Franklin O. Onyekwum (fonyekwum3)
CS 7641 – Machine Learning (Spring 2017)

Project 3 – Unsupervised Learning and Dimensionality Reduction

Datasets Analysed and Selection Rationale: I analysed the 2 datasets from Assignment 1, namely:

1. **Abalone**[1]: This dataset was obtained from the UCI machine-learning repository. The objective is to predict the age of abalones (a kind of sea snail), based on various physical measurements. The dataset has 8 attributes (categorical, real and integer), contains 4177 instances, and its output class has 29 members. To allow for faster cluster-to-label comparisons and computations, I defined a new class label named **Ring Class** with 5 possible values (A to E), and mapped this to the original class label (**Rings)**, such that **Rings 1 – 5 = Ring Class A**, **Rings 6 – 8 = Ring Class B**, **Rings 9 – 11 = Ring Class C**, **Rings 12 – 17 = Ring Class D** and **Rings 18 – 24 = Ring Class E**. The Abalone dataset is interesting for the same reasons as previously described. It is deceptively simple yet notoriously difficult to classify, and I would like to see if clustering and the various other techniques explored in this project lead to any improvements.

2. **Wine Quality**[2]: This is also from the UCI repository. The goal here is to classify wine qualities, based on various physiochemical tests and properties of the wine. The dataset has 12 real attributes, with 4898 instances and an output class containing 11 members (2 of which have no member instances). The Wine quality dataset is generally well behaved, and I am interested in seeing how much improvement will be gained by applying the techniques being investigated, in contrast with the Abalone dataset.

Implementation Methodology: All the experiments described in this report were carried out using **WEKA** version 3.8.1.

Clustering: Clustering is an unsupervised learning methodology that attempts to group together or cluster similar instances in a given dataset. There are various clustering techniques, with each using a different measure of similarity. Please note that to make this a valid unsupervised learning experiment, the class attributes for each dataset was removed/ignored when performing clustering. The clustering techniques investigated are K-Means Analysis (KM) and Expectation Maximization (EM).

- **K-Means Analysis**: K-Means works by taking the specified number of cluster (K), randomly picking K points as the initial centres of the K clusters, and assigning each point in the dataset to the cluster closest to it, based on a distance metric (Euclidean, Manhattan etc.). The algorithm then iteratively finds the centres of the new clusters, resets the centres as required and reassigns points to clusters, until no more improvements are made. In these experiments, I decided to use Euclidean distance as the distance metric, as it is very well understood and would not introduce any extra complexity.

  ***KM and the Abalone Dataset*:** The experiment here involved using the add-on **XMeans** classifier in WEKA, setting K to increasing numbers from 2 to 15, and measuring the output Bayesian Information Criteria (BIC) and Distortion values for each run. The resulting clusters where then compared with the actual class label, and the clustering accuracy was also recorded. My understanding is that BIC is supposed to reduce as the number of clusters is decreased. However, for a number of my experiments I found that BIC increased with K. For such scenarios, I used the distortion relationship instead, which fortunately behaved as it should (i.e., decrease as K increases). **Figure 1** below shows the results of these experiments.

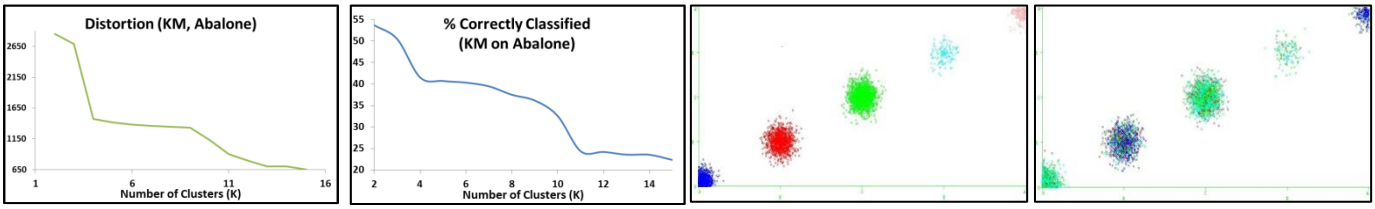

Figure 1: KM on Abalone. Left-to-Right: (a)K vs Distortion; (b)K vs Classification Accuracy; (c)Clusters with Original Labels; (d)KM Clusters vs Labels

Using the elbow rule on **Figure 1a**, the optimal number of clusters (K) was set at 4. **Figure 1b** supports this selection, as it also has an elbow at K = 4, and levels out from K = 4 onwards. **Figure 1b** also shows that the accuracy at K = 4 is **41.50%**. **Figures 1c** and **1d** compare the clustering of the data using the original labels to the clustering arrived at by KM with K = 4. In **Figure 1c**, the clusters are perfect and only points/instances with the same colour are in each cluster, as expected. **Figure 1d** has 4 different colours, with each colour representing a KM cluster. It shows that KM does a fair job, as it seems to have identified

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

3 of the original class labels (the centre 3 clusters are predominantly dark blue, green and light blue). It doesn't do so great on the other 2 clusters however. Also, the optimal number of cluster chosen by KM (K = 4) is different from the actual number of class labels (5).

***KM and the Wine Dataset*:** The same experiment was conducted with the wine dataset experiment, and **Figure 2** below shows the results. BIC also increased with K here, so I used the distortion measure instead.

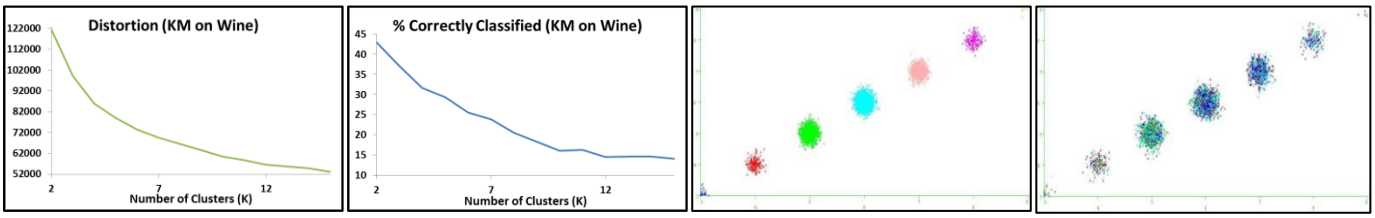

Figure 2: KM on Wine Dataset. Left-to-Right: (a)K vs Distortion; (b)K vs Classification Accuracy; (c)Clusters with Original Labels; (d)KM Clusters vs Labels

Using the elbow rule on **Figure 2a**, K is somewhere between 4 and 6, so I settled on K = 5. **Figure 2b** again shows a downward trend in accuracy as K is increased. However, from **Figure 2b** the clustering accuracy at this number of clusters is **29.30%**, which is rather poor. **Figures 2c** and **2d** are similar to those previously described. Here, K-Means does not do as good a job at clustering the data as shown in **Figure 2d**. Again, the optimal number of clusters here (K = 5) is less than the actual number of class labels (7).

- **Expectation Maximization (EM)**: EM performs clustering using probability distributions. It takes the specified number of clusters (K) and outputs a probability (maximum likelihood) that there are actually K clusters. As such, a higher likelihood value is preferable.

