Distribution of data, using 3 clusters.

*B. Breast Cancer Dataset*

*1) K-means*

Using the breast cancer dataset, k-means determined the optimal number of clusters to be 2. Because this is a binary classification dataset, having 2 clusters corresponding to diagnosis is not surprising. As seen on the right in figure 7 below, the elbow is present (circled in red) where the number of clusters is 2. To emphasize this, silhouette score is examined for a range of clusters between 2 and 6. Visible below on the right, the average silhouette score is at its highest, 0.345, where the number of clusters is 2, and decreases for all other cluster values.

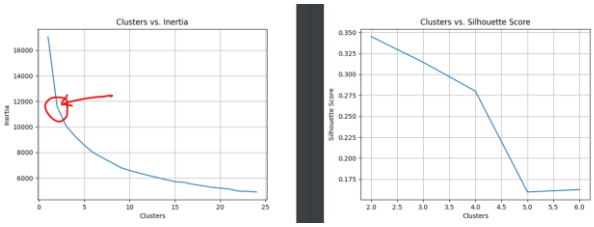


Figure 7 – Clustering Algorithm. EM inertia and silhouette scores.

Diving deeper into the silhouette score with 2 clusters, it's immediately noticeable that the grey cluster holds a lot more points/data. This is indicated by the difference in height when compared to the green cluster. Moreover, from the right plot, it can be concluded that cohesion of the grey points is higher than the green points. This also explains the higher silhouette coefficient values for the grey cluster.

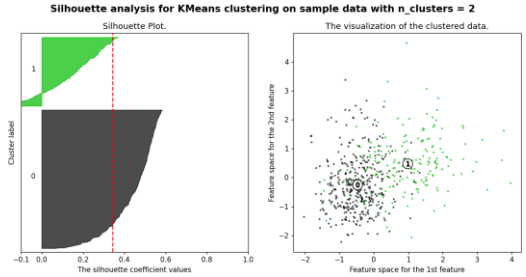


Figure 8 - K-means silhouette analysis with 2 clusters.

Supporting the observations above, the plot below shows a significantly higher number of points belonging to a cluster. When looking back at the original dataset, there is a 63% / 37% split in y-values. As such, the distribution of data per cluster looks appropriate.


Figure 9 - K-means distribution of data, using 2 clusters.

Evaluating the performance of this EM implementation, Adjusted Rand Index (ARI) is used. The EM algorithm, using 2 clusters on the breast cancer dataset, generates an ARI score of 0.511. Being that ARI returns a value between -1 and 1, where -1 indicates

completely dissimilar clustering's, 0 indicates random clustering, and 1 indicates identical clustering's, an ARI value of 0.511 suggests that this k-means clustering solution is better than random, but there is still room for improvement.

*2) Expectation Maximization(EM)*

Like the k-means clustering optimization, EM determined the optimal number of clusters to be 2. Again, this is not surprising given the dataset being used is a binary classification dataset. Starting with a review of the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), both plots below in figure 10 show the rate at which the different covariance variable values decreases to form an elbow where the number of clusters is 2.

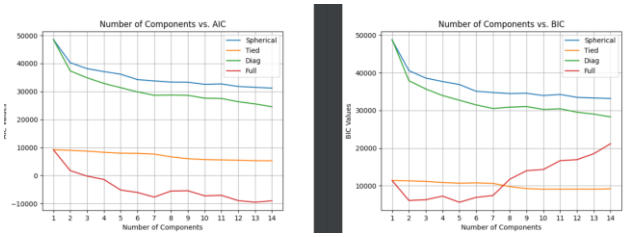


Figure 10 – EM. AIC and BIC scores using varying covariance.

Further analysis of average silhouette scores supports the observation above and can be seen in figure 11 below. On the left, the average silhouette score is the greatest at 0.315 where the number of clusters is 2. The silhouette plot on the right shows both clusters having silhouette scores above the average as well. Moreover, like the silhouette plot produced after clustering using k-means, one cluster holds significantly more data than the other.

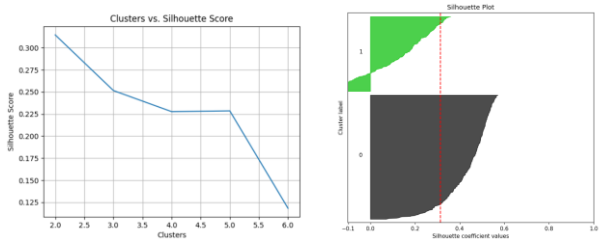


Figure 11 - EM. Silhouette score analysis

Regarding the amount of data in each cluster, figure 12 below visualizes the distribution of samples in each cluster. Like the silhouette plot, we can see about 150 more samples in the cluster on the left. Compared to the same distribution of data produced by k-means clustering, there is less data in cluster 0 and more data in cluster 1. As a result, it is likely that that EM is doing a better job making estimates that maximize the likelihood function. To determine this, ARI is calculated. For the breast cancer dataset with 2 clusters, an ARI value of 0.774 is returned. This is much better than the ARI returned using k-means and indicates a reasonable level of agreement between the two clusters.


