Un-Supervised Learning and Dimensionality Reduction

Anusha Challa – achalla8

1.Introduction: In this assignment two clustering algorithms are implemented. Clustering algorithms are a class of unsupervised learning algorithms. Of the various clustering algorithms available, K-means and Expectation Maximization are chosen to be implemented and the results are analyzed.

In Addition to the clustering algorithms, Dimensionality reduction algorithms have been implemented on the two datasets and the results are analyzed. The dimensionality reduction algorithms used are: Principal Component Analysis (PCA), Independent Component Analysis(RCA), Randomized Projection and Random Forest (algorithm I chose for feature selection). After the dimensionality reduction is applied on the datasets, clustering and neural network algorithms are re-run on the newly projected data and the results are analyzed

2. Datasets: The two datasets used are Breast Cancer dataset and Wine Quality dataset. These are the two datasets were used in the first assignment. Below is an overview of the datasets

a. Breast Cancer Dataset:

According to global statistics, Breast cancer (BC) is one of the most common cancers among women worldwide, making it a significant public health problem in today's society. Thus, the correct diagnosis and classification of patients into malignant or benign groups is the subject of much research. Breast cancer dataset has identified the features necessary to predict whether a solid lump is Malignant or benign. I have used the UCI machine learning repository for breast Cancer dataset from the URL: http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29. This dataset is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital.

Attribute Information: Dataset has 10 dimensions - 1. radius (mean of distances from center to points on the perimeter) 2. texture (standard deviation of gray-scale values) 3. perimeter 4. area 5. smoothness (local variation in radius lengths) 6. compactness (perimeter² / area — 1.0) 7. concavity (severity of concave portions of the contour) 8.concave points (number of concave portions of the contour) 9. symmetry and 10. fractal dimension ("coastline approximation" — 1).

The dataset contains 554 rows, each row representing a person. 'Class' is the column which indicates if the cancer is 4 = malignant or 2 = benign. We can identify that out of the 554 persons, 348 are labeled as 2 (benign) and 206 as 1 (malignant)

b. Wine Quality Dataset:

Wine preservation has been around since the late 1970s, there has been a recent surge in popularity of automated wine preservation systems in bars and restaurants. Intelligent preservation systems can be built to classify different wines based on quality, along with their preservation. Wine quality dataset has identified the features that are necessary to predict the quality of wine. I have used the White Wine Quality dataset from the URL: https://archive.ics.uci.edu/ml/datasets/wine+quality. The dataset is publicly available and is created by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Attribute information: Dataset has 11 dimensions - 1. fixed acidity 2. volatile acidity 3. citric acid 4. residual sugar 5. chlorides 6. free sulfur dioxide 7. total sulfur dioxide 8. density 9. pH 10. sulphates 11. Alcohol.

Output variable (based on sensory data) is 12. quality (score between 0 and 10). Wine quality dataset has quality ratings of 0 or 1. 0 being low quality and 1 being high quality. The dataset contains 4898 rows. We can identify that out of 4898 wines, 3258 are labelled as 1 (High Quality) and 1640 as 0 (Low quality). Wine quality dataset is highly correlated; hence it is expected that it would perform poor with clustering.

3. Clustering Algorithms: Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, Clustering algorithms can be divided into 2 subgroups.

a. Hard Clustering Algorithms: In hard clustering, each data point either belongs to a cluster completely or not.

b. Soft Clustering Algorithms: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

3.1. K-Means Algorithm:

K Means is a hard-clustering algorithm. It performs clustering using the below steps.

1. Chooses K random points as centroids
2. Computes the distance between each data point in the dataset and the k centers chosen in Step 1. Identifies that center which is the closet to each data point and classifies the data point under that center
3. Based on the classified points, re-computes the center by taking the mean of all vectors in the group
4. Repeat steps 2 & 3 until convergence where the center points don't change much between iterations.

3.1. Expectation–Maximization(EM) Clustering using Gaussian Mixture Models (GMM):

EM is a soft-clustering algorithm. GMMs assume that the data points are Gaussian distributed. This is a less restrictive assumption than that of K means algorithm that assumes that data points are circular and using mean. Since Gaussian distribution is assumed, we have 2 parameters to describe the shape of clusters, mean and standard deviation.

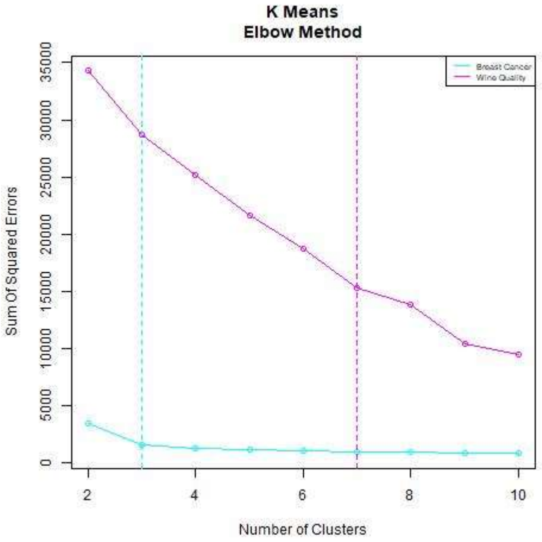
1. Choose K Random Gaussian centers and initialize Gaussian distribution parameters for each cluster
2. Given the initial Gaussian distributions, computes the probability that each data point belongs to a particular cluster
3. Based on these probabilities, computes a new set of parameters for the Gaussian distributions such that the probabilities of data points within the cluster is maximized. These new parameters are computed using a weighted sum of the data point positions, where the weights are the probabilities of the data point belonging in that particular cluster
4. Steps 2 and 3 are repeated iteratively until convergence, where the distributions don't change much between iterations.

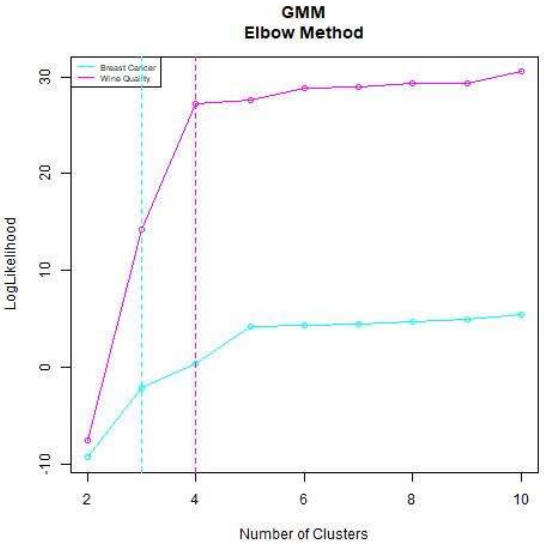
4. Implementation of Clustering Algorithms:

Clustering is implemented using python sklearn Library.The data was randomly split into a training set that included 70% of the observations and a test set that included 30% of the observations. A series of k-means cluster analyses and GMM cluster analyses were conducted on the training data specifying K=2-10 clusters. Given that both the datasets have two classifications each, the value two through 10 are chosen as the cluster sizes to search over. K-Means aims to minimize the distances of each point to the cluster center. GMM, which is the implementation of the EM algorithm, assigns probabilities for each data point. The error is measured using the cluster confidence or the Log Likelihood.

