Unsupervised Learning and Dimensionality Reduction

CS 7641 – Machine Learning

Manish Mehta

Introduction

The purpose of this project is to explore and apply some of the unsupervised learning and dimensionality reduction techniques for two new datasets and one older dataset (from assignment 1), with various combinations of applying clustering techniques on original data, applying dimensionality reduction techniques on original data, applying clustering post dimensionality reduction. We will then compare and contrast the results we get from our earlier experiments to the current one, and see if there are any improvements in accuracy, time, or any other metrics. For the purpose of this assignment, we have selected the data on "Digit Recognition", "Breast Cancer" and "Phishing Websites", as they are some of the most attractive areas of interest and have a large-scale impact, be it in healthcare domain or for any corporation and organization[1,2,3,4]. The aim is to understand, implement and analyze each of these clustering algorithms, k-Means and Expectation Maximization (EM) on the new datasets. Then, apply dimensionality reduction algorithms such as PCA, ICA, Randomized Projections (RP) and another feature selection algorithm (chosen Random Forest Classifier, RFC), and observe what we see. We also train the Neural Net classifier on the Phishing Dataset, after once we have performed dimensionality reduction on the data, and then once when we consider clustering labels as the new features in addition to the older ones for the Phishing Dataset and see if there is any improvement from assignment 1. Along with that, we also graphically report the training, testing accuracies, and fit times against number of iterations for each of the techniques for the sake of completion.

What are our datasets

Breast Cancer (BC) Wisconsin (Diagnostic) Dataset is used for the fact that given features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, the task is to predict the diagnosis of the tissue (benign or malignant). This has tremendous applications in healthcare, which have been outlined here[5]. The handwritten digits dataset poses the problem of recognizing handwritten letters from an image, in which the challenge is to learn information stored in the spatial domain. Computer Vision can be used, however it comes at the cost of simplicity. The problem is chosen because of the eigen digit images as in eigenfaces in face recognition, which can help in better clustering and dimensionality reduction. This data has 5620 rows and 65 features. The Phishing dataset arises from the need of several researchers who have issues obtaining a reliable dataset in the field of Phishing and Website Frauds. This dataset aims to bring to light the features that have proven to be successful in identifying and predicting the cases of phishing websites. Again, the uses have been outlined in the research papers in references section.

Experiment Methodology and Analysis

a. Run the clustering algorithms on the datasets and describe what you see.
b. Apply the dimensionality reduction algorithms to the two datasets and describe what you see.
c. Reproduce your clustering experiments, but on the data after you've run dimensionality reduction on it.
d. Apply the dimensionality reduction algorithms to one of your datasets from assignment #1 and rerun your neural network learner on the newly projected data.
e. Apply the clustering algorithms to the same dataset to which you just applied the dimensionality reduction algorithms, treating the clusters as if they were new features. In other words, treat the clustering algorithms as

if they were dimensionality reduction algorithms. Again, rerun your neural network learner on the newly projected data.

Part 1: Clustering Techniques

After preprocessing and scaling the data, k-Means and EM algorithms were implemented, and their results observed.

Part 1.1: k-Means Clustering

Our aim was to observe the best number of clusters for each dataset using the elbow inspection method on Within-Cluster sum of squares and also on Homogeneity score, describe what attributes make up each cluster and whether it performed well in comparison to original labeling (since we have the true labels for each dataset, but we will provide this analysis for only Phishing Dataset, since this is the dataset which was used in the earlier assignments too and it makes the most sense to comment on the same). Since k-Means is susceptible to get stuck in local optima due to the random selection of initial cluster centers, average metrics over 5 models for each number of k clusters have been reported. The homogeneity score (the closer to 1.0, the better) that describes how each cluster contains members from a single class was used as the ground truth for understanding how well the algo performed, and for sake of completeness, the silhouette score (closer to 1, the better) has also been reported. The plots for training times was not shown, but the training time average has been reported in seconds. The reason it is not shown is because of sake of compactness, and the fact that it always increases with increase in number of clusters, iterations, components, etc.