Figure 12 - K-means distribution of data, using 2 clusters.

## II. PART 2: DIMENSIONALITY REDUCTION

Dimensionality reduction is a technique used to reduce the number of variables or features in a dataset while retaining as much useful information as possible. The goal is to simplify the data without losing too much information and to improve the efficiency and effectiveness of subsequent analyses. 4 techniques used in dimensionality reduction are analyzed below.

### *A. Wine Origin Dataset*

*1) Principal Component Analysis (PCA)*

To determine the amount of reduction that would be most effective on this dataset, explained variance (EV) is used. The EV, which a measure of how much information each principal component contains, shows the cumulative EV on the left and the EV of each component on the right in figure 13 below. As expected, all 13 features of this dataset account for 100% of the EV cumulatively. On the right, the first 8 components account for a total of about 80.6% of the data. For this reason, moving forward, 8 components will be used to verify PCA's effectiveness.

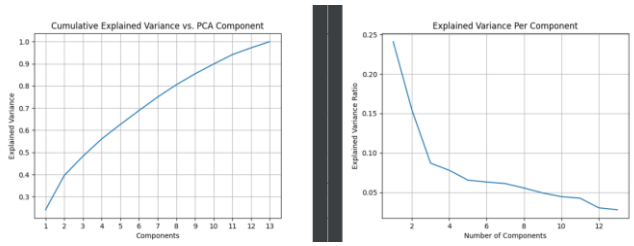


Figure 13 – PCA. Explained variance.

Looking at the distribution of data where the 13 original features are reduced to 8 principal components results in the following plots.

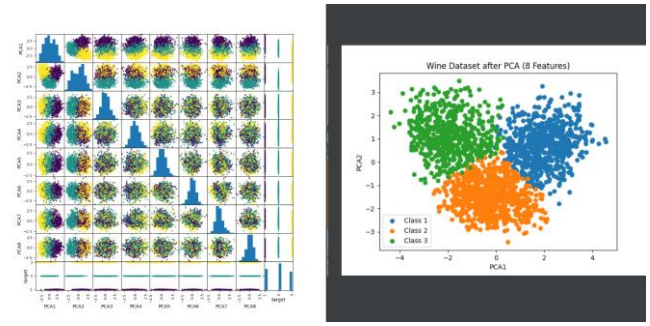


Figure 14 – PCA – Data distribution using 8 principal components.

To the right in figure 14 above, three very distinctive classes exist. Moreover, the scatter plot pictured on the left shows the pairwise relationships between each of the 8 principal components. Most important, the diagonal of

the scatter plot displays a very good distribution of each principal component.

*2) Independent Component Analysis (ICA)*

Using ICA to determine the optimal number of independent components is done by measuring kurtosis and skewness of each component. Plotted below in figure 15, the left image shows normalized mean kurtosis values, and the right image shows skewness values for each component of the dataset.

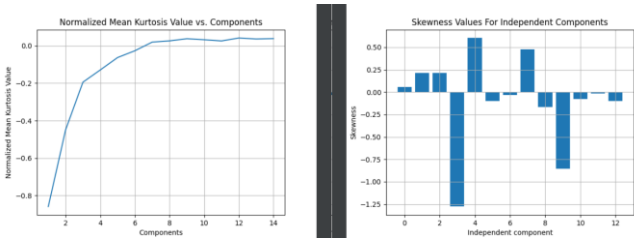


Figure 15 – ICA. Kurtosis and Skewness

By using kurtosis to measure the non-Gaussianity of the independent components, one can derive the components with the highest kurtosis values that are more likely to capture interesting and useful information in the data. In the plot above, the kurtosis values plateau where the number of components is equal to 8. As such, referencing the skewness values, it can be seen that the individual component values from 1-8 cover a broad range of values, both positive and negative. This is ideal because components with high absolute values of skewness, either positive or negative, are likely to be more useful for capturing interesting and important information in the data.

Finally, looking at the 3D scatter plot in figure 16 below, the results from using 8 components with ICA are observed. Note that all the data is grouped tightly and there is no indication of outliers.

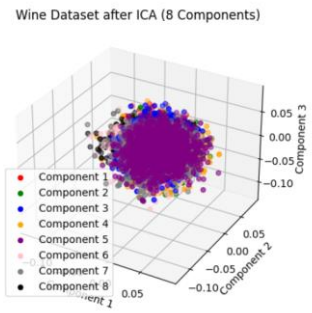


Figure 16 – ICA. 3D Scatter plot using 8 components.