4.1. Choose the Best K (Elbow Method)

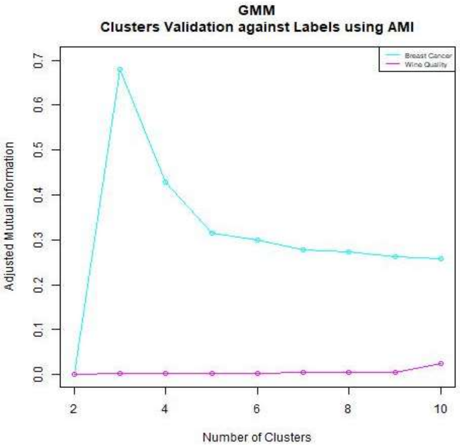
Choice of k impacts the outcomes of clustering significantly. Elbow method is used to determine the best K for each algorithm. The elbow method approximates the best cluster size by finding a point where the metric to optimize sees a significant decline in the rate of change. Even though this is a subjective measurement, it is clear in certain cases. The metric to optimize for K-Means is the Sum of Squared Errors(SSE) and for GMM is the Log-Likelihood. When SSE and log-likelihood are plotted against the number of clusters, the graph looks like an elbow at the optimal number of clusters. The number of clusters is chosen at this point; hence it is called the elbow method.

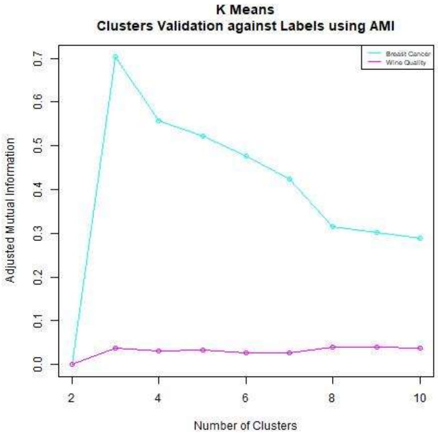




K Means: SSE over all data points was plotted for each of the nine cluster solutions in a curve. As per the elbow method, Ideal K for Breast Cancer Dataset: 3 & Ideal K for Wine Quality Dataset: 7

GMM: For GMM, the performance measure is the Log Likelihood, which should be high for optimal K. The mean log likelihood over all data points was plotted for each of the nine lusters. As per the elbow method, Ideal K for Breast Cancer Dataset: 3 & Ideal K for Wine Quality Dataset is: 4

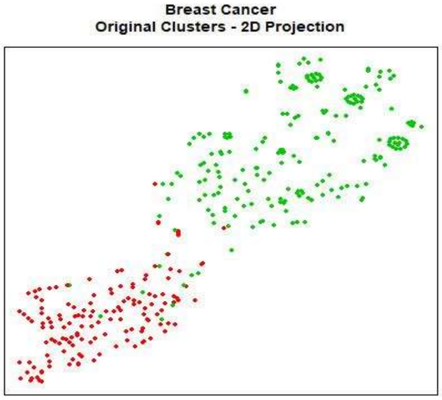


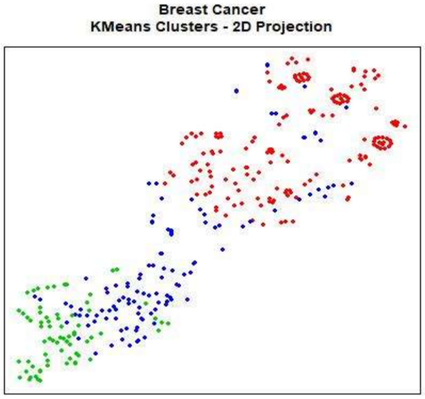


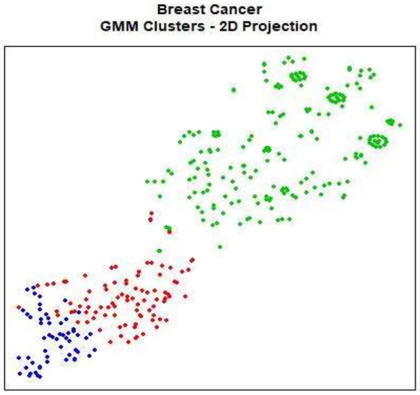
Both K-Means and GMM agree on the optimal K for the Breast Cancer dataset. Both To validate if the number of clusters obtained chosen using elbow method is ideal, labels from the dataset are used. This validation is possible only because the datasets that are chosen have pre-defined labels. The adjusted mutual information, like entropy, measures the similarity of clusters based on their true labels. The adjacent plots give credibility to the argument for three clusters when it comes to the breast cancer data. The mutual information is significantly higher for the case of three clusters compared to the other clusters tested. Other number of clusters like five clusters still do relatively well for K-Means, but fall short in the GMM validation. The Wine Quality dataset performs poorly overall when attempting to recreate the true class labels, never exceeding 0.1 AMI for either clustering algorithm. Since this data reflects close correlation, it makes intuitive sense that clustering is difficult as highly correlated data is nondeterministic.