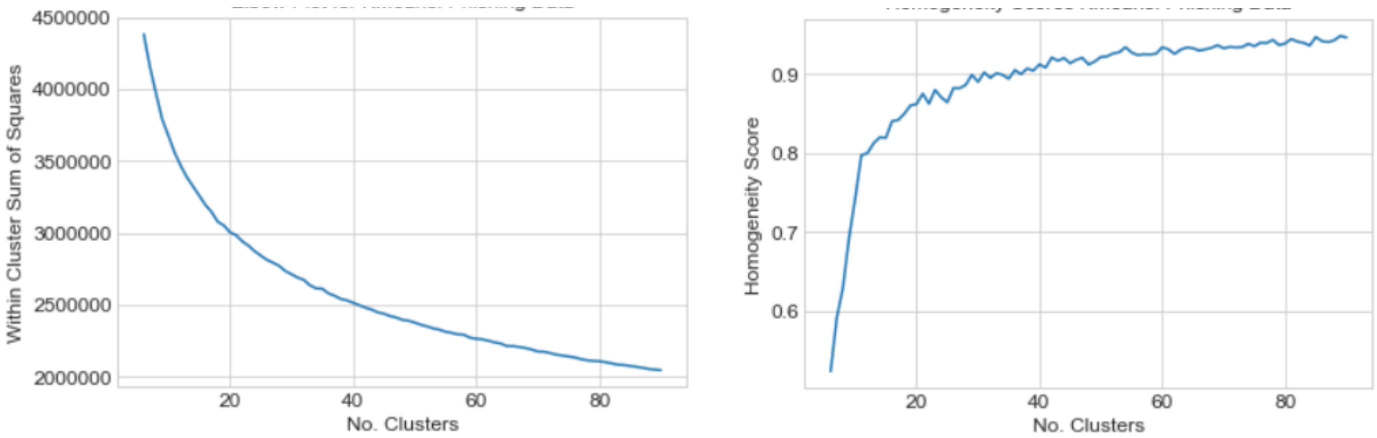


Fig 1: Elbow method for k-Means, WCSS and Homogeneity Score for Handwritten Digits data

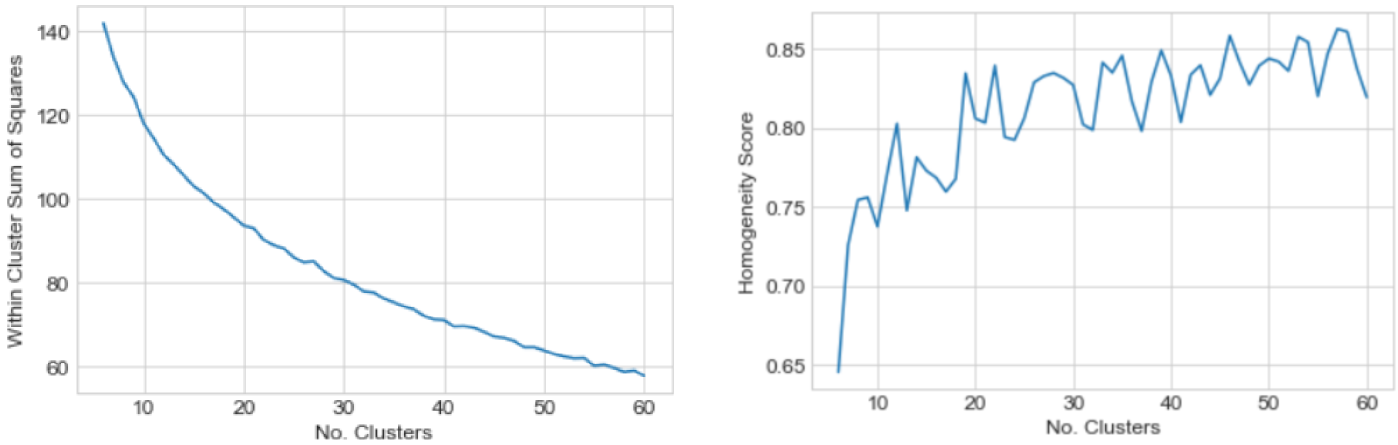
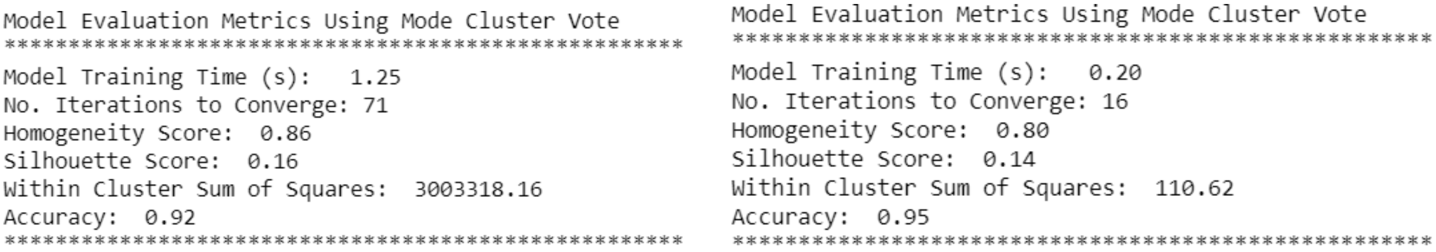


Fig 2: Elbow method for k-Means, WCSS and Homogeneity Score for Breast Cancer (BC) data

Using Elbow criterion, we can see the optimal number of clusters for Digits data = 20, and for BC data = 12. For final run:



```
Model Evaluation Metrics Using Mode Cluster Vote
****************************************************
Model Training Time (s):   1.25
No. Iterations to Converge: 71
Homogeneity Score:  0.86
Silhouette Score:  0.16
Within Cluster Sum of Squares:  3003318.16
Accuracy:  0.92
****************************************************
```

