# CS7641: Machine Learning Assignment 3
# By Yumei Wang (ywang4068)

Experiments on clustering and dimensionality reduction algorithms were explored in the following five steps:
**Step1:** Two clustering algorithms: k-means clustering and Expectation Maximization were evaluated on two datasets.
**Step2:** Four dimensionality reduction algorithms: PCA, ICA, Randomized projections (RP) and Random forest (RF) algorithm were applied to the two datasets.
**Step3:** Two clustering algorithms were reproduced on the dimensionality reduced datasets from step2.
**Step4:** Using the dimensionality reduced wine classification dataset from step2, neural network algorithm was trained to make predictions on test data.
**Step5:** Neural network algorithm was applied on the wine classification dataset transformed by the two clustering algorithms and then compared the performance with those from step4.
Both clustering and dimensionality reduction algorithms are implemented using Scikit-learn. GaussianMixture was used for EM, FastICA was used for ICA, and GaussianRandomProjection was used for RP.

## Datasets description

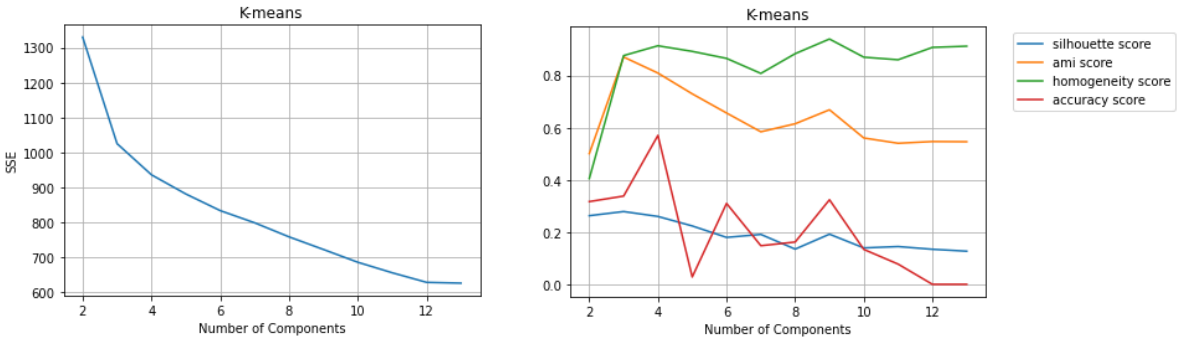
Two datasets can be downloaded from UCI and Kaggle.

1. Wine classification dataset [1] as dataset 1: It was loaded from sklearn datasets [2], and has 178 instances, 13 attributes and 1 output with 3 classes (0-2) and no missing values. All attributes are constituents found in three types of wines and the goal is to determine the origin of wines using chemical analysis. The dataset is balanced with 59 instances of class 0, 71 instances of class 1, and 48 instances of class 2. All attributes are continuous.
2. Breast cancer dataset (dataset 2): It was also loaded from sklearn datasets [3], and has 569 instances, 30 attributes and 1 output with 2 classes and no missing values. It is to classify whether the breast cancer is benign or malignant. The dataset has 357 instances of class 1, and 212 instances of class. It is not unbalanced.

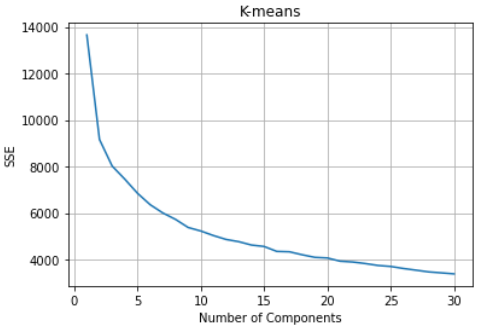
## Step 1: clustering

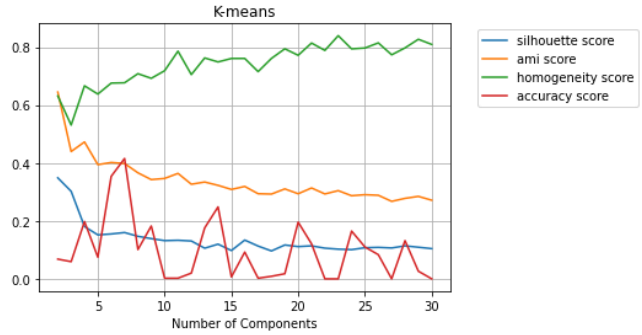
### K-means

To choose a good k, the sum of squared distances (SSE) between centroid and each member of the cluster was plotted against number of clusters. The objective of k-means is to try to minimize this value. The elbow point, where SSE decreases sharply and starts to bend, can be used to determine k.
And to confirm the selection of k, the silhouette coefficient (SC) score, homogeneity score, adjusted mutual information (AMI) score and accuracy score were plotted against number of clusters. The SC score measures how well the data point fits in the assigned cluster as compared to the other cluster. Homogeneity measures how much the sample in a cluster are similar. AMI measures the agreement between the cluster assignments. The higher of these scores, the better. And as the datasets have true labels, the accuracy score between the predicted labels and true labels was calculated.
**Dataset 1:**




For dataset 1, the elbow point was when number of clusters was 3, and the SC, homogeneity and AMI scores were the highest so the best choice of k-means clusters was 3 for dataset 1.
**Dataset 2:**



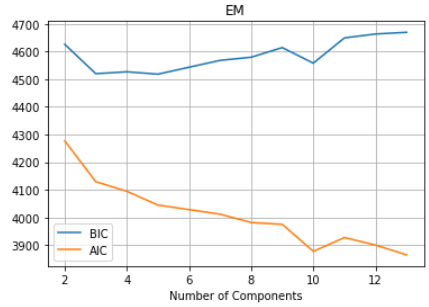



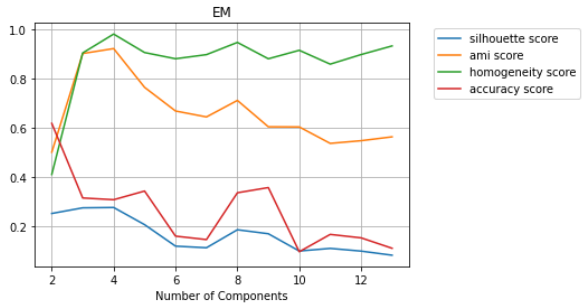

The SSE elbow point was not obvious for dataset 2, and might be 2-5 components, but the SC and AMI scores decreased as the number of clusters increased, so the number of clusters was chosen 2.

## Expectation Maximization (EM)

To choose an appropriate number of components for EM, the Bayes Information Criterion (BIC) and the Akaike Information Criterion (AIC) scores were plotted. Both measure the trade-off between model fit and complexity of the model. BIC is better suited to identify the true model. The lower of the scores, the better. Then the SC, homogeneity, AMI and accuracy scores were evaluated afterward.