***EM and the Abalone Dataset*:** Similar experiments as described above were carried out with the **EM** classifier in WEKA, with the Log Likelihood for each number of clusters being recorded instead. **Figure 3** below shows the results of clustering the Abalone dataset using EM.

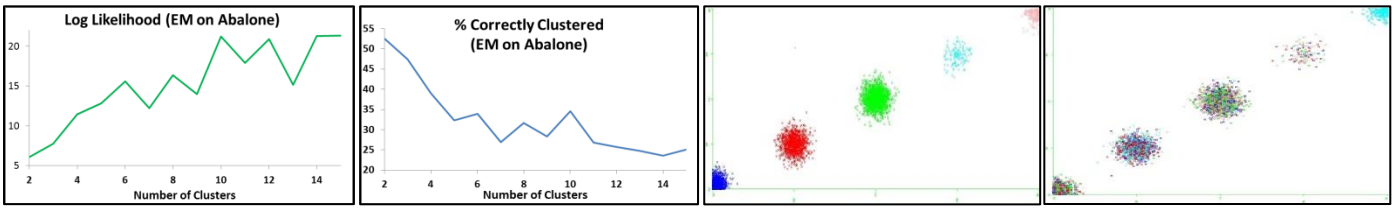

Figure 3: EM on Abalone. Left-to-Right: (a)K vs Log Likelihood; (b)K vs Classification Accuracy; (c)Clusters with Original Labels; (d)EM Clusters vs Labels

It was a bit more difficult here to use the elbow rule to select the best number of clusters, due to the zig-zagging nature of **Figure 3a.** However, I eventually settled on 6 clusters, after observing that the downward trend in classification accuracy in **Figure 3b** seems to level out at 6 clusters, with a classification accuracy of **33.90%**. Using 6 clusters and comparing with the true class label clusters from **Figure 3c**, it appears that EM only accurately found one cluster, from **Figure 3d** (the uppermost cluster is predominantly light blue).

***EM and the Wine Dataset*: Figure 4** below shows the results of clustering the Wine dataset using EM. Similar to the scenario with Abalone, I was unable to find a clear elbow in **Figure 4a**, and eventually settled on 6 based on **Figure 4b** (at which point classification accuracy is a disappointing **34.40%**). The log likelihood curve here is much smoother than that of the Abalone dataset. This suggests that the Wine dataset has attributes that are more deterministically linked than those of the Abalone dataset. Also, EM is able to identify 2 of the clusters here (from **Figure 4d**).

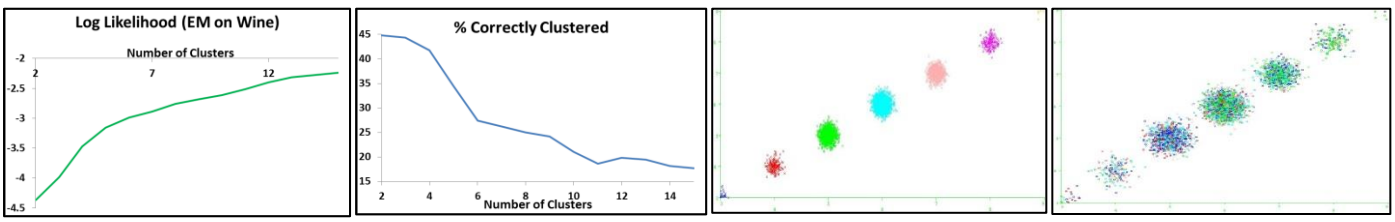

Figure 4: EM on Wine. Left-to-Right: (a)K vs Log Likelihood; (b)K vs Classification Accuracy; (c)Clusters with Original Labels; (d)EM Clusters vs Labels

Dimensionality Reduction: As implied by its name, dimensionality reductions involves attempting to reduce the dimensions or number of attributes in a data set, by isolating attributes that do not encode any/much information in the dataset, and eliminating them. 4 dimensionality reduction techniques are explored in this section, namely **Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections (RP)** and **Information Gain (IG)**.

- **Principal Component Analysis (PCA):** PCA involves attempting to find a linear combination of all the attributes in the dataset that has the most variance. Maximizing variance is important because by finding a combination that maximizes the variance, we are accounting for the variation in the dataset. This combination is known as the first Principal Component (PC), and this routine can be repeated for the projections of the points onto the first PC to find the second PC and so on. The PCA algorithm uses matrices to perform these operations, and returns a list of Eigen values sorted in decreasing order of variance, and Eigen vectors. The higher the Eigen value of a projected attribute, the larger the amount of variance that the attribute is responsible for.

  ***PCA and the Abalone Dataset*:** PCA was run on the Abalone dataset and the 5 resulting principal components and their corresponding Eigen values were recorded. A scree plot was then generated, as shown in **Figure 5**.

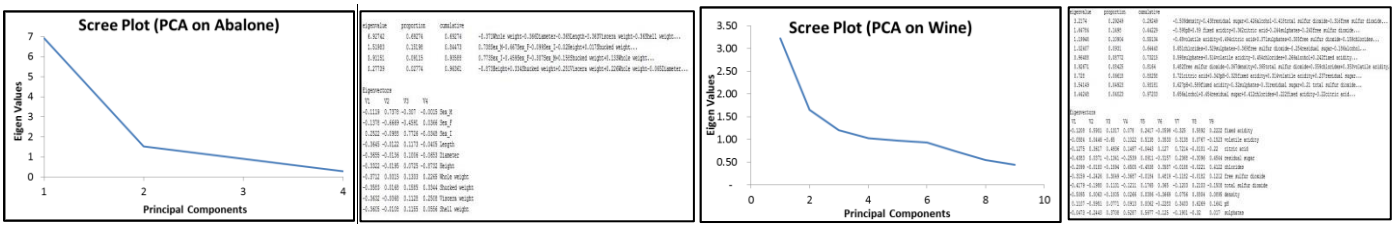

Figure 5: PCA on Abalone and Wine Datasets. Left-to-Right: (a)Abalone Scree Plot; (b)Abalone Eigen Values sorted in Decreasing Order of Variance and Abalone Eigen Vectors; (c)Wine Scree Plot; (d)Wine Eigen Values sorted in Decreasing Order of Variance and Wine Eigen Vectors

From the scree plot in **Figure 5a**, and using the elbow rule, it is clear that the first 2 principal components account for majority of the variation in the dataset. Also, line 3 in **Figure 5b** shows that these 2 components account for about **85%** of the variation in the dataset.

***PCA and the Wine Dataset*:** The same procedure was carried out for the Wine dataset, and the data obtained was used to generate the scree plot shown **Figure 5c**. Using the elbow method, the first 6 principal components were selected here (**Figures 5c**), as they collectively represent **81%** of the variance in the dataset (**Figure 5d**, line 7).
These 2 new reduced datasets were then clustered using K-Means and EM. The following sections discuss the results.