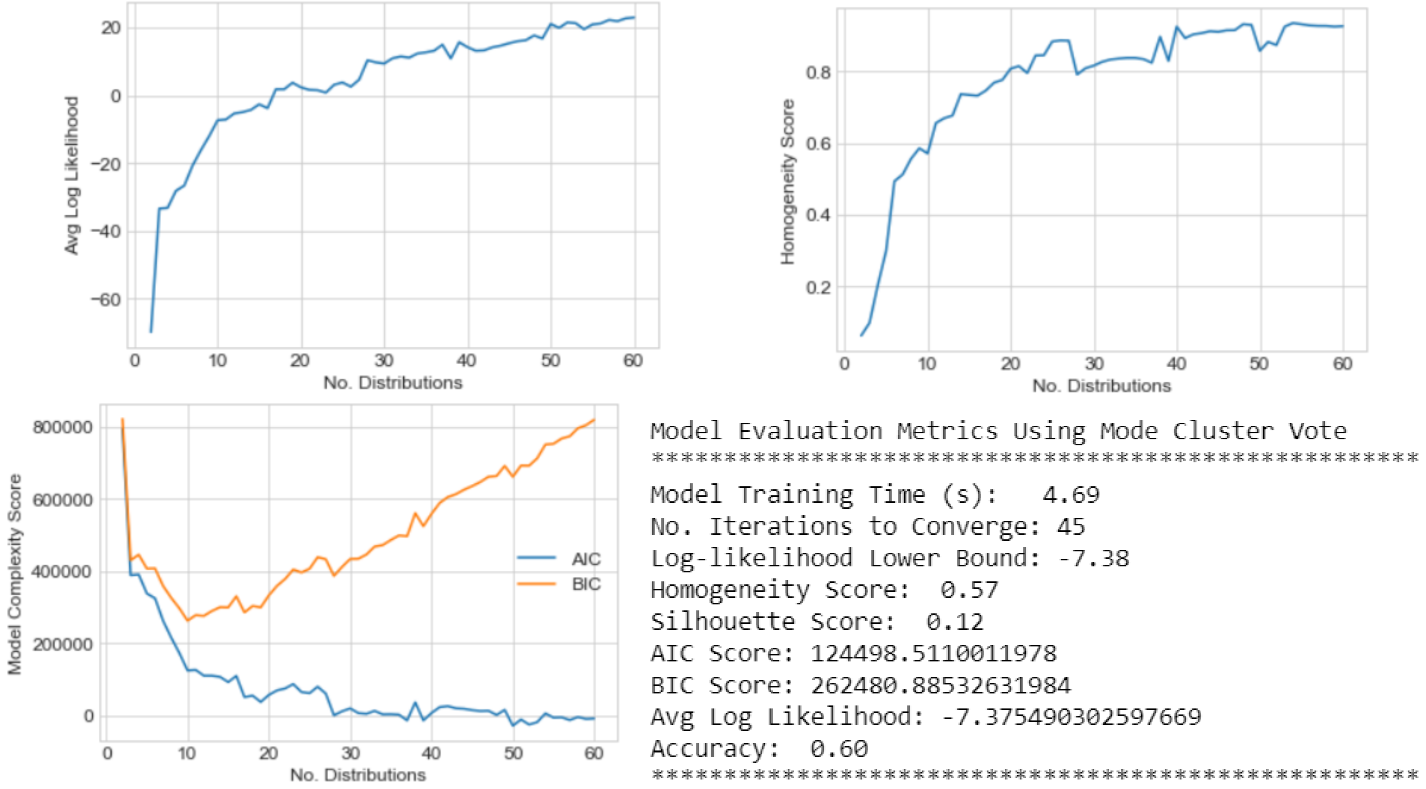
```
Model Evaluation Metrics Using Mode Cluster Vote
****************************************************
Model Training Time (s):   0.20
No. Iterations to Converge: 16
Homogeneity Score:  0.80
Silhouette Score:  0.14
Within Cluster Sum of Squares:  110.62
Accuracy:  0.95
****************************************************
```

Fig 3: Results from final run for Digits and Breast Cancer data respectively (left to right)

The Accuracy signifies how well k-Means performed with respect to the original labels, and high homogeneity scores indicate better clustering, though there may be overlaps in the clusters due to which the silhouette scores are low. But for all practical purposes, we will solely focus on Homogeneity score as our ground metric. The algo used k-means++ as its method of initialization, since it was observed to perform best and speeds up convergence. The similarity/ distance metric used was Eucledian, since it gave the best clustering and highest accuracy across various other metrics that were tried ('L1', 'L2', 'Manhattan', etc.)