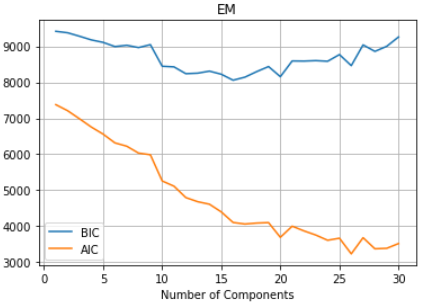
**Dataset 1:**

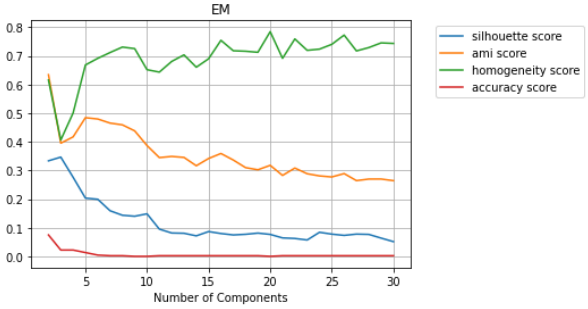



The number of components was set as 3 due to the lowest BIC and highest SC and AMI scores.

**Dataset 2:**



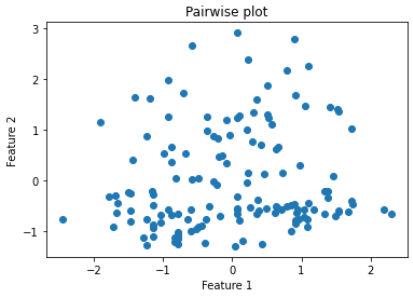



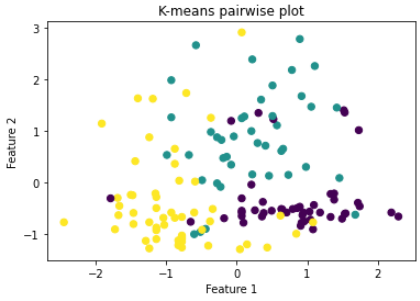

The BIC score decreased and was lowest when number of components was 16, but the SC and AMI scores decreased as number of components increased, so 2 components were chosen.

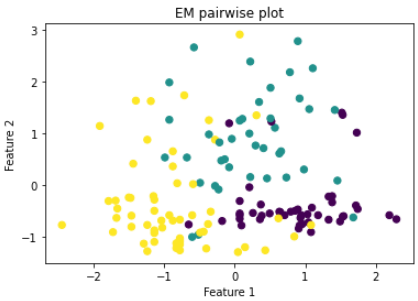
## K-means vs EM

To validate the performance of k-means and EM clustering, pairwise plots of the first two features were compared.

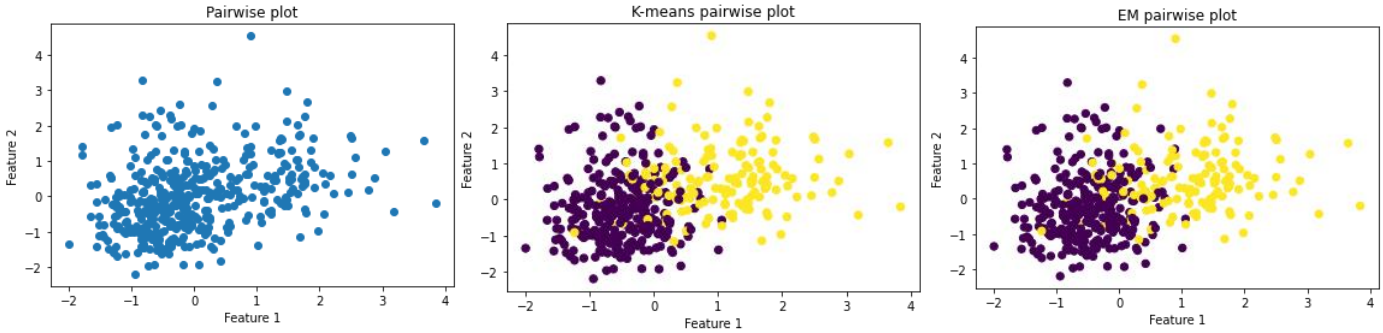
**Dataset 1:**



**Dataset 2:**



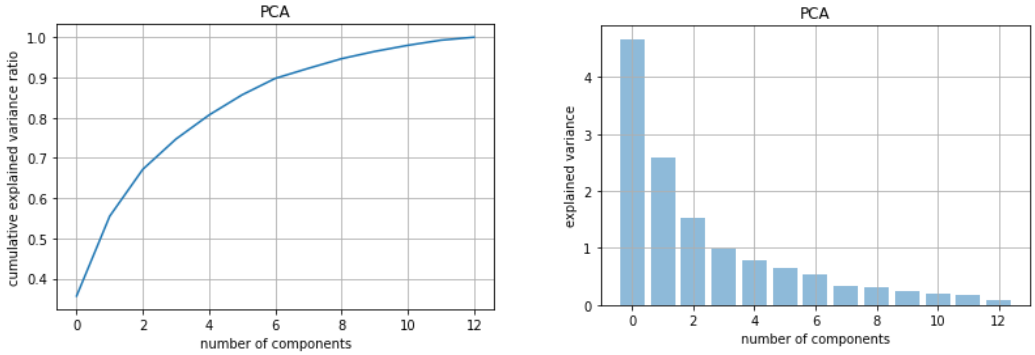

For both datasets, from left to right, it was the pairwise plot of two features from original dataset, after k-means and after EM. K-means and EM were very similar in clustering both datasets and the accuracies of labeling were also similar. For dataset 1, the accuracy was around 30%, and for dataset 2, the accuracy was only 10%. This suggested clustering was not suitable to make classifications for these datasets.

## Step 2 & 3: Dimensionality reduction and clustering

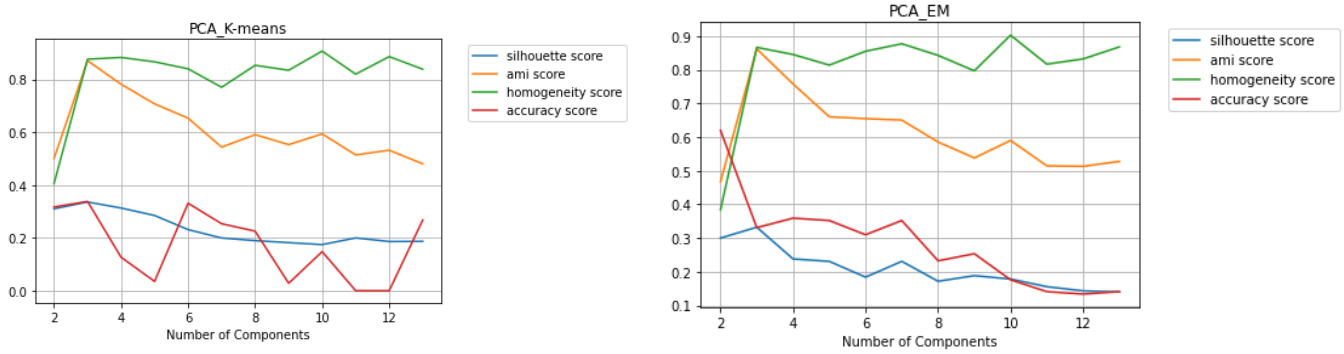
Cumulative explained variance ratio, kurtosis, reconstruction error was plotted for PCA, ICA and GRP respectively to choose the appropriate number of components. Random forest was used as a feature selection algorithm to reduce dimensions, which selected features based on feature importance. K-means and EM clustering were then applied to the dimensionality reduced datasets and evaluated by SC, AMI, homogeneity and accuracy scores. Pairwise plot of the first two features were also plotted.