***K-Means and the PCA-Reduced Abalone Dataset*:** KM clustering was performed on the PCA-reduced Abalone dataset, and the resulting data is plotted in **Figure 6** below. The BIC plot here (**Figure 6a**) acted as expected, and in conjunction with the elbow rule, suggested 3 clusters (K = 3). The clustering accuracy here also shows a significant improvement compared to that of the unmodified dataset (**46.70%** vs. 41.50%). Also, with K = 3 the KM successfully found 3 clusters (i.e. the 3 middle clusters in **Figure 6d**), which is quite similar to the result for the unreduced Abalone dataset (see **Figure 1d**).

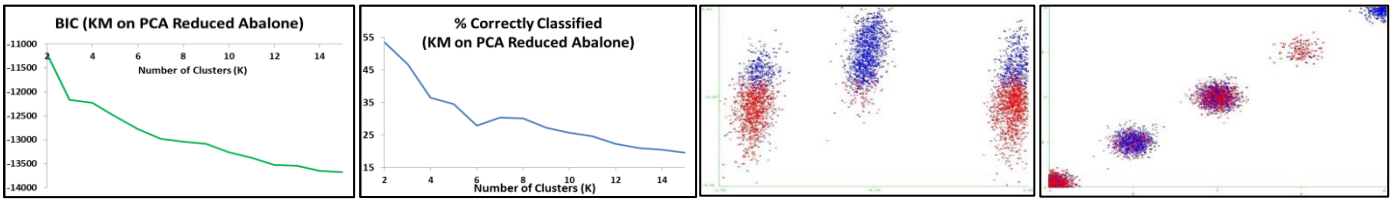

Figure 6: KM on PCA-Reduced Abalone Dataset. LTR: (a)K vs. BIC; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***K-Means and the PCA-Reduced Wine Dataset*:** **K**M clustering was also performed on the PCA-reduced Wine dataset, and the results are shown in **Figure 7** below. The BIC plot here (**Figure 7a**) also reduced as K increases, and using the elbow rule along with **Figure 7b**, K was set at 5. This represents a reduction in clustering accuracy when compared to the selected number of clusters for the unmodified dataset (**26.30%** vs. 29.30%). **Figure 7c** shows the result of clustering the reduced Wine dataset at K=5. Unfortunately however, KM is not as successful at clustering the reduced wine dataset, as shown in **Figure 7d.**

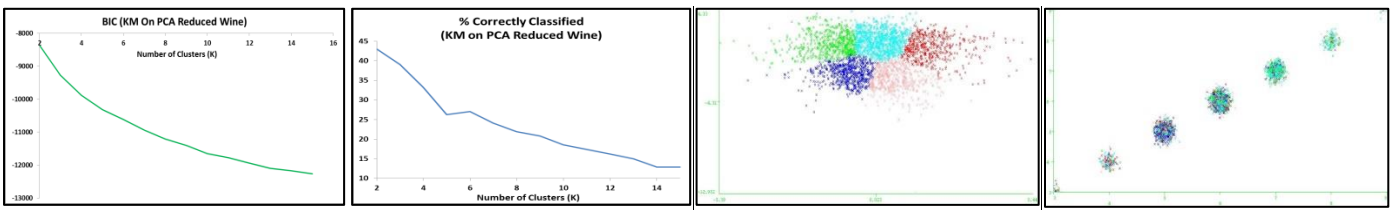

Figure 7: KM on PCA-Reduced Wine Dataset. LTR: (a)K vs BIC; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***EM and the PCA-Reduced Abalone Dataset*: Figure 8** below shows the results of running EM clustering on the PCA-reduced Abalone dataset. The Log Likelihood plot (**Figure 8a**) shows a clear elbow at K = 3, and this value was used for subsequent analysis. Also, at K = 3 the classification accuracy of EM is at its highest (**46.04%**), which is higher than than the accuracy on the unadjusted dataset (33.9%). **Figure 8c** shows the resulting 3 clusters found by EM, while **Figure 8d** shows the EM clusters compared to the true class labels.

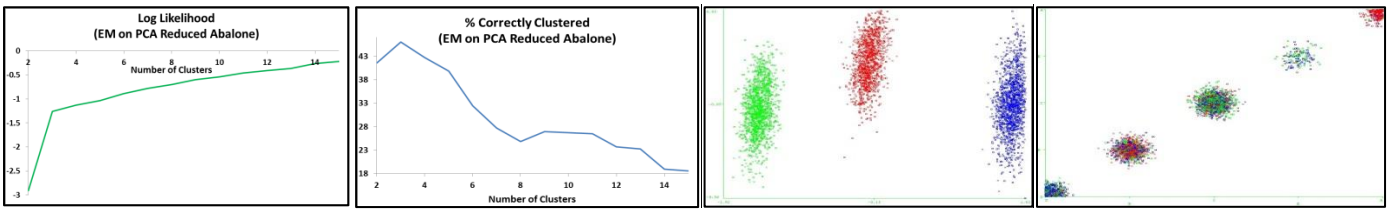

Figure 8: EM on PCA-Reduced Abalone Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***EM and the PCA-Reduced Wine Dataset*: Figure 9** below shows the results of running EM clustering on the PCA-reduced Wine dataset. The number of clusters selected is 5 (from **Figure 9a**). Also, at K = 5 the classification accuracy of EM is lower at **33.03%** than it was for the unmodified data set (34.40%). **Figure 8c** shows the resulting 5 clusters found by EM, while **Figure 8d** shows the EM clusters compared to the true class labels.

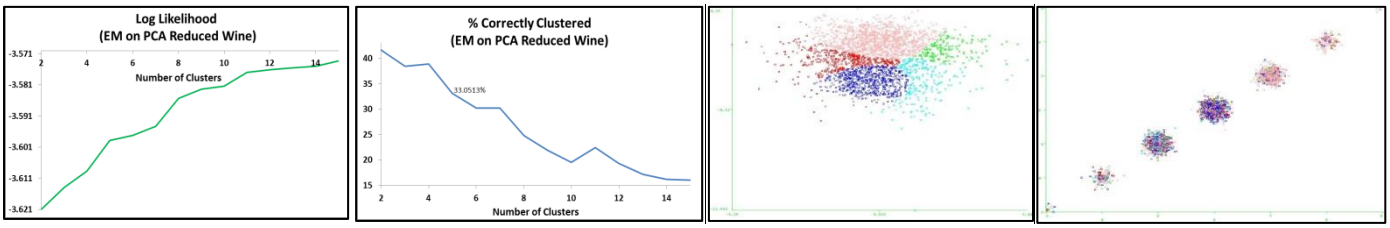

Figure 9: EM on PCA-Reduced Wine Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

- **Independent Component Analysis (PCA):** I see ICA as an attempt to "reverse engineer" the central limit theory. The central limit theory essentially states that all things being equal, most distributions tend towards the normal distribution. Given a dataset, each attribute can be thought of as representing a different distribution, and the entire dataset can be thought of as a combined distribution. This combined distribution, according to the central limit theory, will tend towards the normal distribution. ICA attempts to extract the individual/"independent' distributions within this combined distribution by looking for a set of component distributions which are as unlike a normal distribution as possible. One measure of the degree of dissimilarity of a distribution from the normal distribution is known as **Kurtosis**. The normal distribution has a kurtosis value of 3.