*3) Randomized Projections (RP)*

Dimensionality reduction using RP involves projecting high-dimensional data onto a lower-dimensional space in a randomized manner, while preserving some of the structure and relationships among the variables. To determine the optimal number of components, reconstruction error and variance are measured. Both can be seen plotted in figure 17 below. On the left is reconstruction error, and on the right is reconstruction variance. Both are measured with respect to the number of components.

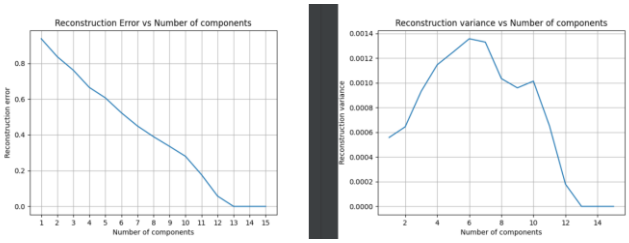


Figure 17- RP. Reconstruction error and variance.

The plot showing reconstruction error does not give much information aside from how the reconstruction error decreases as the number of components increase. However, using the reconstruction variance shows the amount of variance preserved by the reduced-dimensional space as a function of the number of components. Where the number of components equals 6, the reconstruction variance value is peaked. Hence, the optimal number of components suggested by RP is 6. With this information, referencing the reconstruction error plot where the number of components equals 6, it can be determined that the reduced-dimensional representation is only able to capture 50% of the variance in the original data. In other words, half of the information in the original data has been lost in the projection.

Lastly, plotted below in figure 18 is a 3D scatter plot of the 6 components. Again, all the components appear to be grouped closely with very few points being considered outliers.

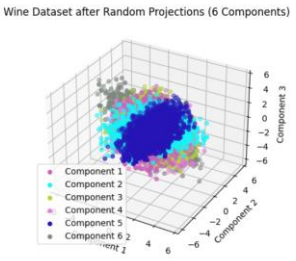


Figure 18 – RP. 3D scatter plot using 6 components.

*4) ExtraTreeClassifier (ETC)*

The last dimensionality reduction algorithm to explore is ETC. ETC works by constructing a large number of decision trees and combining their predictions to make a final prediction. In the context of dimensionality reduction, ETC is used to select the most important features in the original dataset and discard the rest. Plotting the importance of each feature is depicted below in figure 19 below. Referencing the figure on the left shows there are 4 features that stand out as having the greatest importance. The plot on the right shows the reduced-dimensional representation obtained using ETC using the top 5 most important features.

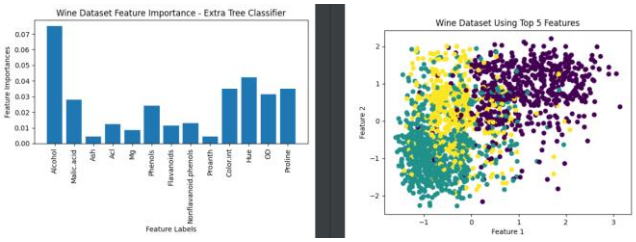


Figure 19 – ETC. Feature Importance

## B. *Breast Cancer Dataset*

### 1) *Principal Component Analysis (PCA)*

Like the PCA analysis on the wine dataset, cumulative explained variance and explained variance per component are used. As expected, the 30 components that make up this dataset account for 100% cumulatively. Plotted below on the right, the per component explained variance shows that the first 4 components account for almost 80% of the explained variance, and that the explained variance decreases very slowly after 10 components. As such, 7 components appear to be the optimal number of components as suggested by this PCA algorithm.

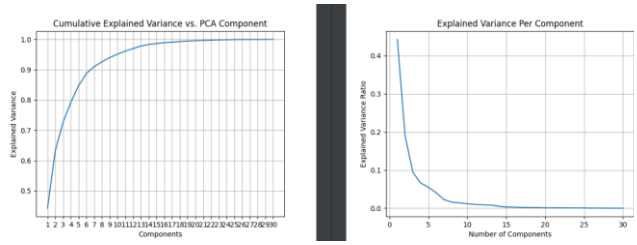


Figure 20 - PCA. Explained variance.

Examinging the scatter plots below, using 7 of the 30 original features captures the data well enough to effectively classify this dataset. Furthermore, 2 distinct classes exist, and very little overlap is present after classifying the data.