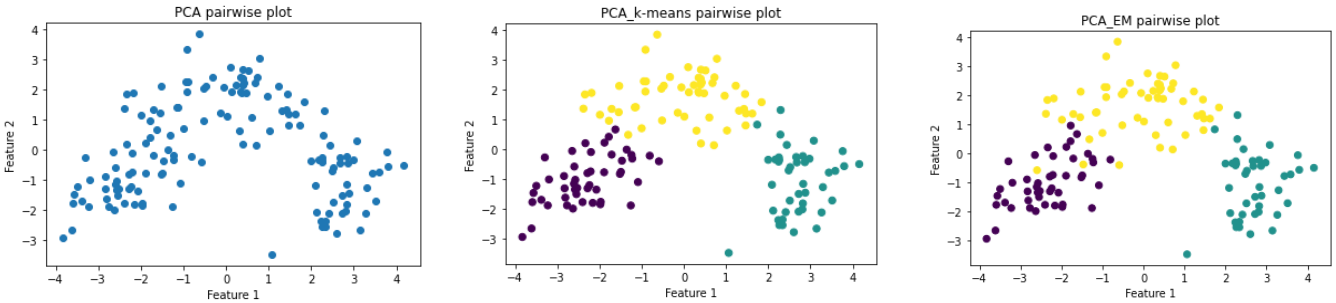
**Dataset 1:**

**PCA with clustering**



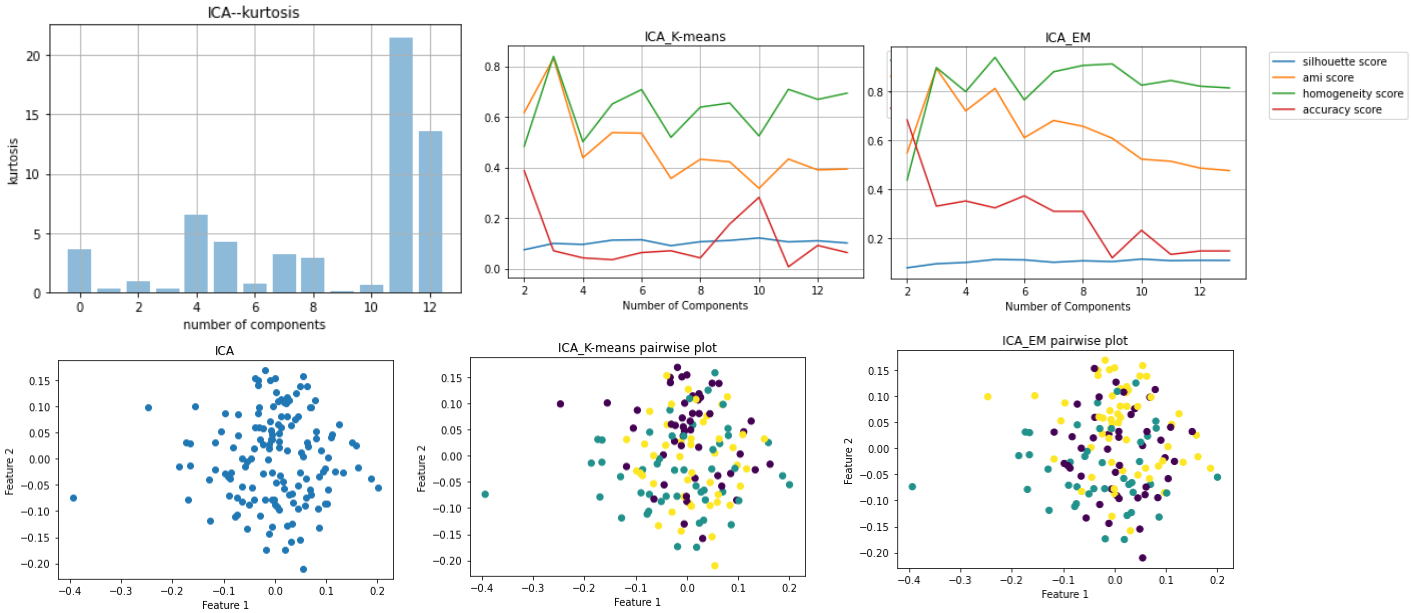

Number of components were chosen 6 for PCA to capture 90% of the variance.



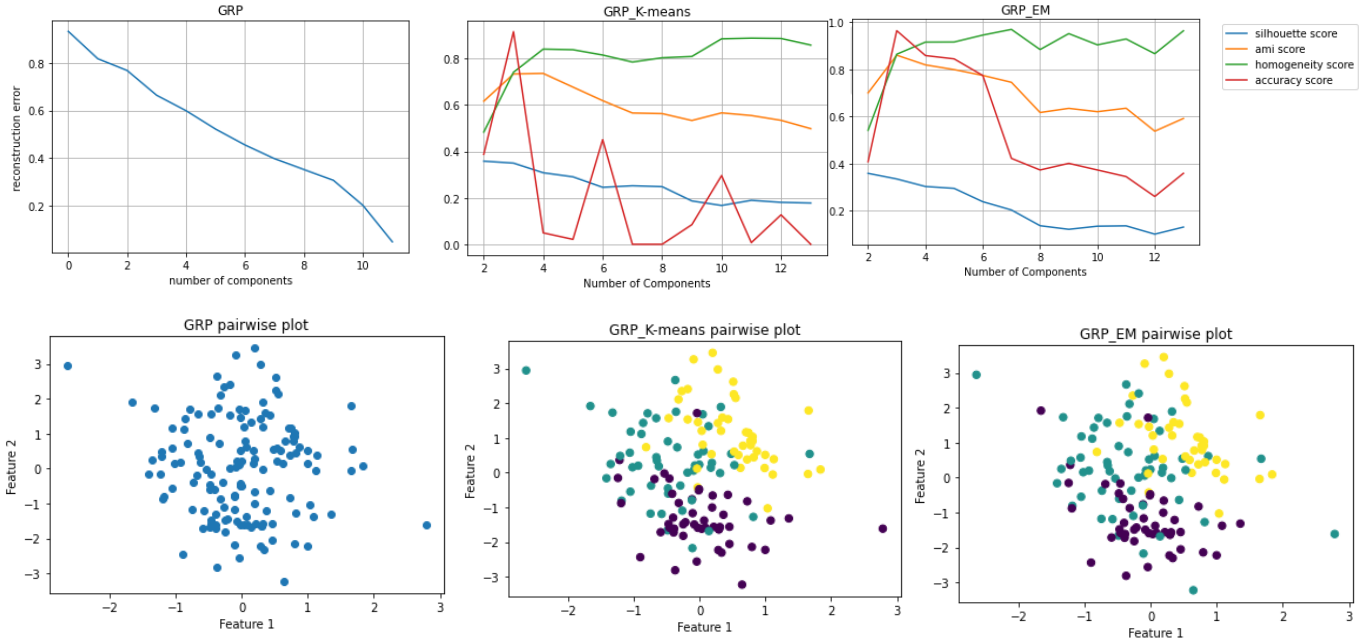
The number of clusters were both chosen 3 for PCA_K-means and PCA_EM and both separated the points well into 3 clusters. The clusters were similar to the original dataset, but separated better. The accuracy of labeling for both were similar, and no better than the no reduction original dataset.