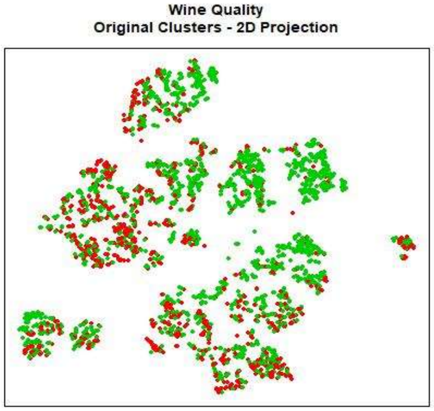
4.2: Description of clusters Obtained: The clusters from original dataset, the clusters obtained after implementing k-means and GMM clustering on datasets are projected into a 2D space using T-SNE to visualize the data. The resulting data is plotted below

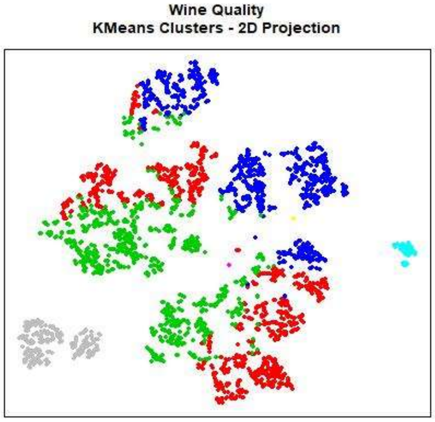
Breast Cancer Dataset: The original dataset has 2 labels. The distribution of the labels from the dataset is shown in the Original clusters graph. Right side of the original clusters plot is Malignant(Green) and the left side is Benign(Red). After applying K-means clustering, the clusters obtained line-up to the original clusters to a great extent. However, the output clusters have some

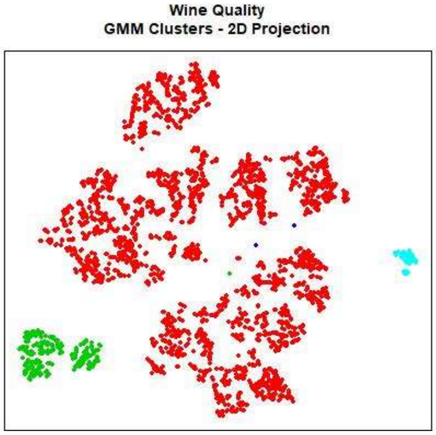












degree of overlapping and k-means couldn’t separate the clusters that are intuitive to human eye. Thus, the inter cluster distance that K-Means has produced is low. GMM on the other hand has done a great job in visually separating the clusters. GMM could group all the data points to the right side of the graph into one cluster which is lined-up well with that of the original data. Thus, the inter cluster distance produced by GMM is higher. The left side of the graph is divided into two subgroups. These subgroups might be representing classes within the cluster which can reveal important information about the cluster to researchers. GMM is soft clustering and the data points at the borders are given the color of that cluster with high probability.

Wine Quality Dataset: The original grouping of data is High Quality Wine (Green) and Low-Quality Wine (Red). Unlike Breast cancer dataset, wine quality dataset doesn’t have visually intuitive clusters in the original dataset. This hint high level of correlation between dimensions and hence it becomes difficult to perform clustering in general. The clusters obtained using k-means clustering look visually distinctive. However, Similar to that of breast cancer, the inter cluster distance are less compared to GMM. Like the K-means clusters obtained for Breast cancer dataset, there is overlap between clusters. GMM on the other hand produced visually more intuitive clusters. All data points on the left most corner, center and right most are grouped into their individual groups. However, as pointed earlier, there is high error when tested against the original labels. This can be explained by the fact that the original data labels have high degree of overlap.

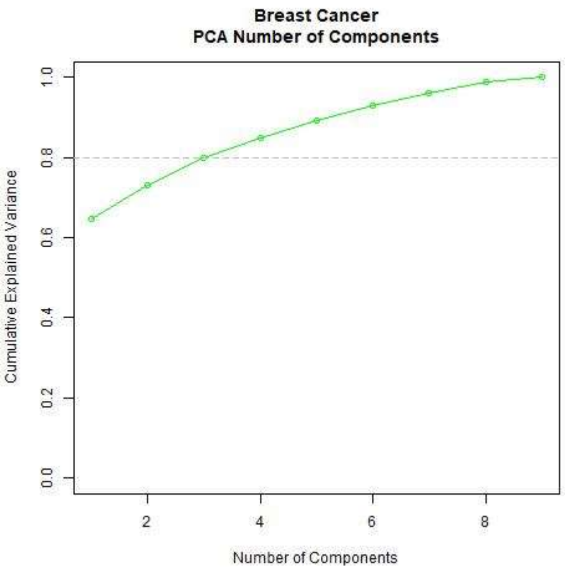
Overall GMM has done a better job at clustering both the datasets. As k-means assumes the clusters as spherical, it couldn’t cluster the breast cancer data properly where the clusters appear to be in a complex shape. On the other hand, since GMM can identify complex shaped structures it could cluster both the datasets more efficiently. Also given that GMM performs soft clustering and lets the data points be in multiple clusters with a probability value assigned, the error for GMM Is lower than that of k-means, as observed in the AMI graph.

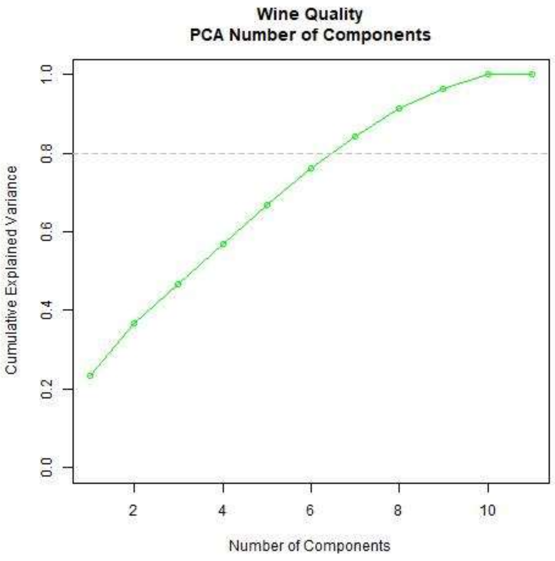
5. Dimensionality Reduction:

Dimensionality reduction is the process of reducing the number of features under consideration, while retaining as much relevant useful information as possible. A major issue that learning algorithms face is the curse of dimensionality. Dimensionality reduction can be used as a technique to mitigate this issue. This is beneficial from the standpoint of computational and time complexity along with the confidence one has in a model. Dimensionality reduction algorithms can also unveil hidden variables in the transformed space. The Adhoc information retrieval problem from the course lectures is a great example, as dimensionality reduction addresses the polysemy and synonymy issues.