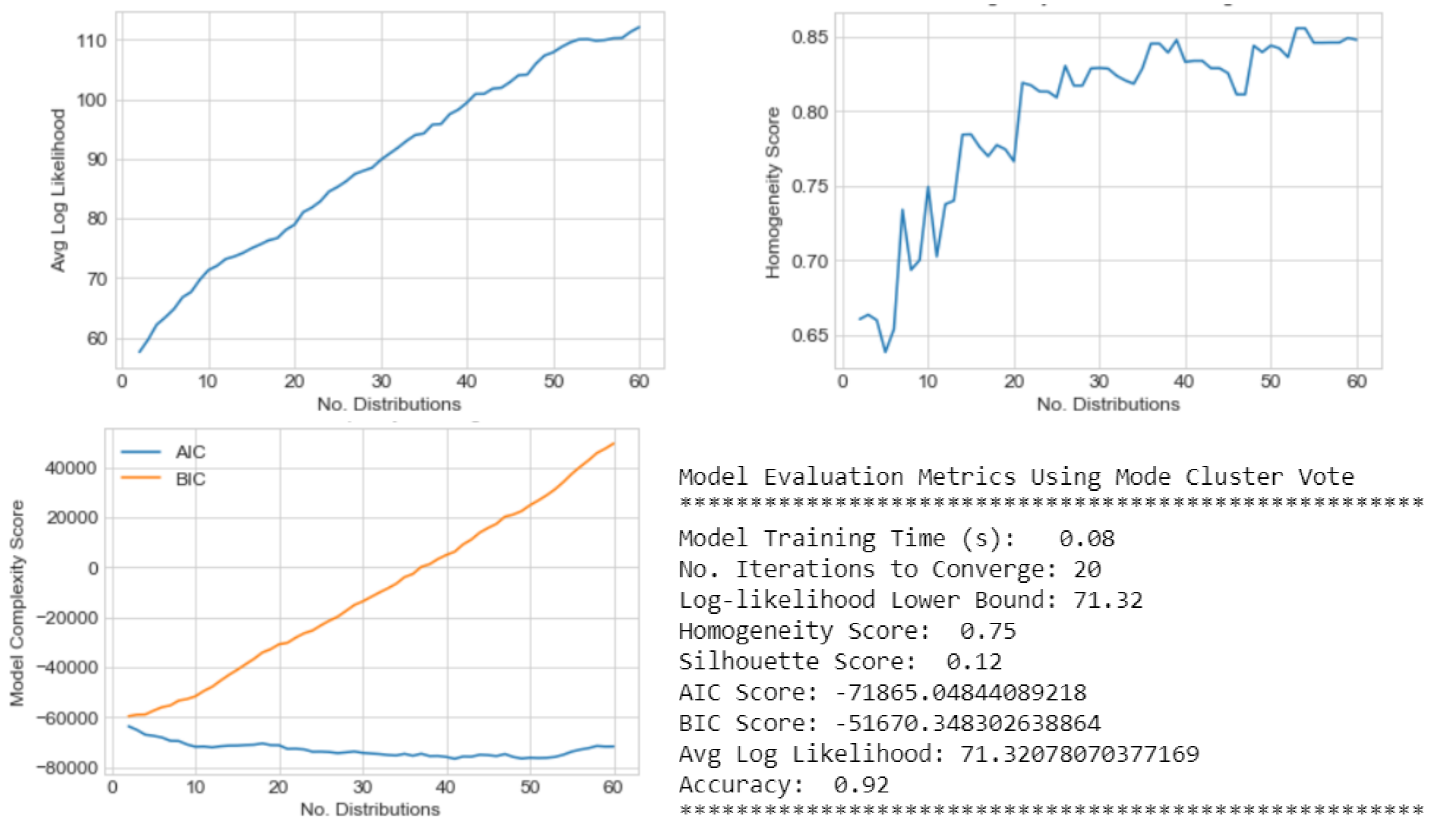
Part 1.2: Expectation Maximization

Gaussian Mixture model was used which assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. It implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. The number of components and the covariance type parameter are tuned using grid search and the optimal model is selected using Elbow criterion on Avg. log likelihood, along with observing Homogeneity score and Akaike and Bayesian information criterion (AIC/BIC) score which is an estimate of a function of the posterior probability of a model being true. BIC tries to maximize log-likelihood of data while penalizing the number of parameters in the model to prevent overfitting. In theory, it recovers the true number of components in the asymptotic regime where the data contains an infinite number of samples generated iid from a Gaussian mixture. Lower BIC values are better.



```
Model Evaluation Metrics Using Mode Cluster Vote
****************************************************
Model Training Time (s):   4.69
No. Iterations to Converge: 45
Log-likelihood Lower Bound: -7.38
Homogeneity Score:  0.57
Silhouette Score:  0.12
AIC Score: 124498.5110011978
BIC Score: 262480.88532631984
Avg Log Likelihood: -7.375490302597669
Accuracy:  0.60
****************************************************
```