**ICA with clustering**




Number of components were chosen 11 for ICA due to the highest kurtosis. The number of clusters were both chosen 3 for ICA_K-means and ICA_EM and both were not able to separate the points as the clusters were overlapped. The clusters created from ICA reduced dataset were dense, different from the original dataset which were scattered. The accuracy of EM was the same with original dataset, but accuracy of k-means (10%) was lower than original dataset.

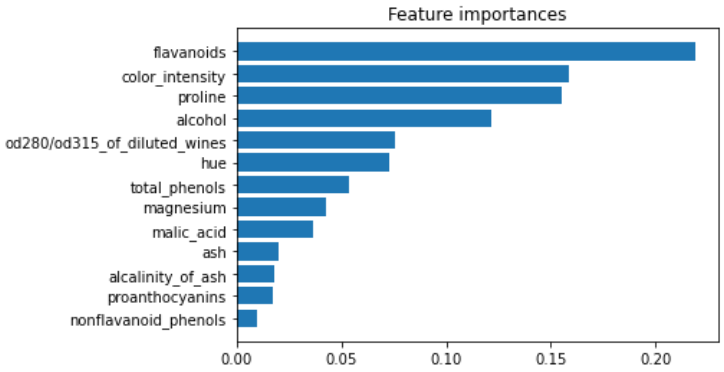
**GRP with clustering**

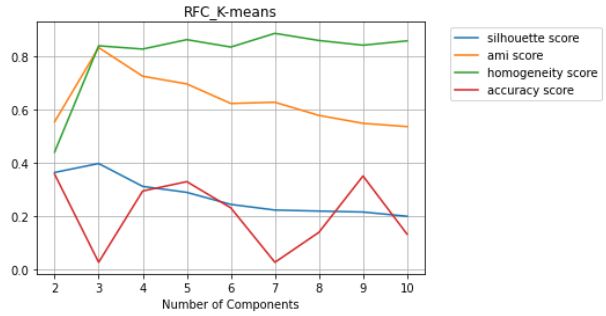


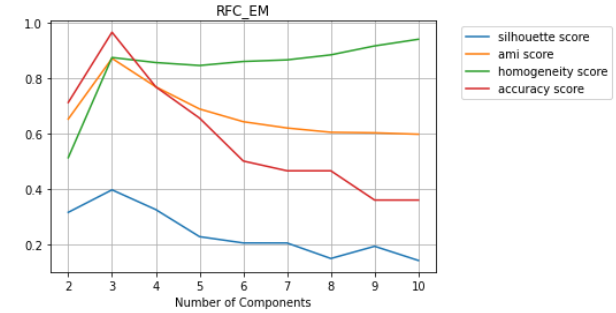

Number of components were chosen 12 for GRP for the lowest reconstruction error. The number of clusters were both chosen 3 for GRP_K-means and GRP_EM and both can separate some of the points. The clusters were very different form the original dataset. But the accuracy of labeling for both were over 90%, which were 3 times of that for the original dataset.

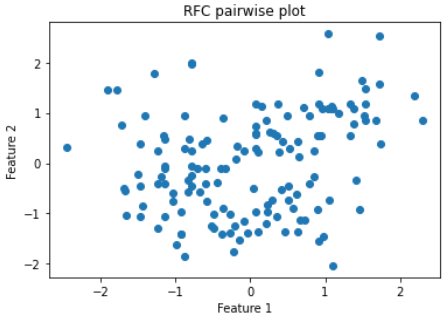
**RFC with clustering**

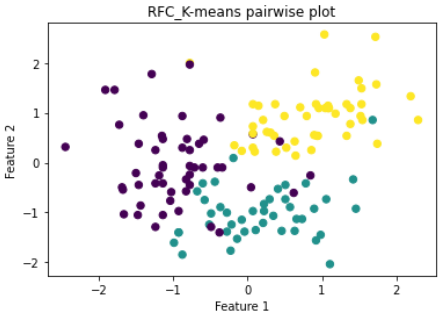
Feature importance was investigated and features with importance greater than 0.05 were chosen only, so the original dataset was reduced from 13 to 7 features.

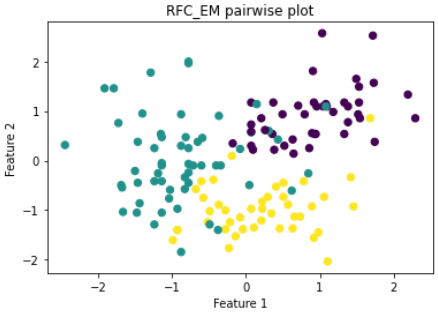


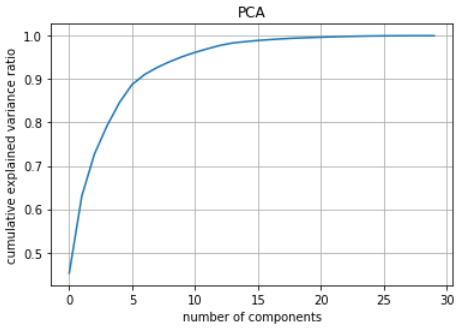



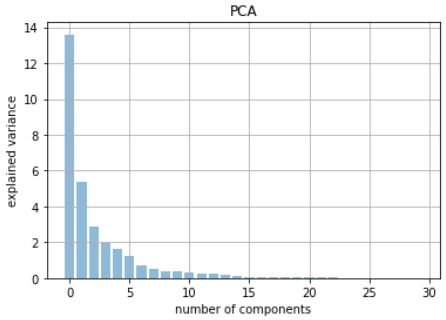




The number of clusters were both chosen 3 for RFC_K-means and RFC_EM and both can separate most of the points. The clusters and accuracy scores were different from the original dataset. The accuracy of labeling for EM was over 95%, but the accuracy of k-means was very low (less than 10%).