5.1. Principal Components Analysis (PCA):

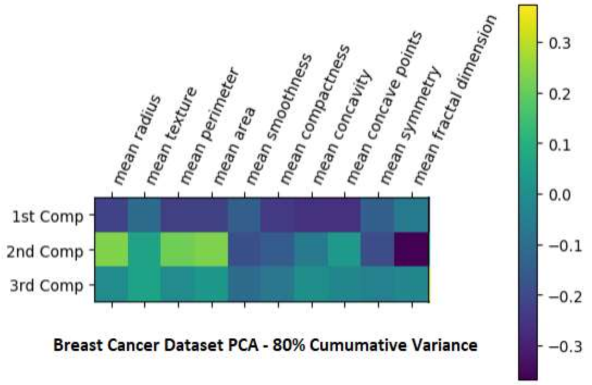
PCA is essentially a method that reduces the dimension of the feature space in such a way that new variables are orthogonal to each other. It reduces dimensionality by finding directions of maximum variance and minimize reconstruction error This can be interpreted as capturing correlations, and then using linear combinations of original variables to represent new variables.





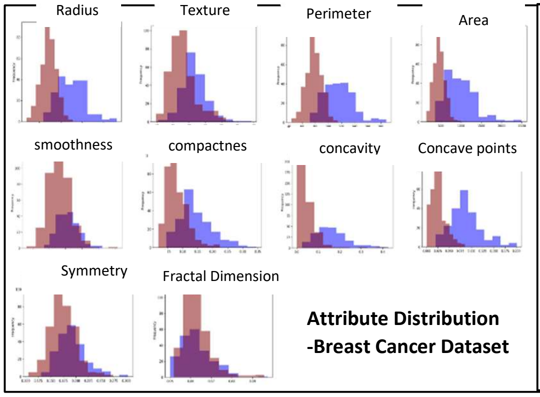
Analysis on datasets: Input for PCA is m, the number of components. Each dataset is tested for m values ranging from 1 to the number of features in that dataset. I.e., Breast cancer dataset is tested for m values from 1-10 features. Wine quality dataset is tested for 1-11 features. For each run and for each dataset, the cumulative variance is captured and plotted in the adjacent graphs. For this analysis, 80% of cumulative variance has been chosen as cutoff. By analyzing the graphs, you can observe that more than 60% of the Breast cancer data can be explained just by considering 2 features. 80% of the data can be explained by considering 3 out of 10 features. It can also be observed that only 20%-30% of the data can be explained by considering 2 featured for wine quality dataset. And 6 out of 11 features should be considered to explain 80% of the Wine quality dataset.

For the **breast cancer dataset**, the below heatmap shows how the features are mixed up to generate principal components. The heatmap shows the contribution of each of the 10 features in breast cancer dataset to the three principal components obtained to achieve 80% cumulative variance. It can be noticed that the features that contributed the most to the first principal component are Concave Points and Concavity. The features that contribute the least are Fractal Dimension, Symmetry, Smoothness and texture. These results agree with the data distributions obtained in assignment 1. As per the distributions of features, Concave points and concavity featured divide the two classes more distinctly compared to other features. And are Fractal Dimension, Symmetry, Smoothness and texture don't show any significant demarcation between the two classes.

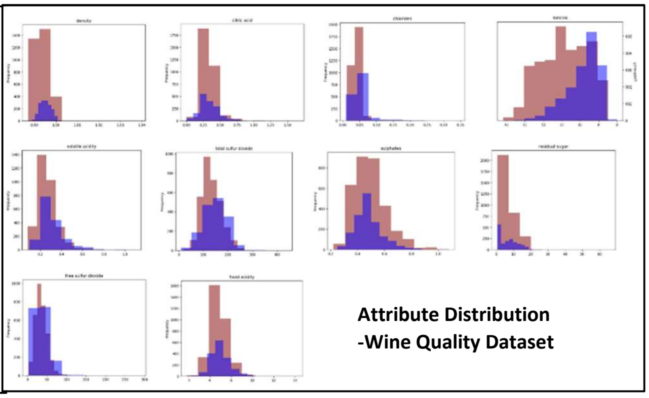


Breast Cancer Dataset PCA - 80% Cumumative Variance

The number of principal components obtained for Wine Quality dataset can as well be explained from the attribute distribution chart. The Wine Quality dataset has no clear distinction between the output labels for any of the input features as shown in the attribute distribution Chart. As a result, more Principal components are needed to describe data. And the first Principal component has contribution from all the features almost equally. Hence Principal Component Analysis works best with the datasets that have features that separate the data distinctly.



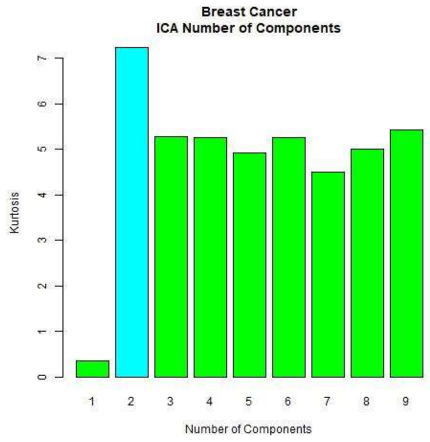
Attribute Distribution -Breast Cancer Dataset

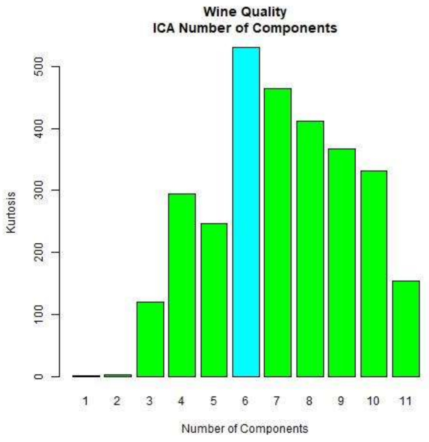


Attribute Distribution -Wine Quality Dataset

5.2. Independent Component Analysis(ICA): Independent component Analysis algorithm finds a linear transformation of features into a new feature space such that the new features are statistically independent of each other. ICA Algorithm assumes that the data can be separated into a new set of non-Gaussian, statistically independent variables. When run with larger values of K, PCA will produce the same first components. However, ICA can produce completely unique components for each value of k. Kurtosis, which measures non-Gaussianity, is used as a selection criterion for the value of k.