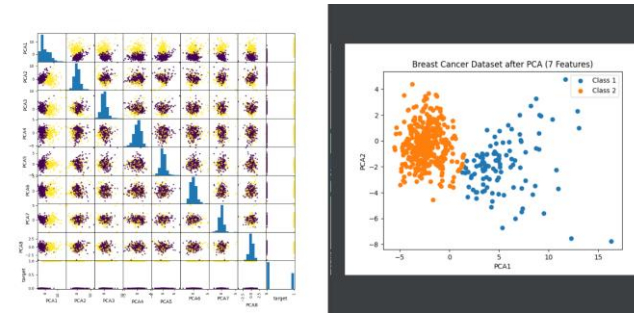


Figure 21 - PCA – Data distribution using 7 principal components.

### 2) *Independent Component Analysis (ICA)*

ICA performs very different on the breast cancer dataset compared to Principal Component Analysis. Using kurtosis and skewness, ICA determines the optimal number of components to be 12. As can be seen in the plot below, the image on the left shows the highest kurtosis value, 1.730 with 12 components. Knowing the best kurtosis value is 3, a value of 1.730 suggests that the distribution of the components are closer to a normal distribution than to a distribution with a higher kurtosis.

The plot on the right, displaying skewness values of individual components, shows that the distribution is mostly symmetric. This is because most of the individual component skewness values are close to 0. Where positive skewness values are seen is an indication the distribution is skewed to the right, while components with a high negative skewness value indicate that the distribution is skewed to the left. While there aren't many, a few of the components demonstrate this well.

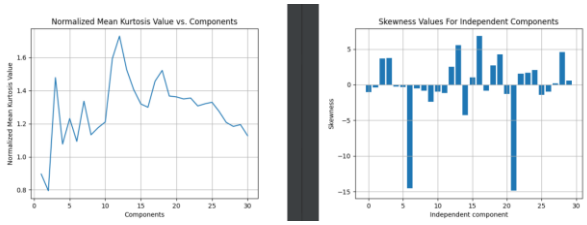


Figure 22 - ICA. Kurtosis and Skewness

Finally, plotting the breast cancer dataset after dimensionality reduction results in figure 23 below. By examining the plot, we can see that the data points for the two classes (malignant and benign) are generally separated from each other in 3D space, with some degree of overlap between the two classes. This indicates that the independent components extracted by ICA are able to capture some of the underlying differences between the two classes in the breast cancer dataset.

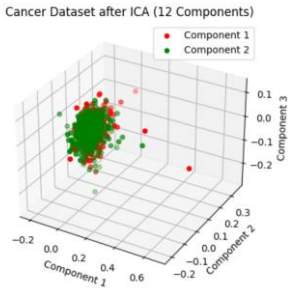


Figure 23 - ICA. 3D Scatter plot using 8 components.

### 3) *Randomized Projections (RP)*

Using RP for dimensionality reduction also suggests a different number of components compared to PCA and ICA. To determine the optimal number of components, reconstruction error and reconstruction variance are reviewed using the entire breast cancer dataset. Figure 24 below shows that the highest reconstruction variance is reached where the number of components is 17. Hence, where the number of components is 17, the greatest proportion of variance in the original data is preserved in the reduced-dimensional data after reconstruction. On the left, we can see that the reconstruction error for 17 components is about 0.5. This value means that the reduced-dimensional data, obtained after applying RP to the original data, has a high level of distortion compared to the original data. If this turns out to be an issue, it might be necessary to select a higher number of components where the reconstruction variance is also lower.

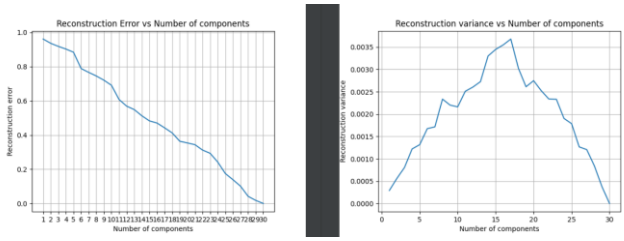

Figure 24 - RP. Reconstruction error and variance.

Plotting the classes after dimension reduction results in the figure below. It can be seen that the data points are distributed somewhat evenly throughout the plot, with a mix of blue and red points in each region of the plot. This suggests that RP has done a reasonable job of separating the data points into the two classes, but there may still be some overlap between the classes.

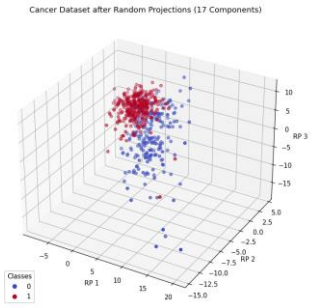

Figure 25 - 3D scatter plot using 17 components.

*4) ExtraTreeClassifier (ETC)*

By visually inspecting the plot below, the separation between the two classes and the structure of the dataset can be evaluated after reducing the number of dimensions using the image on the right. On the left, each component and its corresponding feature importance value is displayed. Based on the plot of the breast cancer dataset using 10 features, it can be seen that RP is able to capture the most relevant information for the classification task, while reducing the dimensionality of the dataset to a manageable number of features.