Fig 4: Avg. Log-Likelihood, Homogeneity score, and AIC, BIC for Digits data vs No. of distributions with results for final run



```
Model Evaluation Metrics Using Mode Cluster Vote
****************************************************
Model Training Time (s):   0.08
No. Iterations to Converge: 20
Log-likelihood Lower Bound: 71.32
Homogeneity Score:  0.75
Silhouette Score:  0.12
AIC Score: -71865.04844089218
BIC Score: -51670.348302638864
Avg Log Likelihood: 71.32078070377169
Accuracy:  0.92
****************************************************
```

Fig 5: Avg. Log-Likelihood, Homogeneity score, and AIC, BIC for BC data vs No. of distributions with results for final run

Here, we can see that optimal no. of components = 10 for digits data and breast cancer data, and the results for final run are also shown in the figure. Here, we see that EM does not perform so well on the digits data overall since its accuracy and homogeneity scores were low but the training time is high. This might be because of selecting lesser number of components, thus having a tradeoff with BIC. Even though on the plot for breast cancer the BIC seems to be increasing, the AIC keeps going down and we attain an accuracy of 92% with respect to the original labels. The training time is also low for the EM for breast cancer data. The covariance type used was 'full' since it provided with the best results when each component had its own covariance matrix, as compared to 'diag' and 'tied'. Silhouette Coefficient scores cannot be used to evaluate the performance of GMMs because they work well only on spherical clusters and GMMs do not necessarily produce spherical clusters. Overall, the results indicate that many samples are close to the decision boundary between neighboring clusters and the clusters aren't well-separated. The absence of a prominent elbow suggested that k-means may not be able to find good clusters and this observation reinforces the point. The mediocre clustering performance might be because of outliers in the data or convergence to a local optimum.

Part 2: Dimensionality Reduction

This section will implement 4 different dimensionality reduction techniques on the new datasets. Then, k-means and EM clustering will be performed for each dataset and dimensionality reduction combination to see how the clustering compares with using the full datasets. The 4 dimensionality reduction techniques are:

a. Principal Components Analysis (PCA). Optimal number of PC chosen by inspecting % variance explained and the eigenvalues.
b. Independent Components Analysis (ICA). Optimal number of IC chosen by inspecting kurtosis.
c. Random Components Analysis (RCA) (otherwise known as Randomized Projections). Optimal number of RC chosen by inspecting reconstruction error.
d. Random Forest Classifier (RFC). Optimal number of components chosen by feature importance.

In several of the plots below, the questions like distribution of eigenvalues for PCA, how kurtotic the distributions are for ICA, how well is the data reconstructed by using randomized projections, etc. have been answered via various plots. For the sake of compactness and readability, since we are using 3 datasets and repeating a whole bunch of experiments on each of them, not every plot will be provided with a footnote. For RFC, feature importance was used as a criteria for selecting the features, the higher the better. Since that was done using array manipulation, the graphs have not been shown. Also, RP was run multiple times, and here we provide the average plots of several restarts.

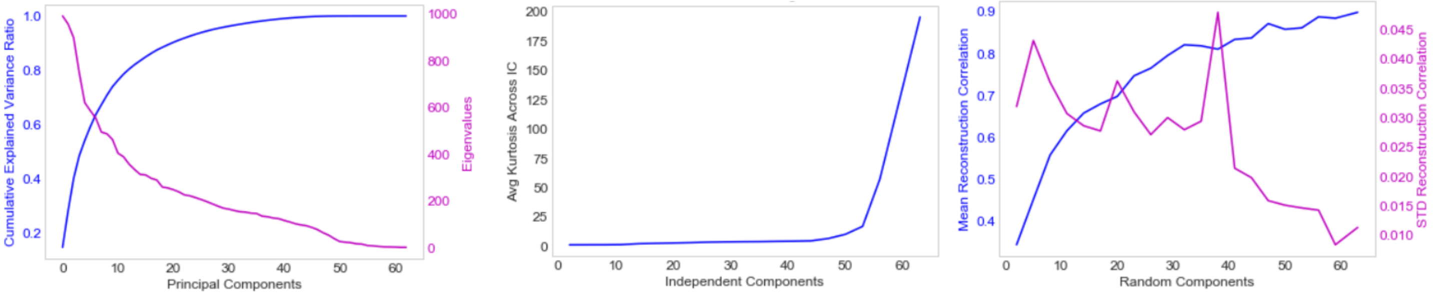


Fig 6: Cumulative Explained variance, Avg. kurtosis, and Mean Reconstruction Correlation vs No. of components for Digits