ICA is run for Breast Cancer and Wine Quality datasets for varying number of components. Each dataset is tested for m values ranging from 1 to the number of features in that dataset. I.e., Breast cancer dataset is tested for m values from 1-9 features. Wine quality dataset is tested for 1-10 features. For each run and for each dataset, kurtosis is captured and plotted in the graphs below. Kurtosis is a statistical measure that is used to describe distributions. It measures extreme values in either tail. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distributions. Distributions with low kurtosis exhibit tail data that are less extreme than the tails of the normal distribution. In ICA, the number of components that returns the maximum value for kurtosis is chosen as the ideal number of components. From the graphs below, it can be observed that



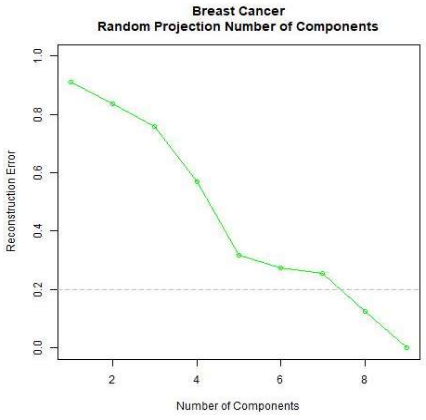


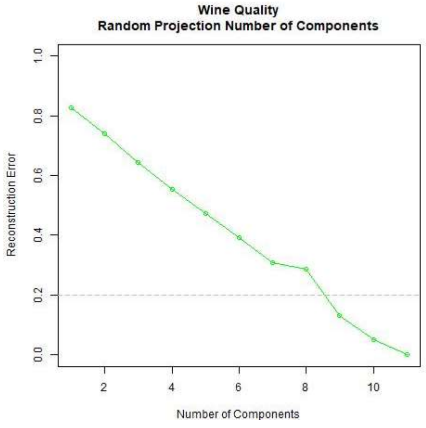
the breast cancer dataset has showcased highest kurtosis when the number of components is 2, which is lower than that of PCA. Whereas the Wine Quality dataset showcased the highest kurtosis when the number of components is 6, which is the same as that of PCA. The Number of components selected for each dataset is highlighted in Blue in the adjacent graph. The kurtosis of the breast cancer data increases to its maximal value at m=2, then remains almost constant at 5 for remaining values of m. On the other hand, the kurtosis for the Wine Quality data exhibits an increasing trend from m=1 to m=6 and then reduces for the remaining values of m. This complements the results of PCA well. The Wine Quality data shows low variance explained by its first few principal components and climbs slowly in kurtosis vs. m because it has underlying phenomena that are independent and require additional components to express. The opposite can be observed when comparing the breast cancer data's PCA vs ICA results.

5.3. Random Projections(RP): Random projection algorithms project the input n-dimensional dataset into a reduced dimensional space along random vectors or sometimes increased dimensional space, in a way which approximately preserves the distance between the points. This can also be said as projecting in a way that the reconstruction error is minimized. Reconstruction error is a measure of similarity between the original data and the reduced-dimension data when it is projected back into the original space.

Random projection is computationally simple: Wherein, it involves creating a random matrix "R" and project the d × N data matrix into K dimensions of order O(dkN). Unlike PCA, which requires processing all the data points beforehand in order to compute the projection, Random Projections let you choose the projection before actually seeing the data. This enables Random Projections to handle Realtime streamed data when needed.

Pythons sklearn module allows 2 different random matrices to choose from. For this analysis, Gaussian Random matrix has been chosen. Random projections were run ten times for varying number of components from 1 to the total number of features. I.e., For breast cancer dataset, Random projections were run 10 times for each of the component sizes varying from 1 to 9. Mean reconstruction error of the 10 runs is captured for each of the component sizes. Below are the plots obtained.



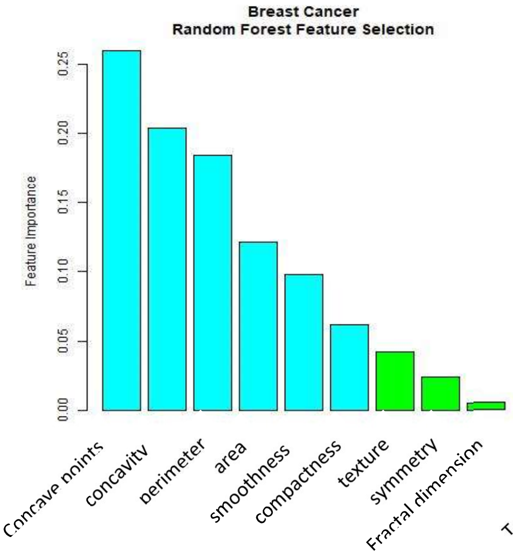


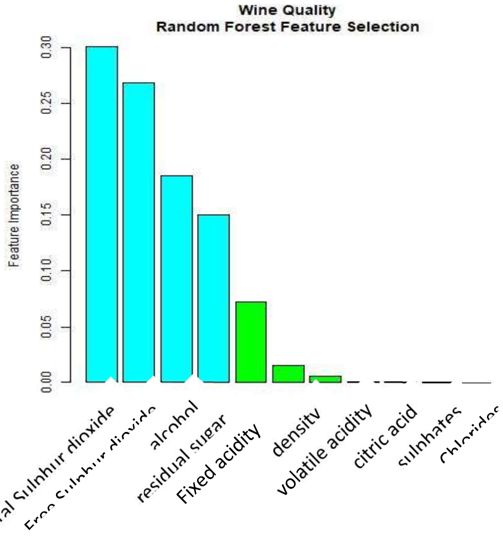
From the adjacent plots, it can be observed that the reconstruction error decreases as the number of components climbs towards the total number of features and eventually ends up at 0. It is expected since as the number of components increases, there is more scope to preserve the original data. An upper threshold of 20% has been considered for his Analysis. Breast cancer dataset shows exhibits the cutoff reconstruction error at 7 components and Wine Quality dataset exhibits 20% reconstruction error at 8 components. Random projections didn't benefit Breastcancer dataset as much as it benefitted Wine Quality dataset. This can be explained by the fact that the original breast cancer dataset already has features that can divide the data into their corresponding lables eficiently. Whereas, winquality dataset has features that don't separate the data well individually. When Random projections are run multiple times, a slight variance is observed between runs. Hence mean of 10 runs is considered.

5.3. Random Forest(RF): Random forest algorithm has been chosen as a feature selection algorithm. As the name suggests it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, mostly trained with the bagging method. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Feature Importance: Feature importance is used as a measure of relative importance of each feature on the prediction. Sklearn module in python measures a features importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so that the sum of all importance is equal to 1. Through looking at the feature importance, you can decide which features you may want to drop, because they don't contribute enough or nothing to the prediction process.

Below are the number of features selected for each data set, sorted in decreasing order by feature importance. A 90% cumulative importance cutoff established which variables to remove. The breast cancer data has selected six variables with feature importance reducing linearly for each feature. The Wine Quality data retained 4, but most of the importance lies within the first two features. These results compare favorably with the attribute distribution as will be explained in the next section. It seems that the breast cancer data shares importance amongst its features, which is why PCA favors it compared to the wine quality data. The wine quality data used the first 2 features to predict most of its data as per the Random forest feature selection.