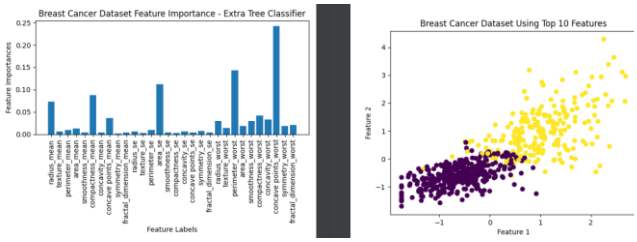

Figure 26 – ETC. Feature Importance

## III. CLUSTERING WITH DIMENSIONALITY REDUCED ALGORITHMS

### A. Wine Origin Dataset

*1) K-means Observations*

Using the k-mean clustering algorithm after dimensionality reduction using PCA, ICA, RP, and ETC exposes strengths and weaknesses for each of the dimensional reducing algorithms. The results of this are plotted in figure 27 below.

Immediately noticeable is the inertia score of ICA versus the other algorithms (top left). In general, a lower inertia score is better, as it indicates that the data points are more tightly packed within their assigned clusters. As such, it is not unexpected that the adjusted rand index (ARI) score (bottom) is also higher than the other algorithms. This is a clear indication that the clustering produced by k-means using ICA are both compact and well-defined, and also meaningful and similar to a truth set or reference clustering. Unfortunately, ICA suffers from a low silhouette score indicating there may be some overlapping or ambiguity in clusters.

ETC is also a strong performing algorithm with a low inertia score and high ARI score. ETC also has a silhouette score, 0.33, which is more than twice as much as ICA, 0.15.

The highest inertia score belongs to RP, signifying that the points within the k-means clusters using RP are spread out over a larger area, which means that the clusters are less compact and well-defined. As a result, k-means clustering using RP did not effectively group the points in a meaningful way which results in a lower ARI score.

Lastly, PCA is a mix consisting of a high inertia score, a better than average silhouette score, and a strong ARI score. This combination suggests that the k-means clusters are less compact and well-defined, but the overall clustering solution is still meaningful and similar to the truth set reference clustering.

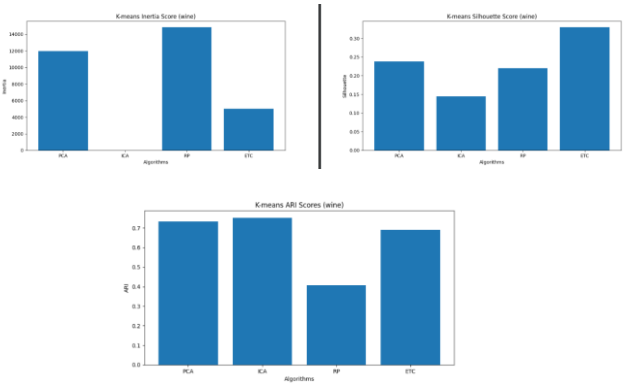

Figure 27 - K-means clustering Inertia (top left), Silhouette (top right), and ARI (bottom) score using reduced wine dataset.

TABLE I. K-MEANS WINE STATS

| K-Means Wine Stats | | | |
|---|---|---|---|
| Algorithm | Inertia | Silhouette | ARI |
| PCA | 11962.32 | 0.24 | 0.73 |
| ICA | 5.51 | 0.14 | 0.75 |
| RP | 13042.96 | 0.24 | 0.41 |
| ETC | 5016.07 | 0.32 | 0.68 |

*2) Expectation Maximization (EM) Observations*

EM clustering using PCA, ICA, RP, and ETC is evaluated similarly to k-means, however, included in this evaluation is AIC and BIC scores. AIC and BIC scores are plotted in figure 28 below. Figure 29 below shows EM silhouette scores and ARI scores.

Both AIC and BIC are statistical measures that balance the goodness of fit of a model with the complexity of the model, and, as a result, a lower AIC or BIC value indicates a better fit and suggests that the model is more likely to accurately represent the underlying structure of the data. Figure 28 clearly shows ETC having the lowest AIC and BIC scores, followed by RP, PCA, and ICA.

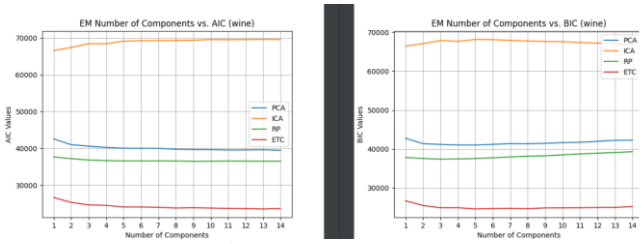


Figure 28 - EM. AIC (Left) and BIC (Right) scores using reduced wine dataset.