  ***ICA and the Abalone Dataset*:** ICA was run on the Abalone dataset in WEKA, using **FastICA** from the **StudentFilters** WEKA add-on, and the resulting transformed data was copied to MS Excel, where the kurtosis of each attribute was calculated. Note that 3 was added to the kurtosis values obtained from Excel, as Excel calculates **Excess Kurtosis** (i.e., amount of kurtosis over the normal distribution kurtosis value). The kurtosis values were then arranged in descending order, and plotted against the corresponding attribute, as shown in **Figure 10** below.

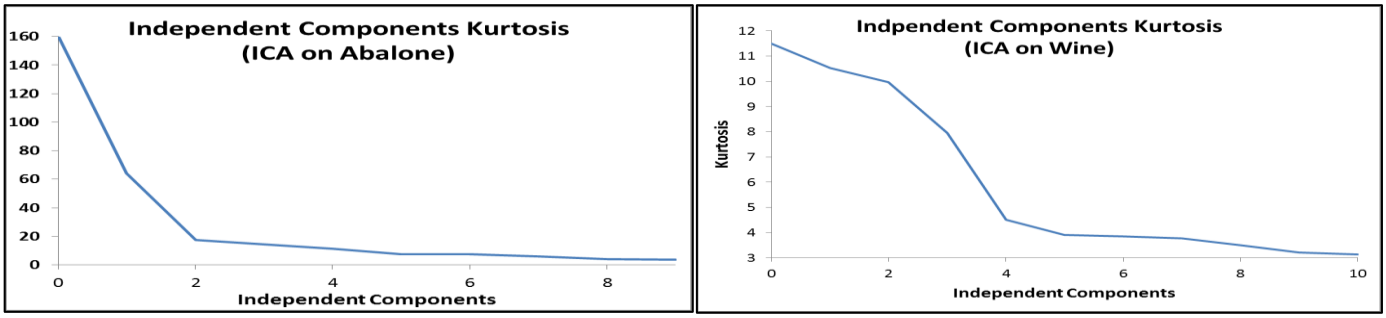

Figure 10: ICA on Abalone and Wine Datasets. LTR: (a)Kurtosis values of ICA Attributes for Abalone dataset in descending order; (b)Kurtosis values of ICA Attributes for Wine dataset in descending order

From Figure 10a, the elbow is clearly at Independent Component 2. Therefore the 3 attributes with the highest kurtosis values (i.e. Components 0 to 2) were used for subsequent analysis.

***ICA and the Wine Dataset*:** ICA was also run on the Wine dataset, and the kurtosis data was calculated and is plotted in **Figure 10b** above. Here, the elbow is at 4, and the corresponding 5 new independent components are used in subsequent analysis.

These 2 new datasets were then clustered using K-Means and EM, as previously described.

***K-Means and the ICA-Reduced Abalone Dataset*: Figure 11** below shows the results of performing KM on the ICA-reduced Abalone dataset. Analysing the BIC plot (**Figure 11a**) in conjunction with the classification accuracy (**Figure 11b**), with the elbow rule suggested 6 clusters (K = 6). The clustering accuracy of **29.60%** at this point is significantly worse than that obtained on the unmodified data (41.50%). **Figures 11c** and **11d** also show that the clusters formed are not as distinct as desired, and there is a good amount of overlapping of clusters.

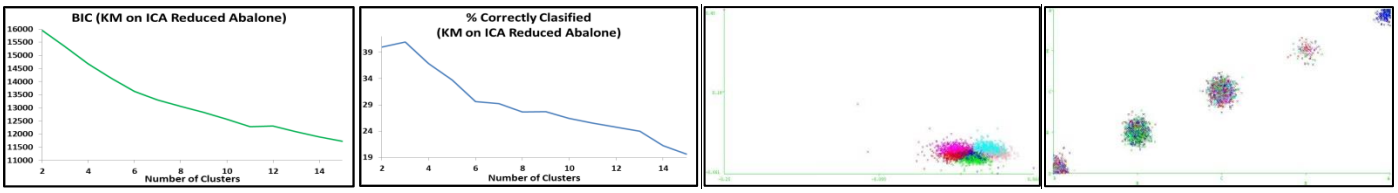

Figure 11: KM on ICA-Reduced Abalone Dataset. LTR: (a)K vs BIC; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***K-Means and the ICA-Reduced Wine Dataset*:** In **Figure 12** below shows the results of performing KM on the ICA-reduced Wine dataset. The BIC plot (**Figure 12a**) shows a distinct elbow at 4 clusters, which is the selected number of clusters for this analysis. Furthermore, the classification accuracy from **Figure 12b** at 4 clusters is **30.25%**, which Is more or less the same as the accuracy for the unmodified Wine dataset (29.30%). The clusters obtained here are also not very distinct and exhibit a lot of overlap (**Figures 12a** and **12b)**.

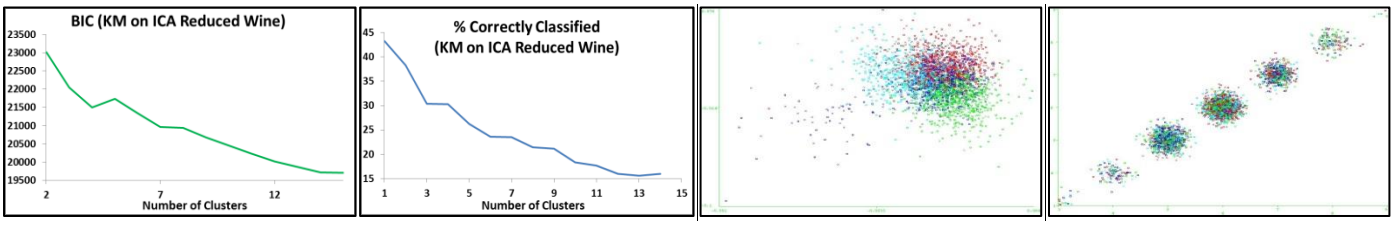

Figure 12: KM on ICA-Reduced Wine Dataset. LTR: (a)K vs BIC; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***EM and the ICA-Reduced Abalone Dataset*:** From **Figure 13a** below, the optimal number of clusters is selected as 4. Also, at 4 clusters the classification accuracy in **Figure 13b** is **44.30%**, which is a marked improvement from 33.90% for the unmodified dataset. This improvement is visible in the cluster diagram in **Figure 13c**, where 3 distinct and well separated clusters can be seen (the fourth cluster contains very few points and is scattered around the other 3 clusters).