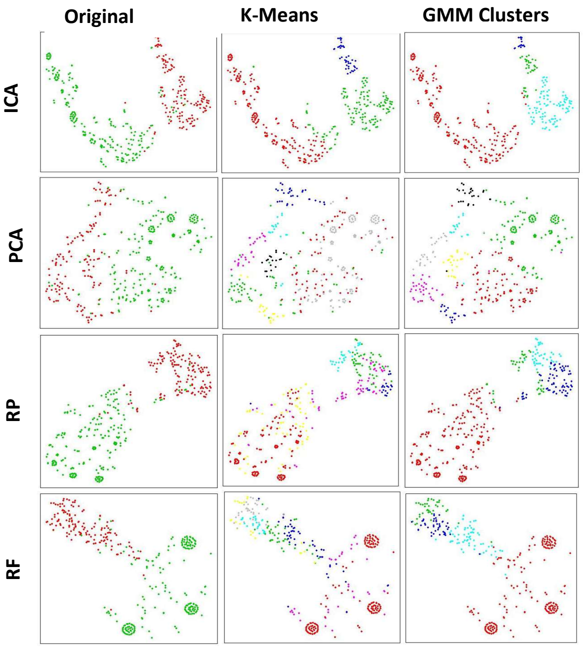


Since the datasets we are contain labels, the results from adjacent graph are compared against the attribute distribution as previously done for PCA. The results are in agreement with the attribute distributions. As per the attribute distributions (plots in section 5.1), the key attributes for breast cancer data set are concave points, concavity, perimeter, area, smoothness and compactness. Random forest algorithm returns the same set of features as the features of high importance. In the case of wine quality dataset, the attribute distribution doesn't visually indicate any particular attribute as the important feature. In this case, Random forest algorithm proves particularly useful. It picks the features total sulphur dioxide, free sulphur dioxide, alcohol, residual sugar as the important features, thereby eliminating more than half of the other features. You can also notice that the feature chlorides has 0 feature importance, which indicates that it is an irrelevant feature

6. Clustering on Dimensionally Reduced Data:

Clustering algorithms, K-Means and EM using GMM are applied on the data after applying the dimensionality reduction algorithms PCA, ICA, RP and RF. Each of the dimensionally reduced sets of data are run for varying number of components varying from 1 to the total number of features. For each clustering algorithm (K-Means and GMM), the best optimal number of Clusters is chosen using the elbow method.

Breast Cancer Data Set Cluster Analysis on Dimensionally reduced data: The 2D projections of the output clusters are shown in the adjacent graph. Original Clusters are the groups created using the original data labels



2D Projected Clusters for Breast Cancer Dataset

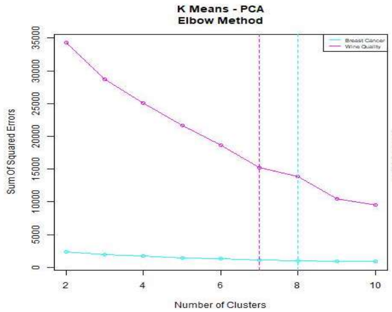
PCA: PCA hs scattered the datapoints in the available space. This is because PCA is a generalization algorithsm. As a result, it produced clusters that a spread in the space, thus repsresenting generality of the features obtained. The intercluster distance has been significantly reduced as a result of generalization. The optimal number of clusters has remained the same at 3 for GMM, as generalization doesn't impact the distribution in gaussian. For k means, it has increased to 4 clusters which can be explained by the dispersion of data points and hence needing more circles to represent data points

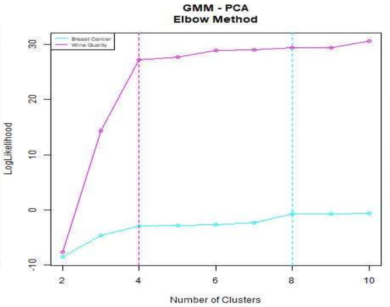
ICA: ICA has created 2 separate visually differentiable clusters. After ICA is applied, the Original data, K-means and GMM clusters obtained are all more distinguishable with larger inter cluster distances compared to that of base data. The optimal number of clusters is found to be 8 for both K-Means and GMM, this is higher than that of the base data. This can be explained by the fact that since the dimensionality is reduced, more distinct clusters were produced by assuming circles for K-Means and gaussian for GMM.

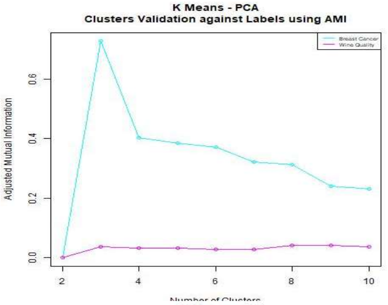
RP: Randomized projections didn't distort the clustering as much as ICA and PCA did. However, as the dimensionality is reduced, the clusters look more spherical. However, the inter cluster distance has increased thus creating better clusters than that of base data. The ideal number of clusters has increased from 3 to 4 and 6 for K-Means and GMM respectively. Significant increase in number of clusters for GMM is due to the increases inter cluster distances.

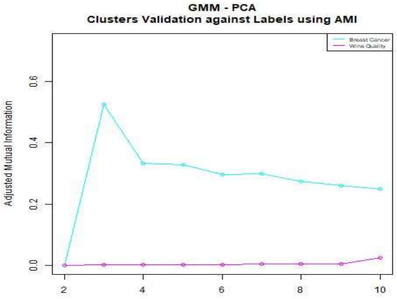
Random Forest(RF): Random forest algorithm has identified concentrated circular patches of data points by eliminating the features three features that are not less important. The output has eliminated noise that was in base data and brought similar data points closer to each other. The optimal number of clusters is found to be 4 for K-Means and 7 for GMM.

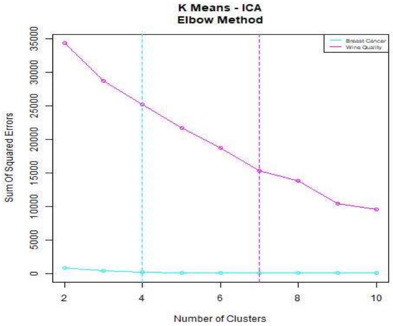
Cluster Validation using AMI for Breast Cancer dataset: It is observed that the overall AMI has reduced for both K-Means and GMM, this increasing the validation error within clusters. The reduction is significant for PCA and Random Projections. Hence these algorithms resulted in more loss of data and thus increasing validation error for breast cancer dataset. AMI has reduced overall for K-Means algorithm but at 3 clusters, the AMI remains the same for k-means algorithm. Similar to that of GMM, the reduction in AMI for clusters over 3 is higher for PCA and RP and not as higher for ICA and RF.