In addition to the lowest AIC and BIC scores, ETC has the greatest ARI and silhouette scores as seen in figure 29 below. Hence, EM clustering, using the dataset reduced by EM, is well-separated and meaningful. Conversely, ICA has the highest AIC and BIC scores, and lowest silhouette and ARI scores. Not only is EM using the wine dataset reduced using ICA the worst performing, but it is also an indication that this EM algorithm implementation has not effectively identified meaningful clusters in the data.

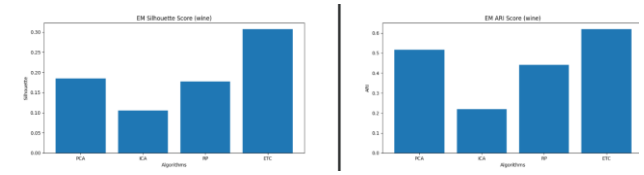


Figure 29 - EM. Silhouette (Left) and ARI (Right) scores using reduced wine dataset.

TABLE II. EM WINE STATS

| EM Wine Stats | | |
|---|---|---|
| Algorithm | Silhouette | ARI |
| PCA | 0.18 | 0.51 |
| ICA | 0.12 | 0.21 |
| RP | 0.17 | 0.44 |
| ETC | 0.31 | 0.61 |

*3) Comparison of The Two*

Both K-means and EM did an effectives job clustering the data in the wine dataset. However, k-means did a slightly better job. All k-means clustering implementations using the 4 dimensionality reduction algorithms performed better than EM. Interestingly, while ETC performed the best, k-means and EM both performed very similarly regarding the silhouette and ARI scores.

Also interesting is the fact that while k-means clustering using the dataset reduced by RP performed the worst, the EM implementation using the dataset reduced by RP was significantly better. This suggests that EM was able to capture the underlying structure of the data more effectively than k-means, even after the data was reduced using RP. This may be because EM is more flexible and can model more complex data distributions than k-means, particularly when the clusters are non-spherical or have overlapping data points.

## B. *Breast Canser Dataset*

*1) K-means Observations*

Similar to the k-mean clustering analysis using dimensionality reduction and the wine origin dataset, ICA and ETC clearly have the lowest inertia score and highest ARI score. Of the two, ICA and ETC, ETC also has the highest silhouette score. In fact, ETC has the highest silhouette score of all k-means clustering implementations. Hence, the k-means clustering using ETC has the most points assigned to the correct cluster and is well-separated from the points in other clusters. Overall, this is the most desirable outcome for a clustering analysis because it indicates that the EM clustering solution is accurate, well-separated, and effectively captures the structure of the data.

On the other hand, RP did not perform so well. Having the highest inertia score, mediocre silhouette score, and lowest ARI score indicates that the k-means clustering using the dataset reduced by RP is well-separated, however, having a high inertia score indicates that the clusters themselves may not be tightly packed together, meaning that the distance between the points within each cluster may be relatively large. This suggests that this clustering solution may not be as effective in capturing the underlying structure of the data.

Overall, like the k-means solution using the wine origin dataset, ETC performed the best. Using the breast cancer dataset, as well as the wine dataset, ETC was the most effective as far a feature selection and reducing the dimensionality of the datasets because of the way it builds multiple decision trees on subsets of the features and random subsets of the data.

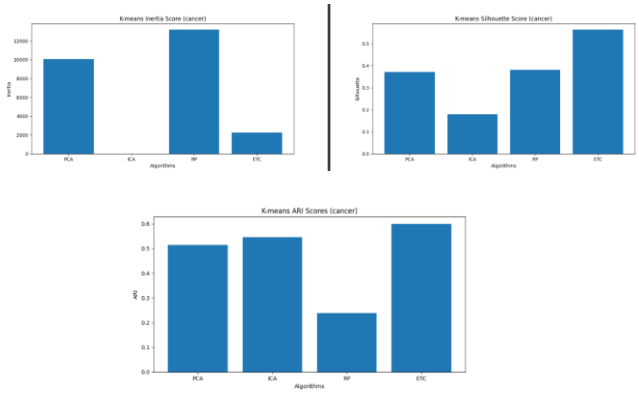


Figure 30 -K-means clustering Inertia (top left), Silhouette (top right), and ARI (bottom) score using reduced cancer dataset.