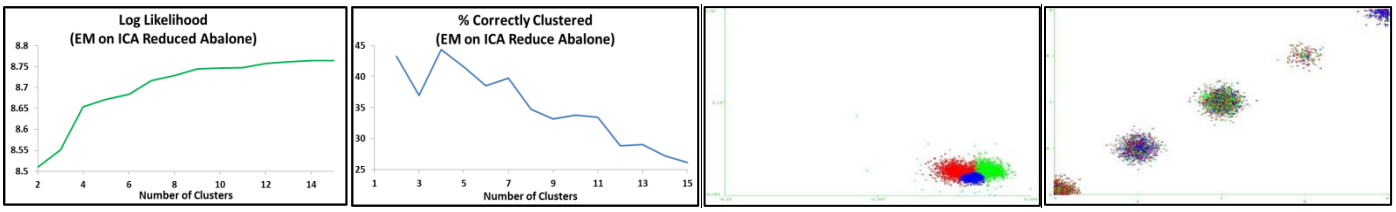

Figure 13: EM on ICA-Reduced Abalone Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)EM Clusters vs Labels

***EM and the ICA-Reduced Wine Dataset*: Figure 14a** below shows that the optimal number of clusters is around 7. With this number of clusters, the classification accuracy in **Figure 14b** is **27.02%**, which is much worse than the 34.40% recorded with the unmodified dataset. **Figure 13c** shows this reduced performance, as the identified clusters are overlapping and not very distinct. The cluster to label plot (**Figure 13d**) is also rather poor, as all the clusters are predominantly purple.

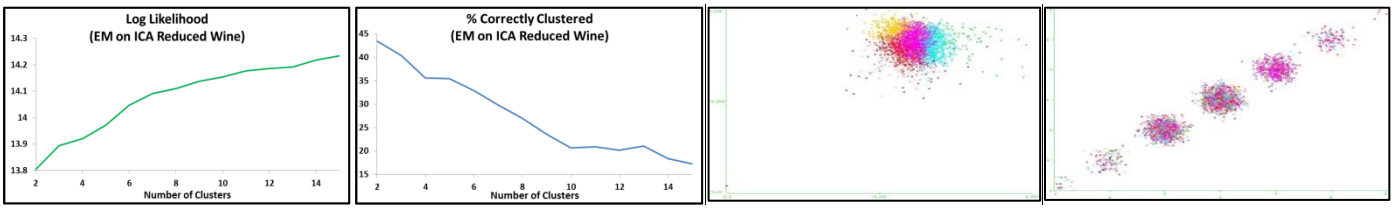

Figure 13: EM on ICA-Reduced Wine Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)EM Clusters vs Labels

- **Randomized Projections (RP):** Compared to the previous dimension reduction algorithms, RP is fairly straight forward. It simply attempts to randomly project the attributes (or dimensions) in the dataset to a lower dimension space (i.e., fewer attributes), in the hope that these random projections somehow capture a good portion of the variance in the original dataset without using as many attributes. RP has a reputation of working surprisingly well, considering that the projections are random and are not based on any underlying logic.

  ***RP and the Abalone Dataset*:** RP was run on the Abalone dataset in WEKA, using **RandomProjection** filter. However, no measure of the fitness of the newly projected points is available for RP in WEKA. I therefore decided to run RP 3 times, with a different seed used for each run (10, 42 and 95). For each set of randomly projected datasets, I then trained a Decision Tree (with 10-set CV), using varying number of attributes from the datasets, and recorded the prediction accuracy and RMSE of the decision tree. I chose to use a decision tree because it is a fairly simple and well understood classifier, and therefore should introduce minimal extra complexity to the analysis. **Figure 14** below shows the results of the experiments.

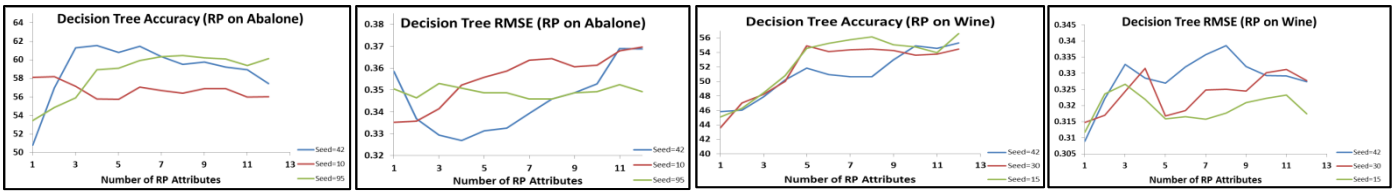

Figure 14: RP on Abalone & Wine Datasets. LTR: (a)Decision Tree Accuracy vs Attribute Count for Abalone; (b) Decision Tree RMSE vs Attribute Count for Abalone; (c) Decision Tree Accuracy vs Attribute Count for Wine; (d)) Decision Tree RMSE vs Attribute Count for Wine

Figures 14a and **14b** show that a fair bit of variance in decision tree accuracy exists for the different seed values used. I eventually settled on using the first 3 components in the new RP dataset, as accuracy was highest for the seed value of 42 at this number of attributes, and I would be using seed 42 (the default seed setting for RP in WEKA) going forward.

***RP and the Wine Dataset*:** RP was also run on the Wine dataset. This time however, based on my experience with the Abalone dataset, I decided to use a set of seed values closer to one another (15, 30 and 42). The datasets were then used to train the same Decision Tree as described above, and **Figures 14c** and **14d** above display the results obtained. The accuracy curves for all 3 seed values are much more similar to one other, and it was fairly easy to pick an attribute count value of 5 here.
The 2 new datasets were then clustered using K-Means and EM, as described below.

***K-Means and the RP-Reduced Abalone Dataset*:** The distortion plot in **Figure 15a** below shows a clear elbow at K = 6 clusters, and this was used in subsequent clustering analysis for this section. Also, **Figure 15b** shows that the clustering accuracy at K = 6 was **41.97%,** which is about the same as the original datasets accuracy value (41.50%). Furthermore, with K = 6, KM successfully found 6 fairly distinct clusters (**Figure 15c**) clusters (i.e. the 3 middle clusters in **Figure 6d**), which is quite similar to the result for the unreduced Abalone dataset (see **Figure 1d**).

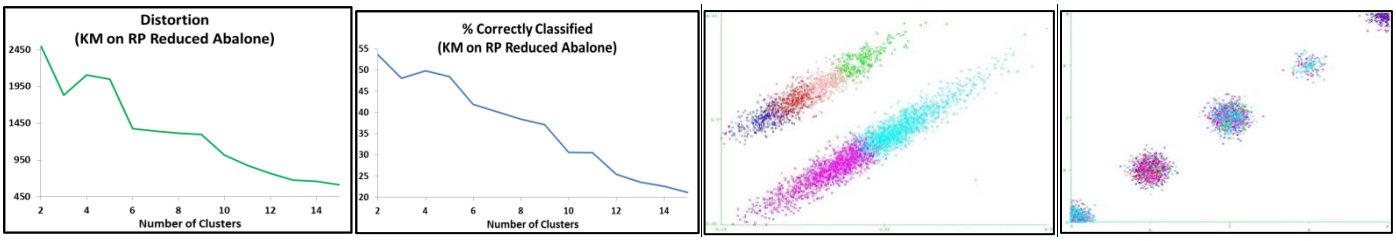

Figure 15: KM on RP-Reduced Abalone. LTR: (a)K vs Distortion; (b)K vs Classification Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***K-Means and the RP-Reduced Wine Dataset*:** The elbow in the distortion plot in **Figure 16a** below is somewhere between 4 and 6, so I picked 5 as the number of clusters. From **Figure 16b**, classification accuracy at 5 clusters is **28.18%**, again almost the same as the accuracy with the original dataset (29.30%). The clusters also seem to be fairly well defined (**Figure 15c**).