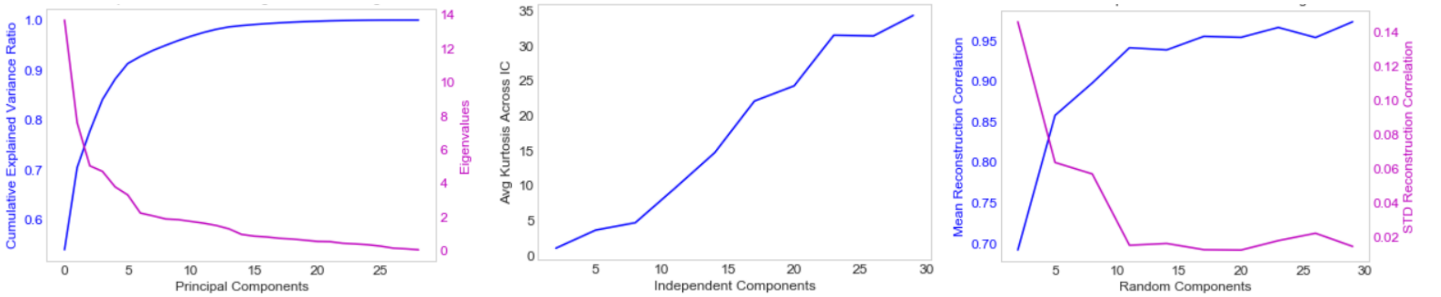
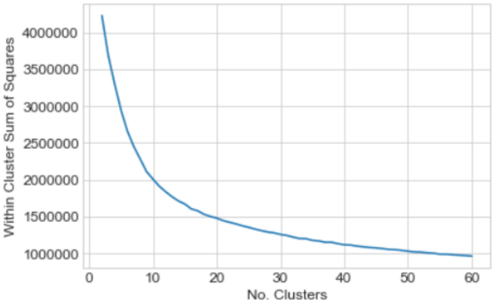
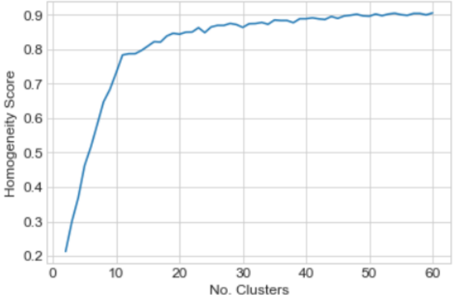
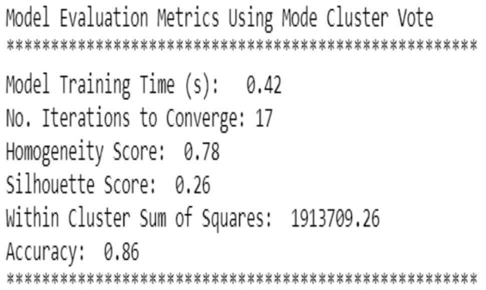


Fig 7: Cumulative Explained variance, Avg. kurtosis, and Mean Reconstruction Correlation vs No. of components for BC data

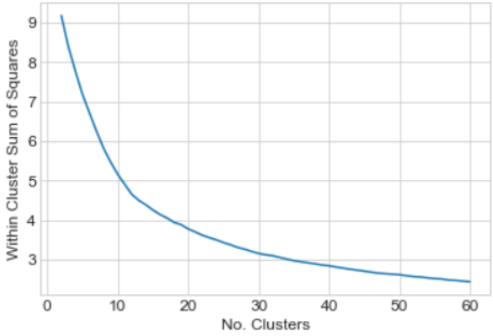
For **k-Means and EM clustering** post dimensionality reduction, choosing # components = 10 for each of PCA, ICA and RCA for Digits data and # components = 4, 5, 6 for BC data, we get the following:

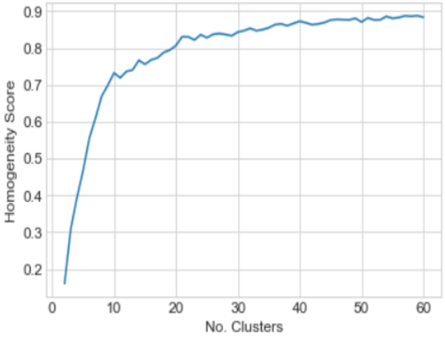


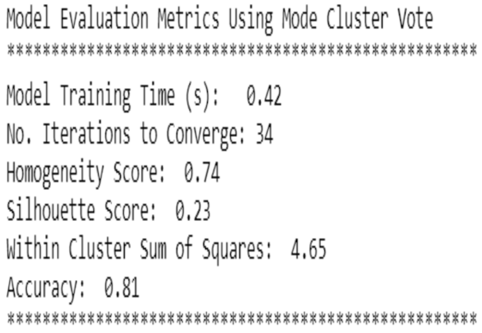




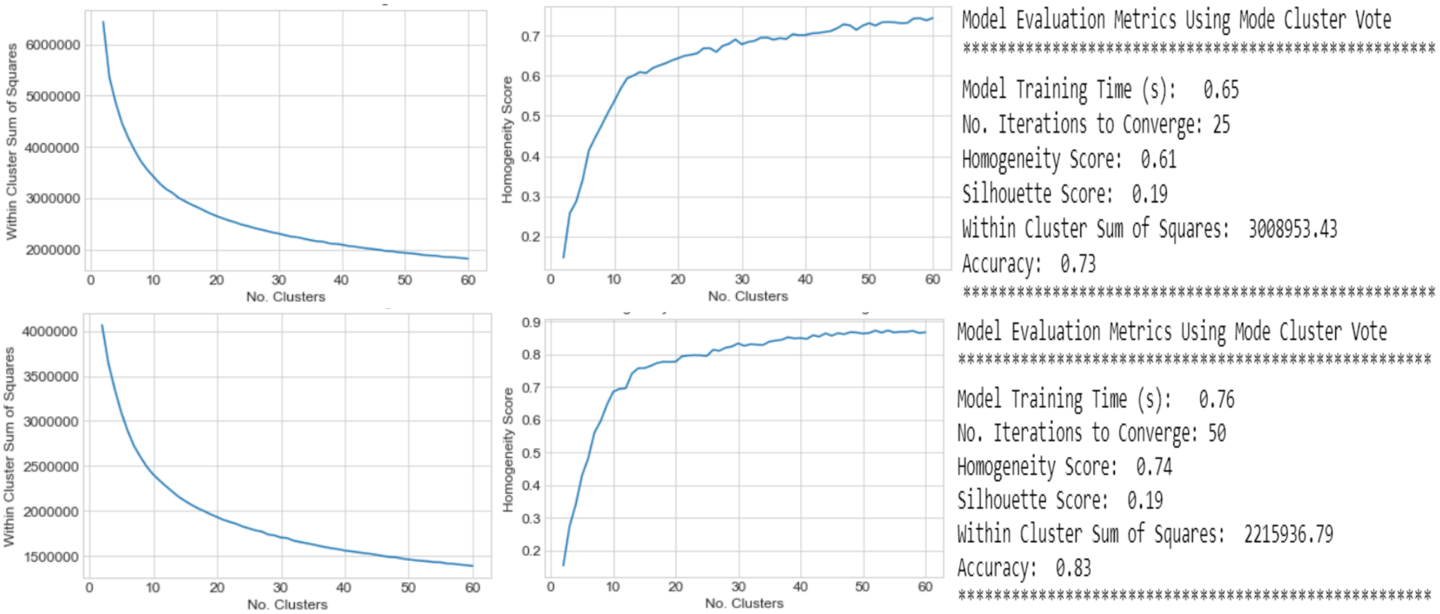
```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************

Model Training Time (s):   0.42
No. Iterations to Converge: 17
Homogeneity Score:  0.78
Silhouette Score:  0.26
Within Cluster Sum of Squares:  1913709.26
Accuracy:  0.86
*****************************************************
```







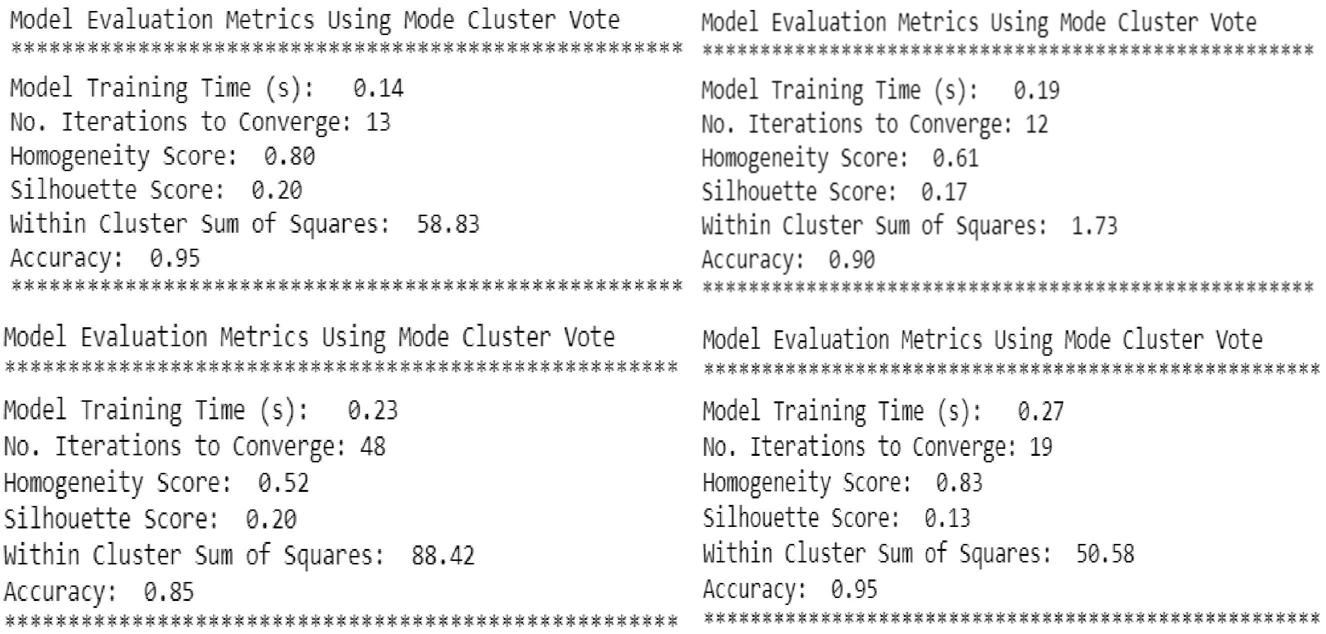
```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************

Model Training Time (s):   0.42
No. Iterations to Converge: 34
Homogeneity Score:  0.74
Silhouette Score:  0.23
Within Cluster Sum of Squares:  4.65
Accuracy:  0.81
*****************************************************
```



```
Model Evaluation Metrics Using Mode Cluster Vote
**************************************************
Model Training Time (s):   0.65
No. Iterations to Converge: 25
Homogeneity Score:  0.61
Silhouette Score:  0.19
Within Cluster Sum of Squares:  3008953.43
Accuracy:  0.73
**************************************************
```

```
Model Evaluation Metrics Using Mode Cluster Vote
**************************************************
Model Training Time (s):   0.76
No. Iterations to Converge: 50
Homogeneity Score:  0.74
Silhouette Score:  0.19
Within Cluster Sum of Squares:  2215936.79
Accuracy:  0.83
**************************************************
```