TABLE III. K-MEANS CANCER STATS

| K-Means Cancer Stats | | | |
|---|---|---|---|
| Algorithm | Inertia | Silhouette | ARI |
| PCA | 10062.94 | 0.37 | 0.52 |
| ICA | 6.27 | 0.18 | 0.55 |
| RP | 13182.56 | 0.38 | 0.24 |

| | | | |
|---|---|---|---|
| ETC | 2252.86 | 0.56 | 0.60 |

*2) Expectation Maximization (EM) Observations*

The reduced breast cancer dataset performed similarly on EM compared to EM using the reduced wine dataset, as well as compared to the feature selection algorithms used with k-means and the reduced breast cancer dataset.

Referencing figure 31 below, ETC clearly has the lowest AIC and BIC score for the entire range of components tested. Interestingly, the BIC score for ETC plateaus between 3 and 7 components, and subsequently begins to decrease for the remaining components. For this to occur, it is possible that the breast cancer dataset has some complex underlying structure that requires a larger number of components to model accurately. As a result, a decreasing BIC score is observed. Also interesting is that, on the left of figure 31, the AIC score for ETC increases. Because AIC and BIC differ in the way they penalize for model complexity, seeing a decreasing BIC score and increasing AIC score is not an impossible outcome.

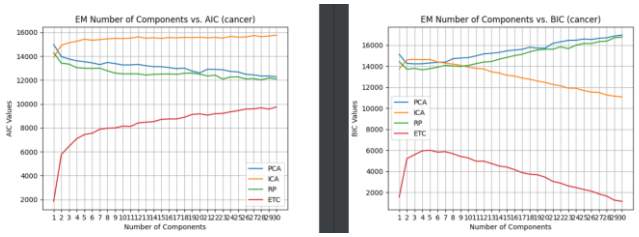


Figure 31 - EM. AIC (Left) and BIC (Right) scores using reduced breast cancer dataset.

Figure 32 below supports the observation above that EM using the reduced dataset produced by ETC is the superior solution. Referencing the table provided, ETC has a silhouette score that is almost double of the other solution, as well as the highest ARI score.

ICA also has a high ARI score, however, the silhouette score of 0.13 suggests that the quality of clustering is poor. ICA also does not have strong AIC or BIC score, suggesting it is likely to be overfitting the data and that the model may be too complex. PCA and RP bridge the gap between ETC and ICA having middle of the road silhouette scores, but RP has a significantly lower ARI score. While RP's ARI score is not terrible, the score of 0.32 suggests that EM clustering is doing a decent job at identifying clusters in the data, but there may be room for improvement.

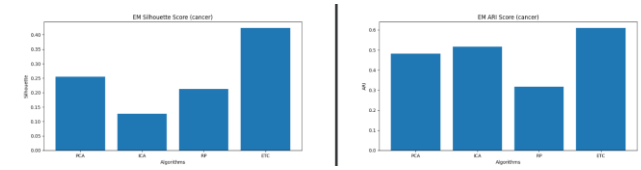


Figure 32 - EM. Silhouette (Left) and ARI (Right) scores using reduced breast cancer dataset.

TABLE IV. EM CANCER STATS

| EM Cancer Stats | | |
|---|---|---|
| Algorithm | Silhouette | ARI |
| PCA | 0.26 | 0.48 |
| ICA | 0.13 | 0.52 |
| RP | 0.21 | 0.32 |
| ETC | 0.42 | 0.61 |

*3) Comparison of The Two*

Because the breast cancer dataset has more than twice as many features as the wine origin dataset, the breast cancer dataset is more complex making it more difficult to extract meaningful information and cluster the data effectively. As such, the scores produced by k-means and EM clustering using the reduced datasets are generally better for the wine origin dataset.

Specific to the breast cancer dataset, k-means did an overall better job of clustering. It is most likely the case that some subset of data within the breast cancer dataset has well-separated and roughly spherical clusters enabling k-means to outperform EM.

## IV. NEURAL NETWORK WITH DIMENTIONALITY REDUCED DATASET

Figure 33 below is a visual representation of the neural network (NN) learning curves using the dimensionally reduced wine origin dataset reduced from 13 features to 7. Intriguingly, the learning curve generated from the NN using ICA shows that it suffers from high bias and low variance for the first 15% of training examples used. This can be seen in the lowest training and test scores as compared to the other learning curves. However, ICA rapidly improves in performance as it learns from the data and, once 20% of the training examples have been used, ICA performs quite well with no signs of high variance or bias. Additionally, the learning curve generated by ICA shows the training and test scores converging once 40% of the training examples are used. At this point, ICA has learned all it can from the available data and further training shows minimal improvement to its performance.

Referencing the other 4 algorithms (NN with the full-sized dataset, PCA, RP, and ETC), it can be observed that all of them have a gap between training and testing scores as compared to ICA, hence higher variance. Additionally, the NN using PCA, RP, and ETC all have testing scores that are increasing in the last 20% of the training example. Because of this, these three algorithms might benefit from more data.