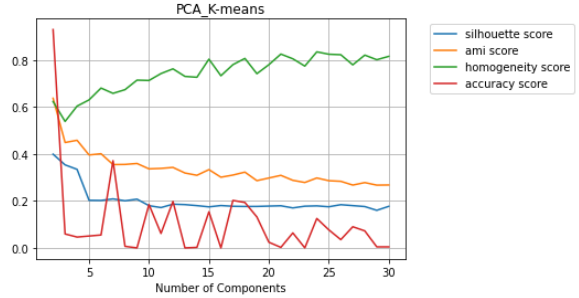
**Dataset 2:**

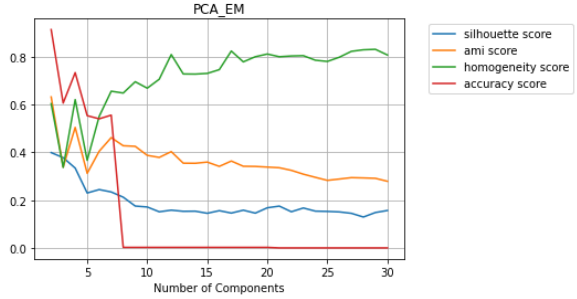
**PCA with clustering**

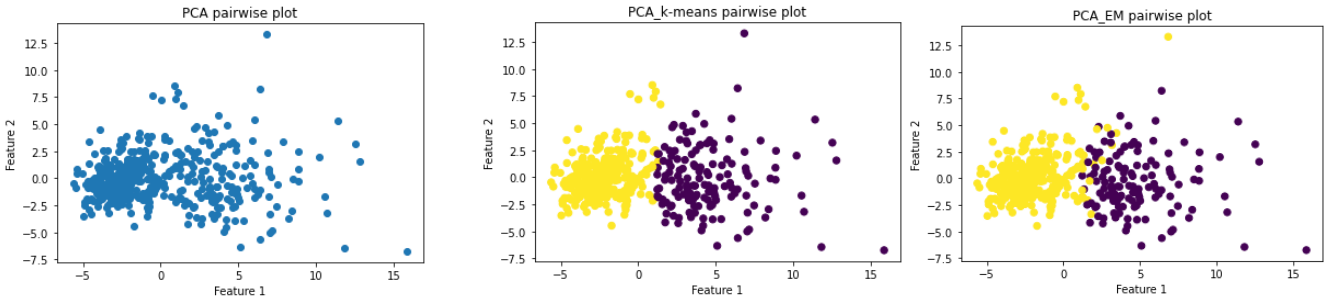



Number of components were chosen 5 for PCA to capture 90% of the variance.

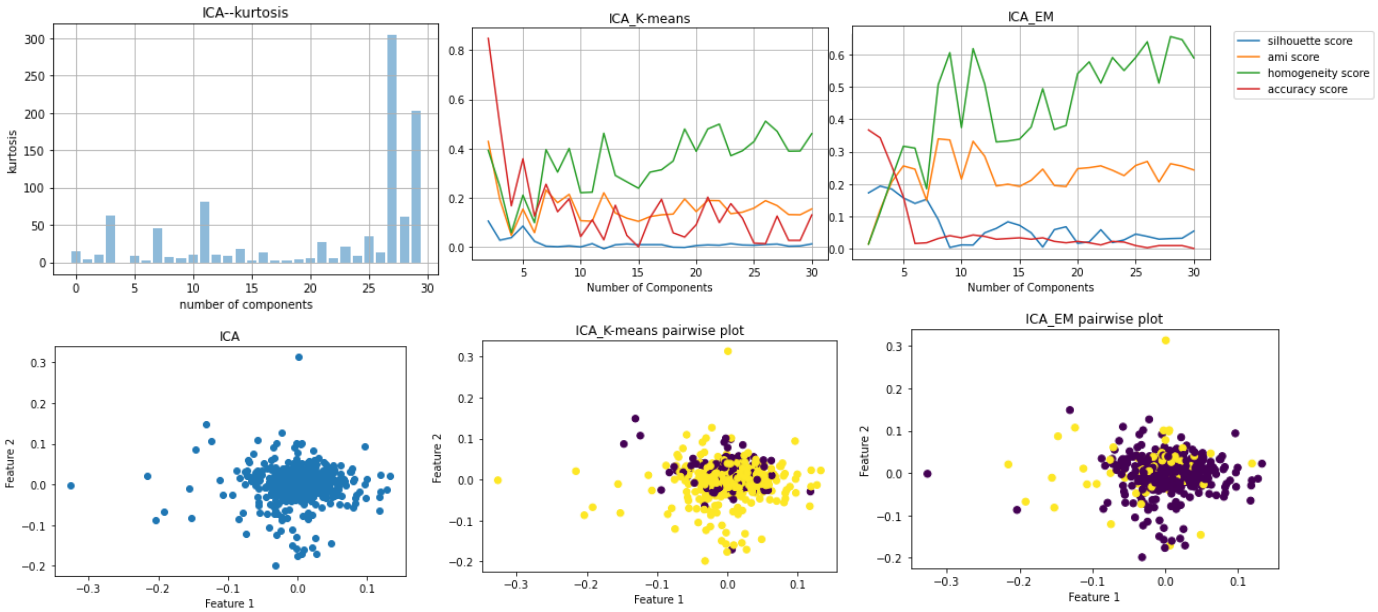




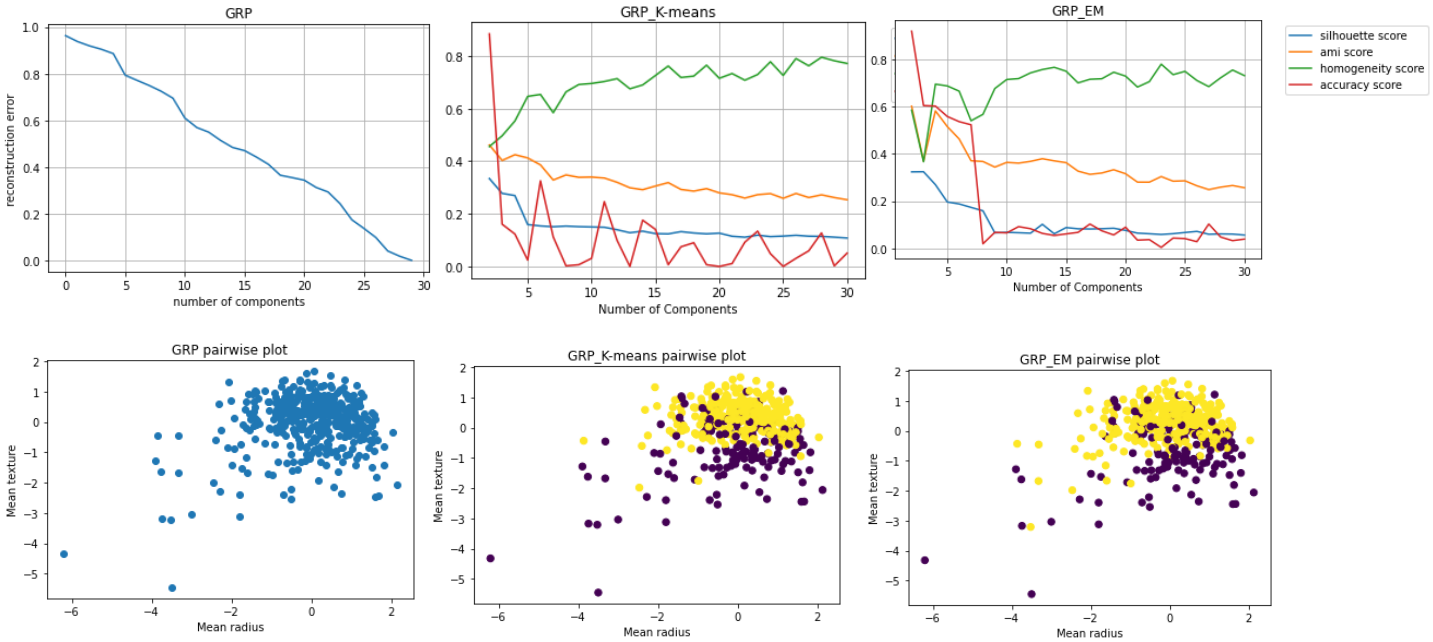

The number of clusters were both chosen 2 for PCA_K-means and PCA_EM and both separated the points well into 2 clusters. The clusters were similar to the original dataset, but separated better. And the accuracy of labeling were both over 90%, which were 9 times of that for the original dataset.