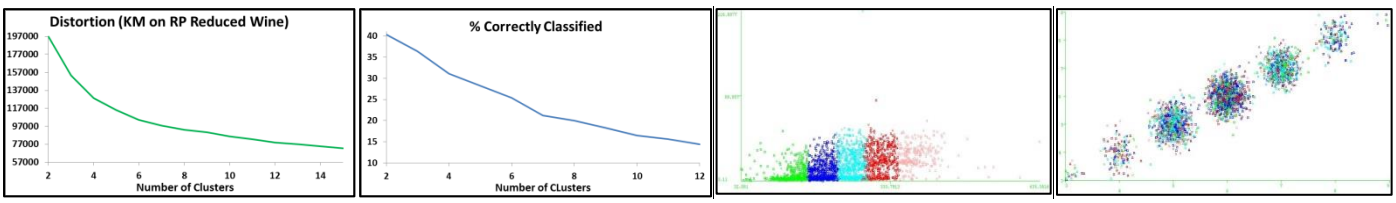

Figure 16: KM on RP-Reduced Wine. LTR: (a)K vs Distortion; (b)K vs Classification Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***EM and the RP-Reduced Abalone Dataset*:** From **Figure 17a** below, the optimal number of clusters is selected as 7. Clustering accuracy at 7 clusters is **32.13%** from **Figure 17b** (similar to 33.90% for the unreduced dataset), and the clusters are fairly distinct in **Figure 17c**.

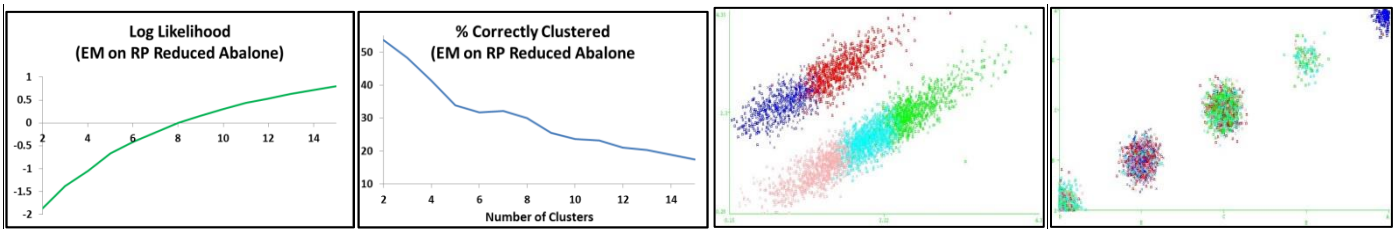

Figure 17: EM on RP-Reduced Abalone Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)EM Clusters vs Labels

***EM and the RP-Reduced Wine Dataset*:** In **Figure 18a** below, the elbow location/number of clusters is selected as 4. Also, clustering accuracy at 4 clusters is **30.00%** from **Figure 18b** (less than 34.40% for the unreduced dataset), and the 4 clusters are still quite distinct in **Figures 18c** and **18d**.

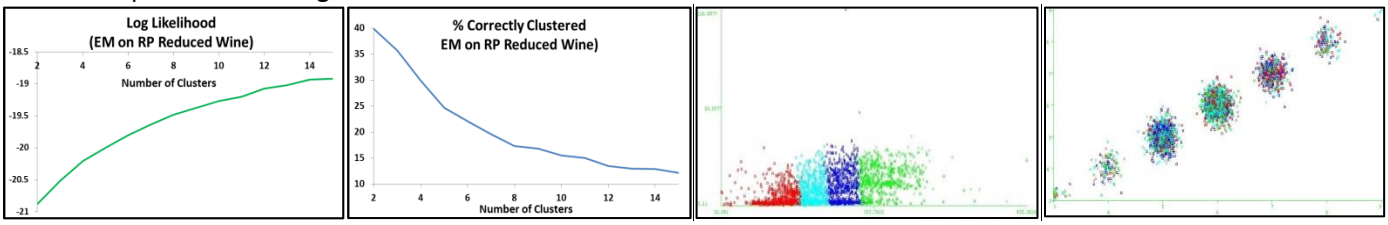

Figure 18: EM on RP-Reduced Wine Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)EM Clusters vs Labels

- **Information Gain (IG):** This is another fairly simple attribute selector that calculates the information gain attributable to each of the attributes in the dataset, and then selects the attributes with the highest information gain.

***IG and the Abalone Dataset*:** IG was run on the Abalone dataset in WEKA, using **InfoGain** WEKA attribute selector. The information gain values for the attributes where then sorted in descending order and plotted, as shown in **Figure 19** below.

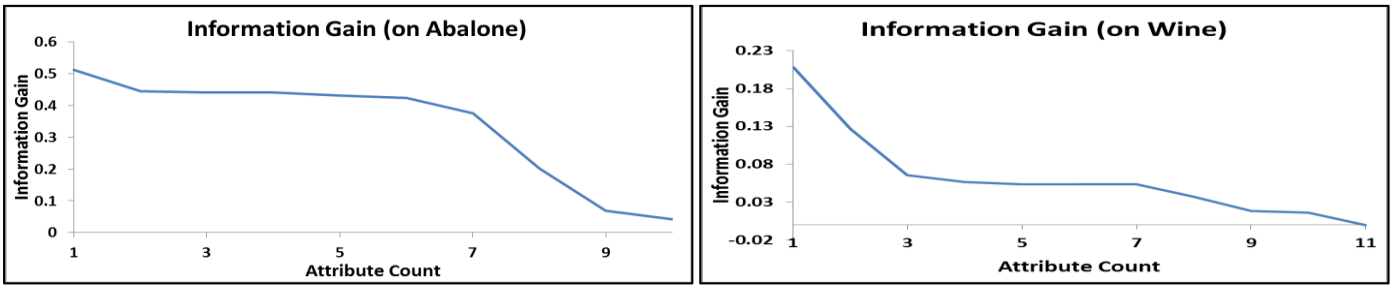

Figure 19: Information Gain on Abalone and Wine Datasets.

From **Figure 19a**, it is clear that the elbow is at 2, so the 2 attributes with highest information gain were used going forward, while all the other attributes were eliminated.

***IG and the Wine Dataset*:** The same procedure described above was performed on the Wine dataset, and the resulting information gain values are shown in **Figure 19b**. Using the same logic as previously described, the 3 with the highest information gain were used for subsequent analysis on the Wine dataset.
The new datasets were then clustered using K-Means and EM, as described below.

***K-Means and the IG-Reduced Abalone Dataset*:** Using the BIC plot in **Figure 20a** below, 10 clusters was selected as the optimal number of clusters. Also, at 10 clusters, **Figure 20b** indicates that the clustering accuracy is around **25.00%** a sharp drop from the 41.50% observed in the original dataset. However, the 10 clusters appear to be fairly well defined in **Figure 20c**, especially considering the large number of clusters.