Overall, all 4 NN solutions using the dimensionally reduced dataset perform well compared to the NN solution using the full-sized wine origin dataset. Moreover, all the NN solutions show little overfitting of the data.

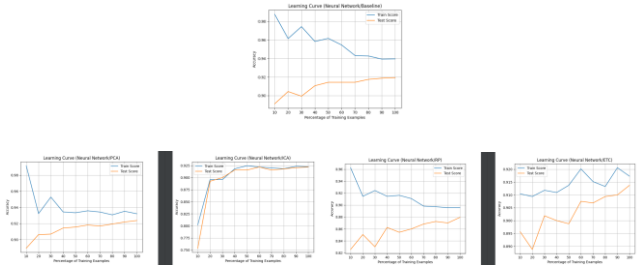

Figure 33 – NN Learning Curve. Baseline – top, Bottom, right to left – PCA, ICA, RP, ETC

Lastly is a look at the training time and accuracy, found in figure 34 below. Unsurprisingly, the NN using the full-sized dataset has the highest accuracy. This is most likely due to the reduced dimension dataset having lost some important information during the dimensionality reduction process. As a result, all the algorithms using this reduced dataset suffer from a lower accuracy score, even if only by a couple percent. This is supported by the fact that the NN using RP has the lowest accuracy overall, about 5% lower than the NN baseline solution using the full dataset. While PCA, ICA, ETC, and RP are designed to identify the most informative features, they may not always capture all the relevant information in the data. As a result, the accuracy is close to, but not quite as good as, where the full dataset it used. Overall, it's likely that the 4 feature selection techniques may have discarded important information that is necessary for accurate predictions when reducing the dataset from 13 features to 7.

Looking at the training times, its clear that there is much more variation than there is with accuracy. The NN baseline solution and NN using ICA stand alone having taken about 400 seconds to train. The NN using ETC and PCA took about half as long, around 200 seconds. The fact that PCA, RP, and ETC took less time to train is expected because running dimensionality reduction on the wine origin dataset simplifies the problem for the NN and makes it easier to learn the relevant patterns in the data. As such, this leads to faster training times.

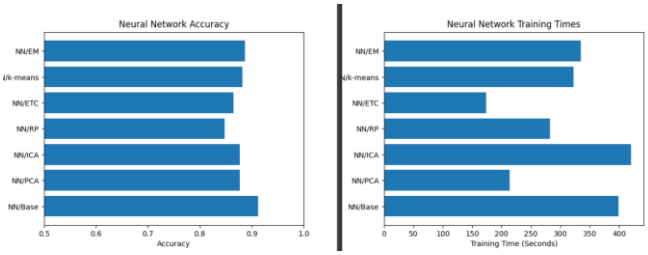


Figure 34 – NN Accuracy, left. NN Training Time, right.

## V. Neural Network with Clustering

The two neural networks (NN) implemented using clustering perform remarkably similarly. When previously discussed in part 1, the k-means clustering silhouette score was 0.17 and Expectation Maximization (EM) about 0.16. However, k-means had and ARI score of 0.74 while EM's ARI score was below 0.70. Using the NN, however, it's probable that both k-means and EM were able to identify the same underlying patterns in the data leading them to perform almost equally. This may be because the wine dataset is a relatively simple dataset with a small number of features, only 13.

Figure 34 below shows each implementations respective learning curve and shows just how similarly these two clustering algorithms performed.

Finally, using the knowledge that k-means and EM are effective ways to reduce the size of the dataset and thus make training the NN faster, its unsurprising that both the k-means and EM solutions had faster training times when compared with the baseline NN. Moreover, the only notable difference when comparing k-means with EM is that k-means has a slightly lower accuracy and training time. The lower training time of k-means is because EM is a more computationally intensive algorithm. However, the lower accuracy is because EM is a more flexible algorithm that allows for more complex modeling of the data, while K-means is a more rigid algorithm that assumes that the clusters are spherical and of equal size. As a result, EM can model the data as a mixture of probability distributions, which allows it to capture more complex patterns in the data and, as seen in figure 34, produce a slightly higher accuracy.

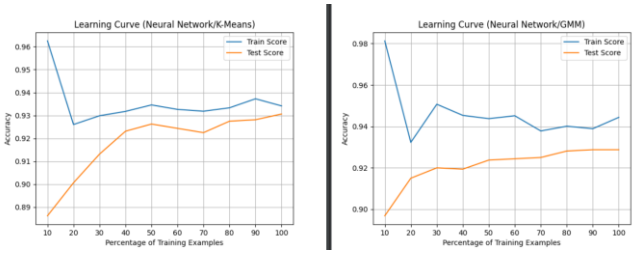


Figure 35 – Neural Network with Clustering. K-means, left. EM, right.