**ICA with clustering**




Number of components were chosen 27 for ICA due to the highest kurtosis. The number of clusters were both chosen 2 for ICA_K-means and ICA_EM and both were not able to separate the points as the clusters were overlapped, which were different from the original dataset. But the accuracy of k-means was 90%, and EM was 35%, and both were higher than the original dataset.

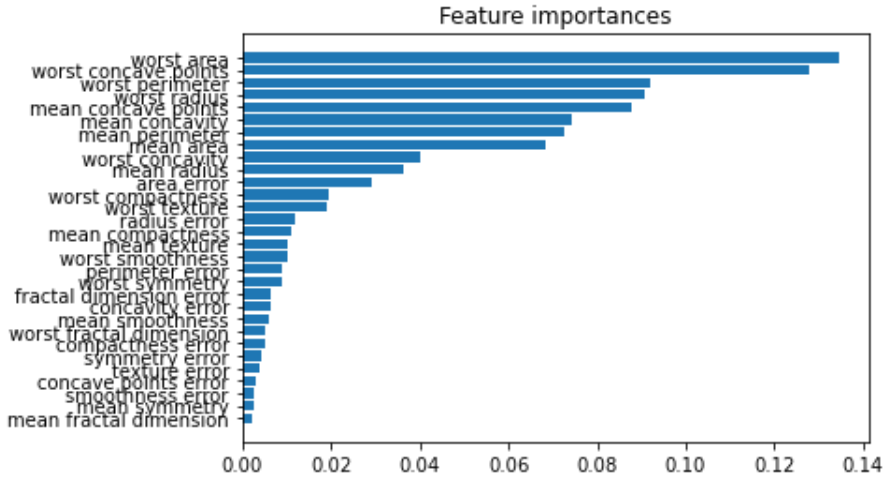
**GRP with clustering**

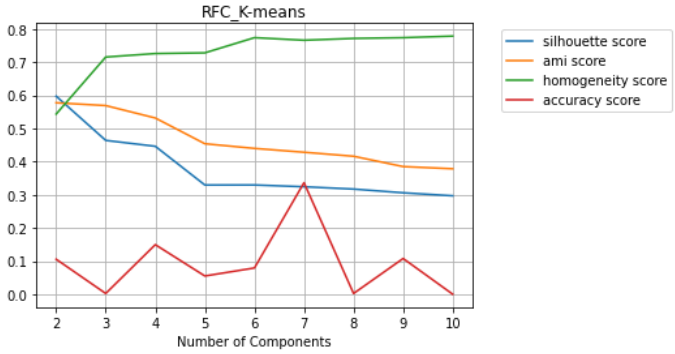


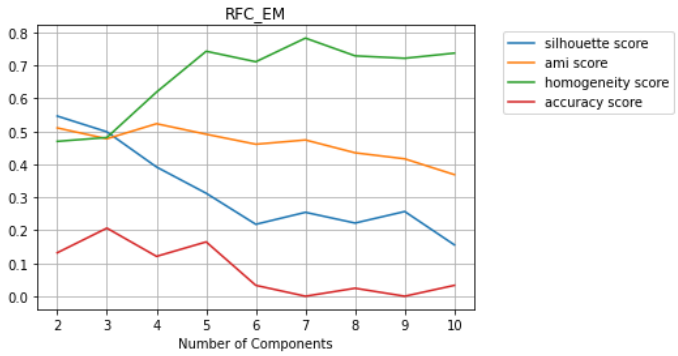
Number of components were chosen 28 for GRP for the lowest reconstruction error. The number of clusters were chosen 2 for GRP_K-means and GRP_EM but the two clusters were overlapped, which were very different from the original dataset. However, the accuracy of labeling for both were over 90%, 9 times of that for the original dataset.

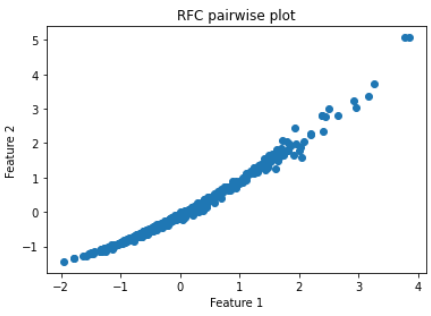
**RFC with clustering**

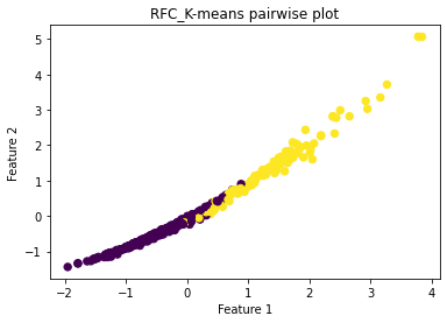
Feature importance was investigated and features with importance greater than 0.05 were chosen only, so the original dataset was reduced to 8 features.

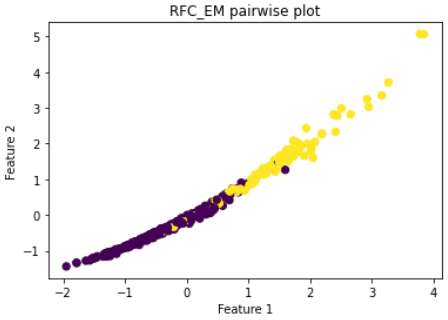



The number of clusters were both chosen 2 for RFC_K-means and RFC_EM and both can separate some of the points, but there was overlap between two clusters. The clusters were very different from the original dataset. The accuracy of labeling of both were very low (around 10%), similar with that of the original dataset.

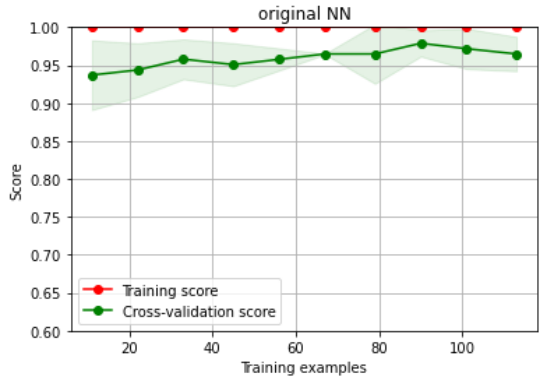
Comparing all four dimensionality reduction algorithms, for dataset 1, EM clustering on RFC reduced dataset achieved the highest accuracy of labeling and produced three separable clusters. For dataset 2, K-means clustering on PCA reduced dataset clearly separated the data points into two clusters and got the highest accuracy of labeling.

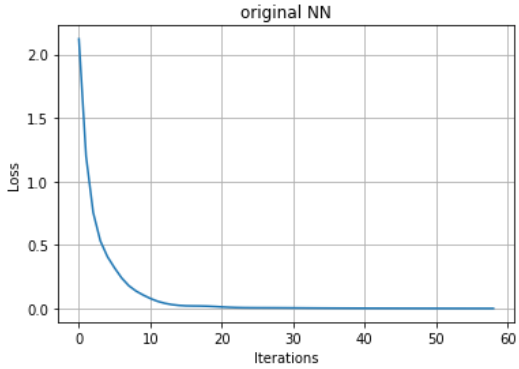
## Step 4 & 5: Neural network exploration on dimension reduced datasets

Dataset 1 was chosen for the evaluation of neural network (NN) algorithm. GridSearchCV was applied to find optimal parameters for each NN algorithm on different dimension reduced datasets by the four dimensionality reduction algorithms and two clustering algorithms.