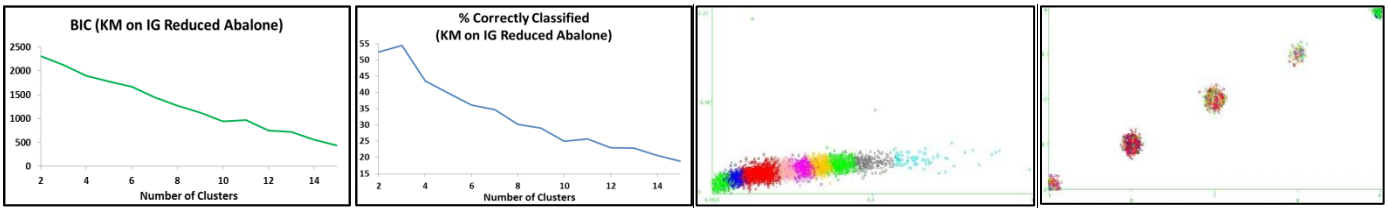

Figure 20: KM on IG-Reduced Abalone Dataset. LTR: (a)K vs BIC; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***K-Means and the IG-Reduced Wine Dataset*:** From the distortion plot in **Figure 21a** below, 3 clusters was selected as the optimal number of clusters. Also, at 3 clusters, **Figure 21b** shows a clustering accuracy of **46.35%** a marked increase from the 29.30% observed in the original dataset. Furthermore, the 3 clusters are well defined in **Figure 21c**.

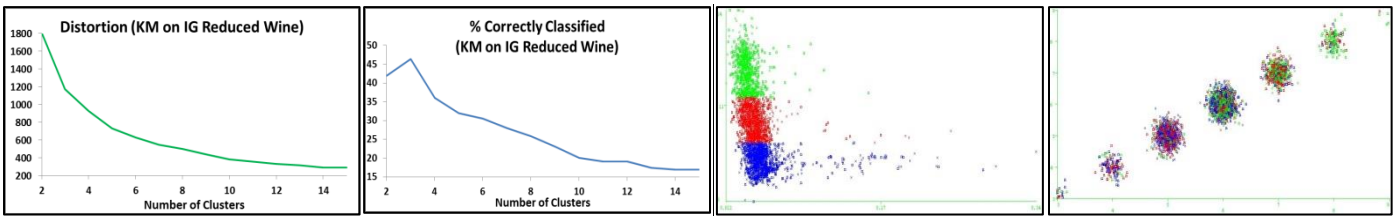

Figure 21: KM on IG-Reduced Wine Dataset. LTR: (a)K vs Distortion; (b)K vs Accuracy; (c)Sample Clusters Image; (d)KM Clusters vs Labels

***EM and the IG-Reduced Abalone Dataset*:** From **Figure 22a** below, the optimal number of clusters is selected as 5 after consulting **Figure 22b**. **Figure 22b** also shows that clustering accuracy at 5 clusters is **46.44%** (which is much better than 33.90% for the unreduced dataset), and the 3 clusters are fairly distinct in **Figure 22c** (although not as much so in **Figure 22d**).

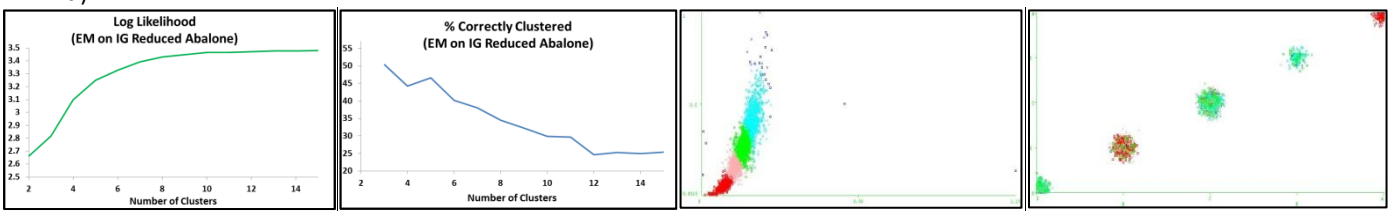

Figure 22: EM on IG-Reduced Abalone Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)EM Clusters vs Labels

***EM and the IG-Reduced Wine Dataset*: Figure 23a** below shows that the optimal number of clusters using the elbow rule is 4 (also supported by **Figure 23b**). **Figure 23b** also shows that clustering accuracy at 5 clusters is **44.34%** (which is again much better than 34.40% for the unreduced dataset). Furthermore, the 3 clusters are fairly distinct in **Figure 23c** (although again not as much as in **Figure 23d**).

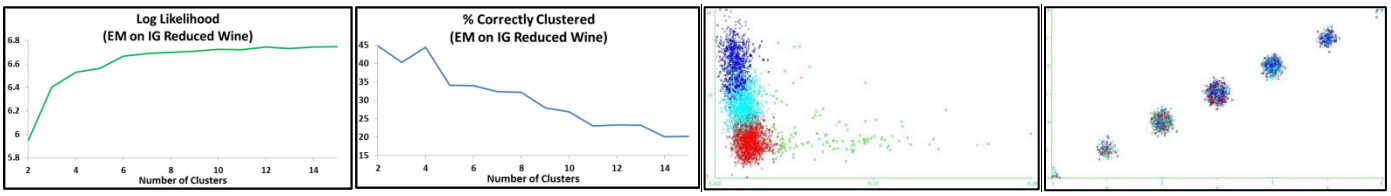

Figure 23: EM on IG-Reduced Wine Dataset. LTR: (a)K vs Log Likelihood; (b)K vs Accuracy; (c)Sample Clusters Image; (d)EM Clusters vs Labels

Dimensionality Reduction, Clustering and Artificial Neural Network Analysis: The final requirement in the project brief is to train a neural network on one of the datasets above, using both the dimensionally reduced version of the dataset, and also to train a neural network using the dimensionally reduced dataset with the clustering data added as an extra attribute. I chose to perform this experiment using the Wine dataset.

To ensure that the process of dimensionally reducing the dataset does not lead to a biased data when training the neural network, I divided the Wine dataset into a training set (70%) and a test set (30%). It was this training set data that was then used to perform all the dimensionality reduction experiments detailed above. I then trained the Artificial Neural Network (ANN) from **Assignment 1** (i.e. Momentum = 0.4, Learning Rate = 0.1, Number of Hidden Layers = t and Number of Iterations = 500), on 3 versions of the different dimensionally reduced datasets (i.e. reduced Wine datasets alone, reduced Wine datasets with K-Means cluster labels, and reduced Wine datasets with EM cluster labels). The ANN was trained with increasing amounts of training data (i.e. from 10% to 100%), and then tested with the previously unseen test data set. The classification accuracy of each ANN was then recorded and, **Figure 24** below shows the plots of these accuracies against the portion of training data used.

From **Figure 24a**, we can see that without including the clusteing information, both the RP and IG reduced data essentially performed as well as the original Wine dataset. However, including the KM cluster labels (**Figure 24b**) and the EM cluster labels (**Figure 24c**) led to much poorer performance form both of these datasets. Furthermore, it is worth noting that both the PCA and ICA reduced datasets performed poorest across the board.