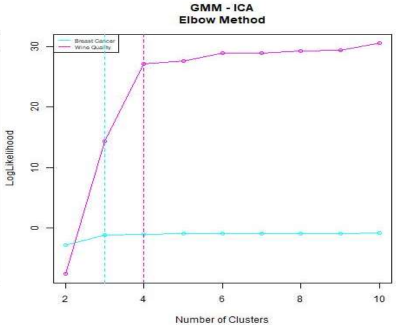


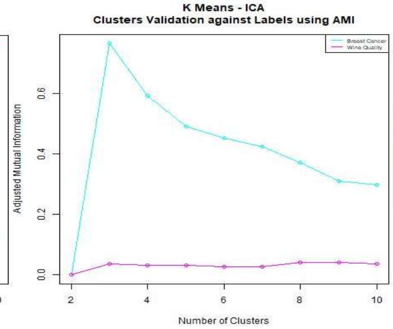


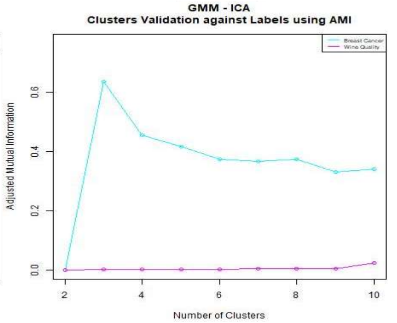


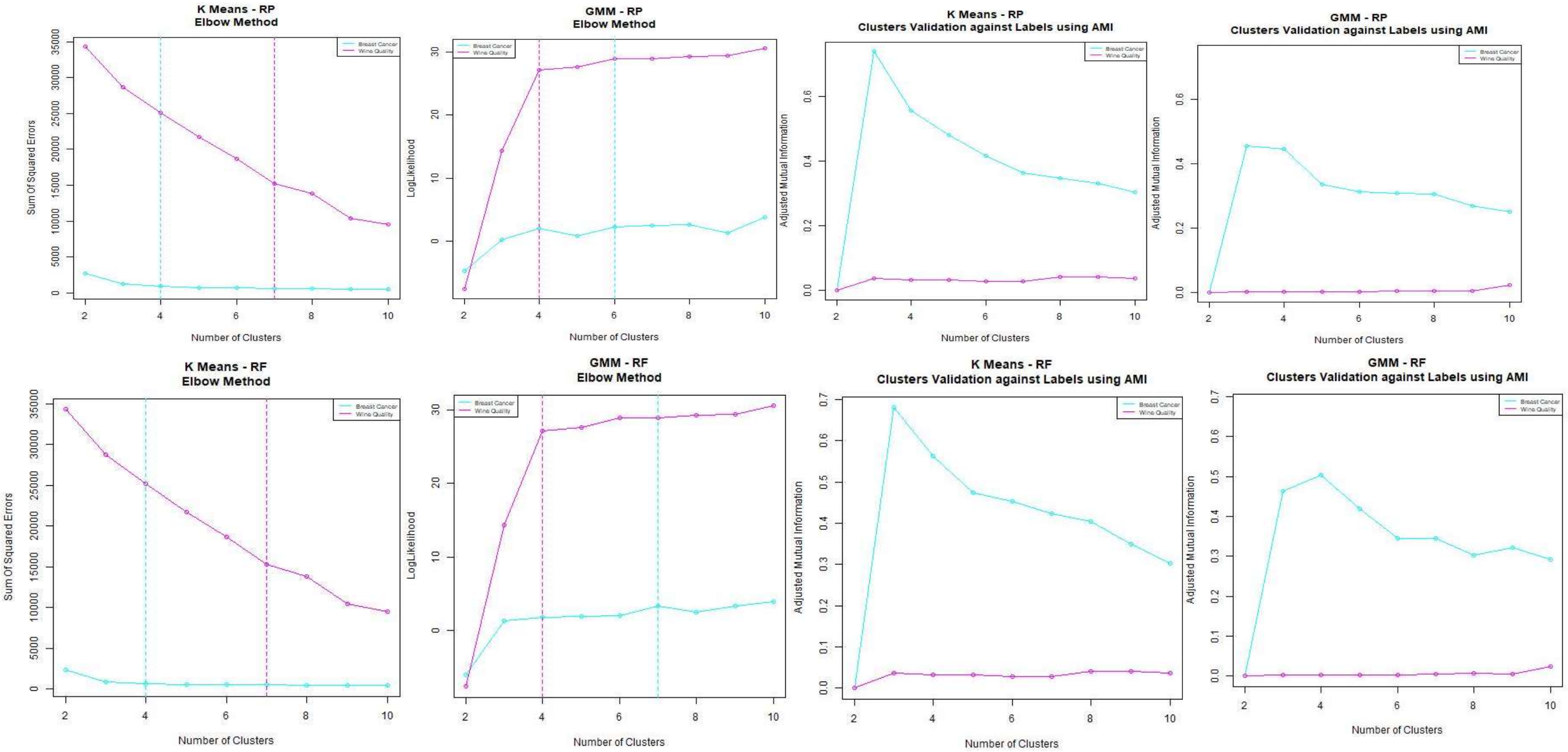












Wine Quality Data Set Cluster Analysis on Dimensionally reduced data: Unlike breast cancer dataset, Dimensionality reduction has not shown any significant impact on the wine quality dataset results. The number of clusters identified by elbow method remained the same for both K-Means and GMM algorithms. The AMI has remained the same as well. This can be explained by the fact that Wine Quality dataset originally has highly co-related features. The results on WineQuality dataset can be seen in purple lines in the graphs above

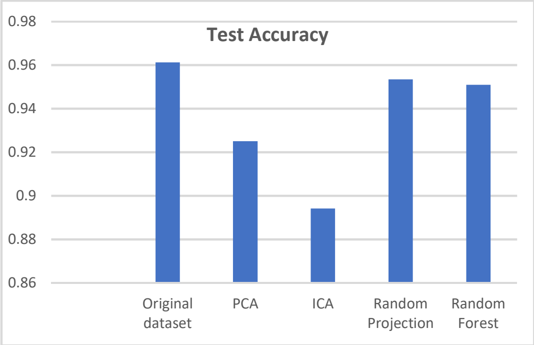
7. Application if Dimensionality Reduction and Clustering outputs to Neural Networks:

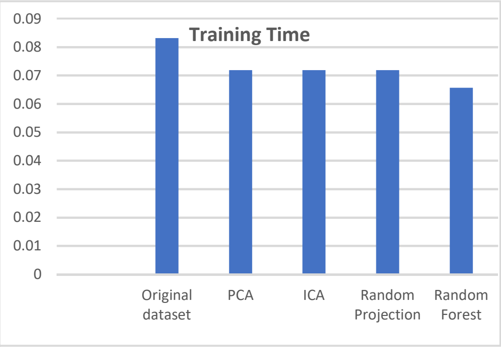
The output of dimensionality reduction algorithms on Breast Cancer dataset is applied to neural network created during assignment 1 and the results are compared for each of the algorithms. Only the optimal number of components identified in section 5 are used for neural network. The neural network created in assignment one was recreated and its optimal architecture used to test the performance of the dimensionally reduced data, using each algorithm respective optimal value of k. The neural

1. With the same optimal neural network parameters that are identified during Assignment 1. I.e., 2 hidden layers with 7 and 5 Nodes each. Test accuracy obtained was 0.9825 at 200 iterations.
2. Identified the optimal parameters again for each of the dimensionally reduced outputs of individual algorithms. Below are the optimal parameters identified for each of the datasets

	#Features	# Hidden Layers	# Nodes	Initial Weight	Test Score	Training Time	Query Time
Original dataset	9	3	12, 12, 12	0.001	0.9612403100	0.08312256811	0.00411259
PCA	3	2	18,18	0.001	0.9250645994	0.07188167578	0.00311799
ICA	2	2	12, 12, 12	0.001	0.8940568475	0.07188014986	0.00312357
Random Projection	7	3	18, 18, 18	0.0031	0.9534883720	0.07187571523	0.00312522
Random Forest	6	3	18, 18, 18	0.0316	0.9509043927	0.06562395099	0.00625491

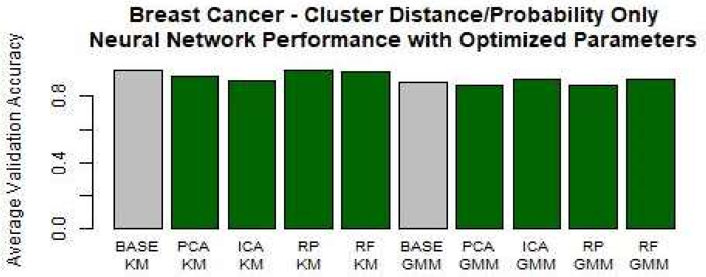
Accuracy: The results are plotted in the bar plot below. It ca be observed that after apply dimensionality reduction on data, the accuracy of the models has not increased. Accuracy of ICA has significantly reduced. Accuracy after PCA has reduces as well but not as much as ICA. Accuracy of Random Forest and Random projection is almost the same and only slightly less compared to that of the original neural network results. This is expected as the dimensionality is reduced, accuracy of the data is impacted

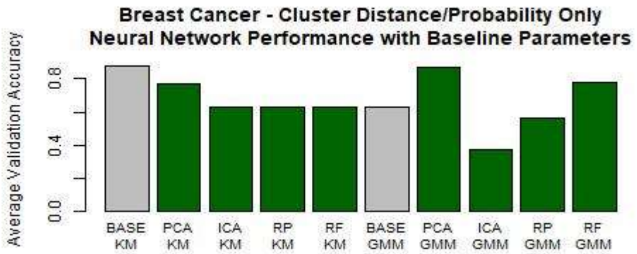




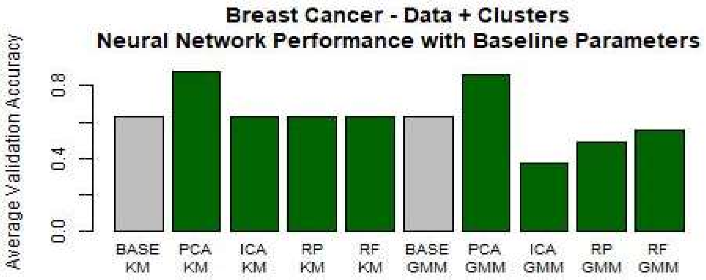
Time: It can be observed that the time taken to Train and query the model has reduced for all the dimensionality reduction models. This is expected as the dimensionality is reduced, the number of features that the neural network has in its input layer is significantly lesser compared to that of the original dataset

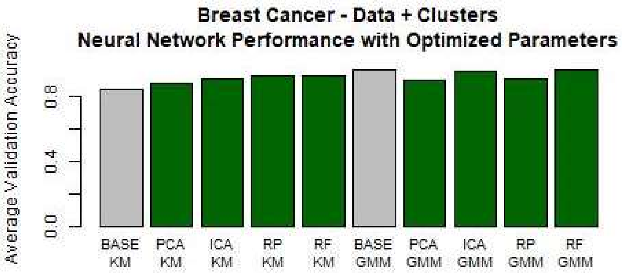
Secondly, the neural networks are run by using the Distance to the center of cluster for K-Means and probability of belonging to a cluster for GMM as predictors. All clustering and dimensionality reduction combinations used to run the neural network use the optimal values of cluster and dimension number identified in the previous sections. Neural networks have their hyperparameters selected through grid search. Similar to the first run, the network is run using the baseline parameters from first assignment and new optimal parameters are identified for each of the dimensional reduced clustered datasets. For both the datasets, it is observed that GMM perform slightly less than average. K-Means with RP and RF performed equally as well for validation accuracy as that of the base data.





Finally, the network was trained on breast cancer dataset along with the cluster assignments as an additional predictor. For the breast cancer data, like the previous run, GMM performs poorly on average across all dimensionality reduction algorithms. It is interesting to note that despite this, the best scoring combination in this category is the baseline data along with GMM clusters. As should be learned when studying dimensionality reduction: fewer features are (generally) preferred.





Overall, the best validation performance for the breast cancer data is 96.64% and it belongs to the network trained on original data with the GMM features attached. The performance of neural network is improved after applying dimensionality reduction. Hence it is proved that curse of dimensionality is a fact and as dimensions reduce, the learners tend to perform well

Interesting groupings on data have been revealed by performing Clustering. The number of groups that clustering identified is often more than the groups contained in the original dataset, which suggests that smaller subcategories may exist the labels assigned. Thus, clustering would be very helpful to Subject matter experts to identify subgroups within identified groups of data