5-fold cross validation was applied and learning curves and loss curves were plotted to explore the performance of the neural network algorithms. And training accuracy, test accuracy and fit time were compared between each algorithm. PCA_NN means NN applied on PCA reduced dataset, ICA_NN means NN applied on ICA reduced dataset, GRP_NN means NN applied on GRP reduced dataset, RFC_NN means NN applied on RFC dataset, K-means_NN means NN applied on k-means transformed dataset, and EM_NN means NN applied on EM transformed dataset.

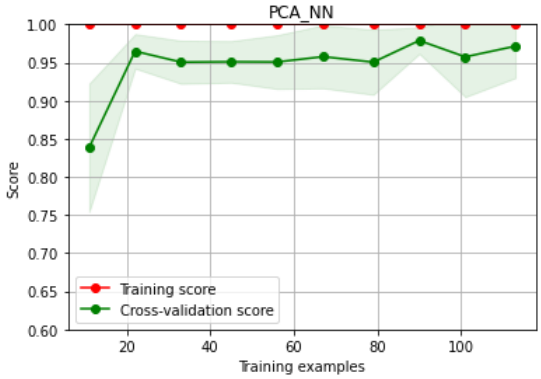
**Original no reduction dataset_NN:**

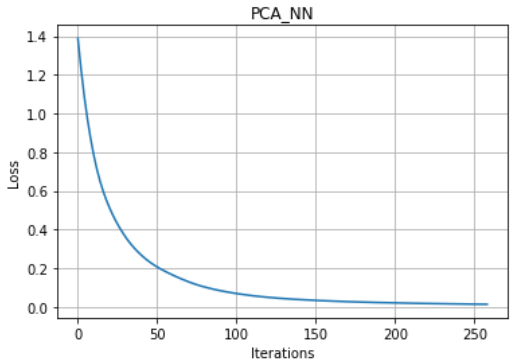



For the original no reduction dataset, the training accuracy was always 1.0 and validation accuracy increased from 0.93 to 0.97 and converged at last, which showed the model had a great performance. The loss decreased sharply at the beginning, after 10 iterations it was almost 0.

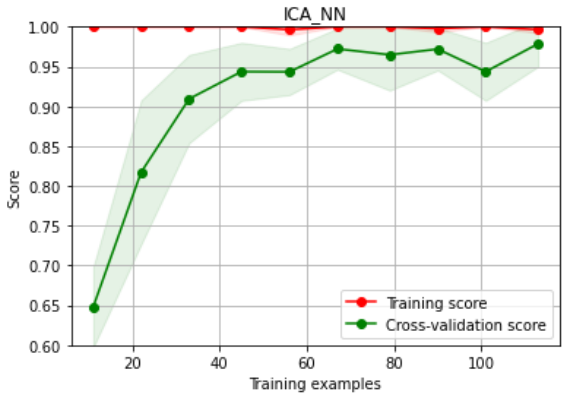
**PCA_NN**

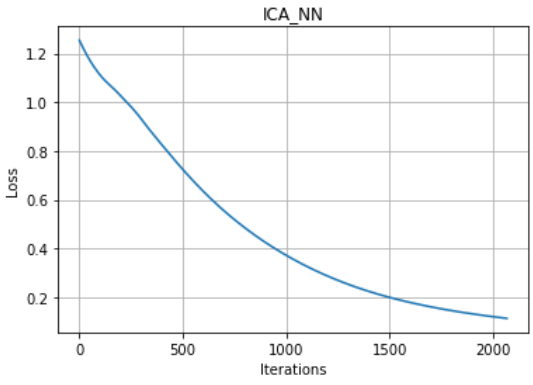



For the PCA reduced dataset, the training accuracy was always 1.0 and validation accuracy increased from 0.83 to 0.97. The loss decreased smoothly but was almost 0 until 200 iterations. This might be caused by the slower learning rate, which was 10 times less than the original dataset.

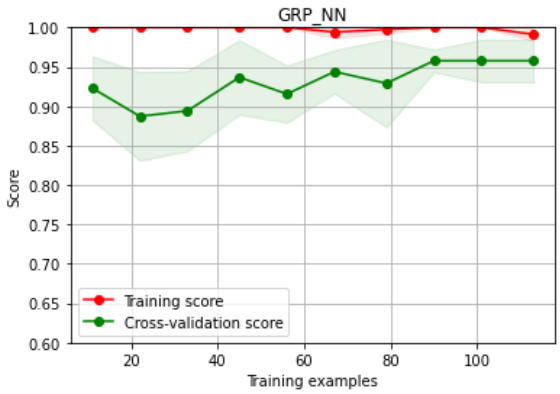
**ICA_NN**

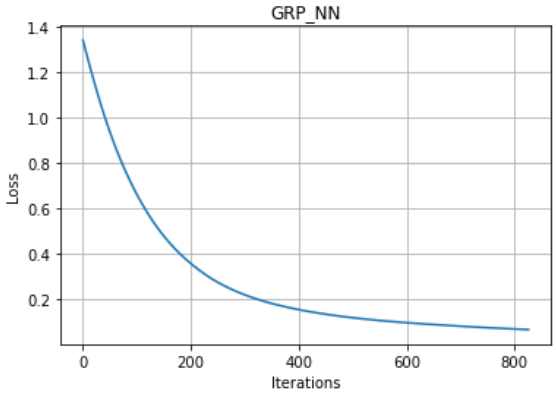



For the ICA reduced dataset, the training accuracy was always 1.0 and validation accuracy increased from 0.65 to 0.97. The loss decreased slowly and didn't achieve 0 even after 2000 iterations.

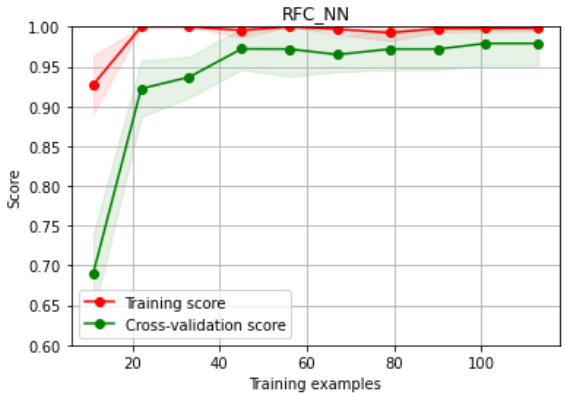
**GRP_NN**

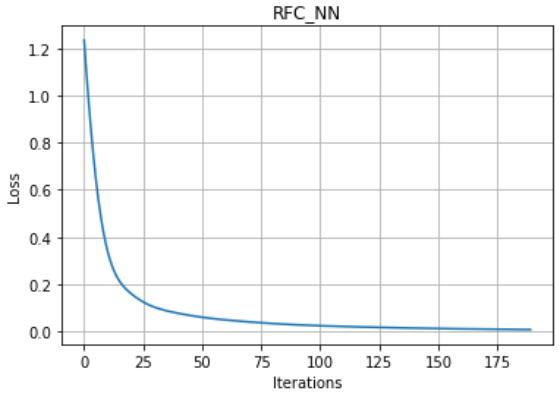



For the GRP reduced dataset, the training accuracy was almost 1.0 and validation accuracy went ups and downs but converged to 0.96. The loss decreased but didn't achieve 0 after 800 iterations, although the learning rate was 200 times less than the original dataset.