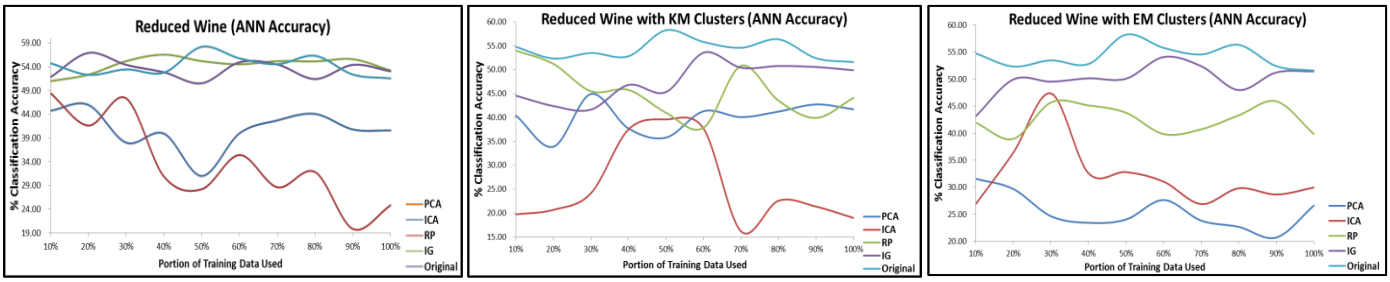

Figure 24: ANN Classification Accuracy. LTR: (a) Classification Accuracy vs % Training Data for Reduced Wine Datasets; (b)Accuracy vs % Training Data for Reduced Wine Datasets with KM Cluster Labels; (c) Accuracy vs % Training Data for Reduced Wine Datasets with EM Cluster Labels

Conclusions: My first major observation in this assignment is that the shapes, sizes and distinctness of the different clusters changes as each dimensionality reduction technique and clustering method is applied (and sometimes not for the better). This leads me to conclude that applying the different dimensionality reduction techniques results in different bits of information being "thrown away" from the datasets. The clustering techniques in turn place different weights/importance on different types of information, and this accounts for the various cluster shapes and sizes shown in this report.

Another important observation I made is that none of the clustering experiments output the correct number of clusters (i.e., the number of labels in the class attribute). This is probably due to the fact that there is noise in the datasets that cannot be perfectly modelled by any clustering algorithm, hence the mismatch.

Regarding the accuracy of the various clustering techniques, **Figure 25** below compares the accuracies of the different clustering methods on the Abalone and Wine datasets. From **Figure 25a**, we see that K-Means outperforms EM on the original Abalone dataset. However, once dimensionality reduction is introduced, EM performs at least as well as or better than K-Means in all but one case.

This is not the case with the Wine dataset however, as EM outperforms K-Means in almost all scenarios, (both before and after the use of dimensionality reduction). These results further buttress the point that there is no "silver bullet" or "one-size-fits-all" methodology/technique in machine learning. Some machine learning algorithms are better suited to certain datasets than they are to other datasets, and one needs to experiment with as many algorithms and techniques as possible to discover which one performs best on a given dataset.

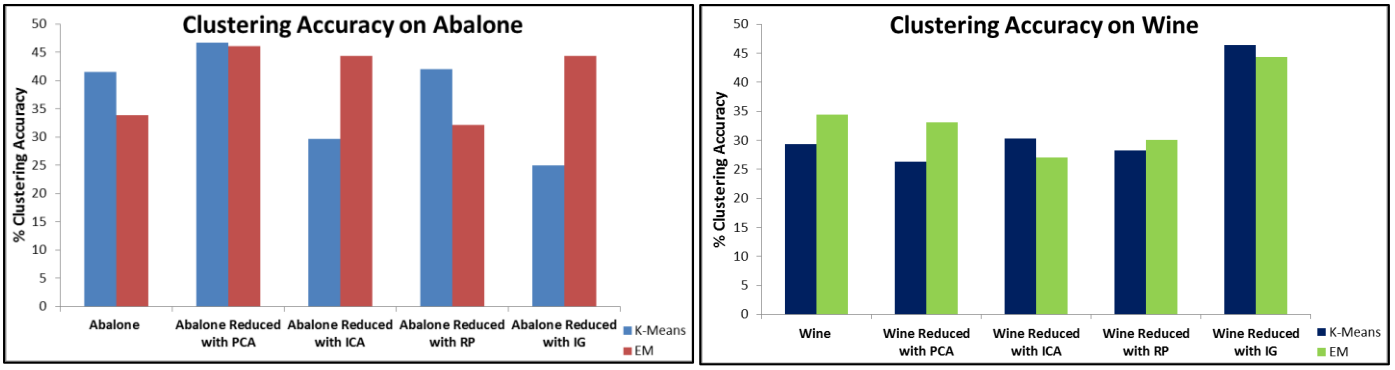

Figure 25: Clustering Accuracy on Abalone and Wine Datasets

Figure 26 below shows the time complexity of the ANN with different versions of the Wine dataset. The IG reduced data was the fastest to train the ANN with, in both scenarios where cluster labels were added. On the other hand, the ANN trained with RP reduced datasets took the most time in almost all scenarios.

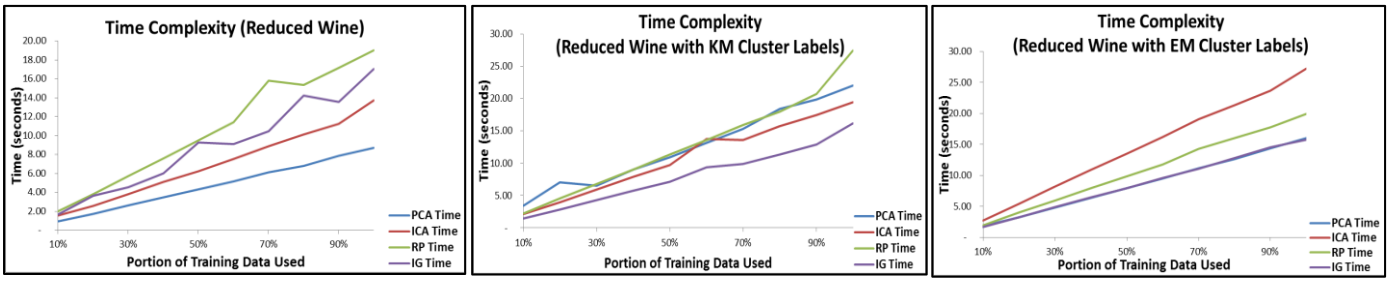

Figure 26: Time Complexity of ANN with Different Versions of the Wine Datasets

Areas for Future Improvement: A number of the results I obtained here lead me to believe that the elbow method might not be the best way to determine the optimal number of clusters. Given enough time, I would like to look into more deterministic methods of performing this analysis.

Also, I would like to spend some more time to understand the implications when the BIC measure increases as the number of clusters increases, as this might be indicative of the fact that he dataset is not suitable for clustering, for instance.

Regarding datasets, I would also like the opportunity to perform these experiments on many more diverse datasets, as the results I obtained here are inconclusive, and not enough to support any generalizations.