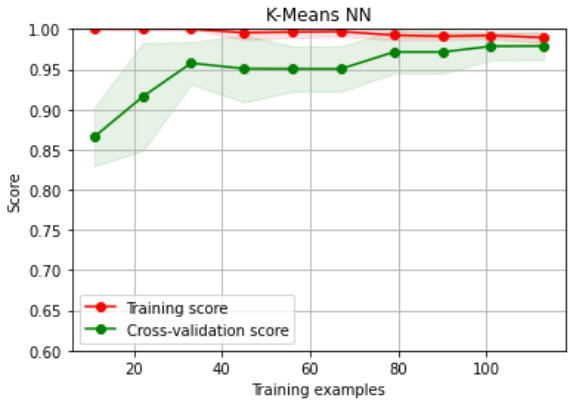
**RFC_NN**

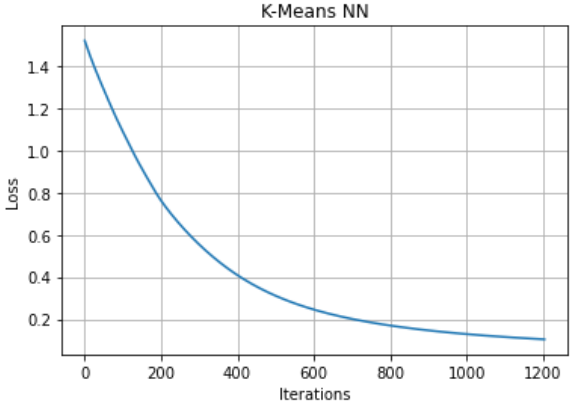



For the RFC reduced dataset, the training accuracy increased first and then was 1.0 and validation accuracy increased from 0.68 and converged to 0.97. The loss decreased fast and achieved 0 after 150 iterations. The learning rate was same as the PCA_NN, but loss decreased faster than it.

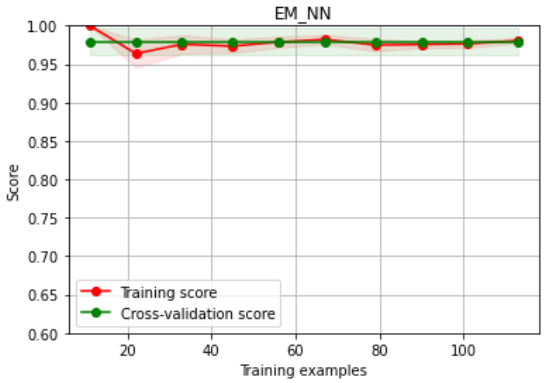
**K-means_NN**

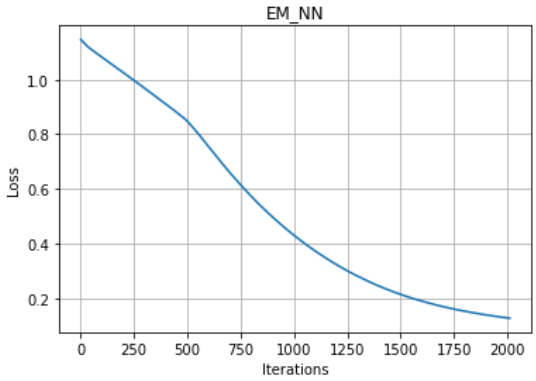



For the k-means reduced dataset, the training accuracy was almost 1.0 and validation accuracy increased from 0.87 and converged to 0.97. The loss decreased slowly and didn't achieve 0 after 1200 iterations.

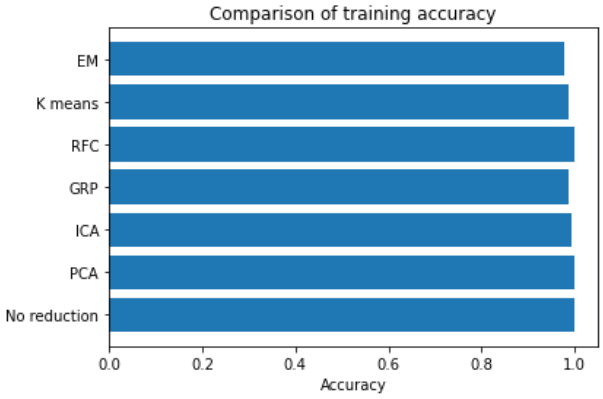
**EM_NN**

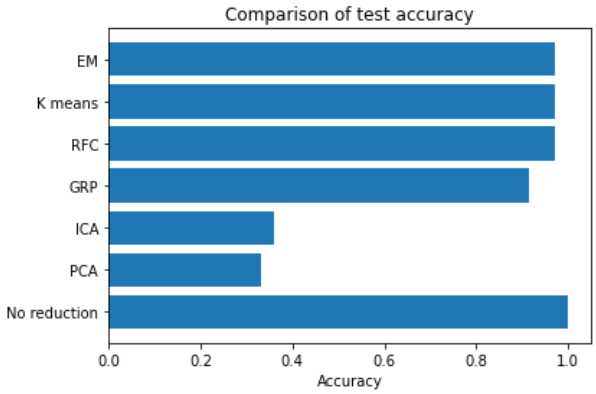


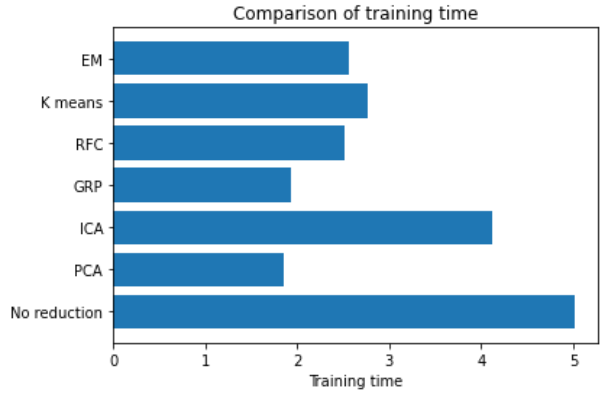




For the EM reduced dataset, the training and validation accuracy converged to 0.97 after running 30 samples, but the loss decreased very slowly and didn't get to 0 after 2000 iterations.

**Comparison**




The correlations between features were checked and it turned out there were no two features with correlation higher than 0.90. The original dataset was 13 dimensions, PCA reduced the dimension to 6, ICA and GRP reduced it to 11, RFC reduced it to 7.

Neural network algorithm on the original dataset and the six dimension-reduced datasets achieved similar training accuracy (0.97-1.0) and similar test accuracy (0.92-1.0) except for ICA and PCA reduced datasets. But PCA and GRP were fastest to fit the data, while no reduction was the slowest. NN on RFC reduced dataset performed the best in terms of loss curve, training accuracy, test accuracy, and training time, compared to all others.

**Reference:**


1. https://archive.ics.uci.edu/ml/datasets/Wine
2. https://scikitlearn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine
3. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html