Fig 8: WCSS, Homogeneity Scores, and Final results for running **k-means** on dimensionally reduced **Digits data** after PCA, ICA, RCA, RFC (top row to bottom row) with appropriate number of clusters selected for each (11, 12, 14, 13 respectively)



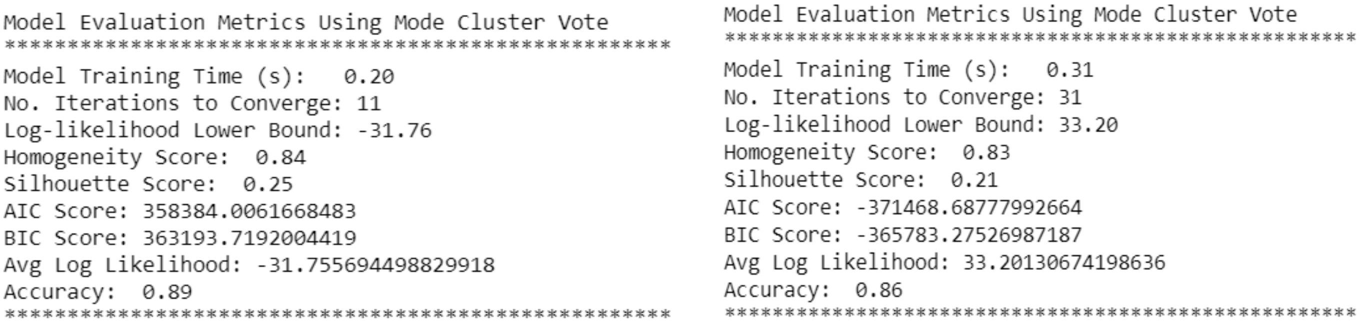
```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.14
No. Iterations to Converge: 13
Homogeneity Score:  0.80
Silhouette Score:  0.20
Within Cluster Sum of Squares:  58.83
Accuracy:  0.95
*****************************************************
```

```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.19
No. Iterations to Converge: 12
Homogeneity Score:  0.61
Silhouette Score:  0.17
Within Cluster Sum of Squares:  1.73
Accuracy:  0.90
*****************************************************
```

```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.23
No. Iterations to Converge: 48
Homogeneity Score:  0.52
Silhouette Score:  0.20
Within Cluster Sum of Squares:  88.42
Accuracy:  0.85
*****************************************************
```

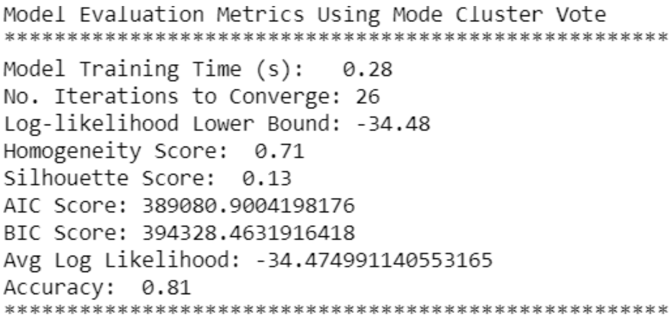
```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.27
No. Iterations to Converge: 19
Homogeneity Score:  0.83
Silhouette Score:  0.13
Within Cluster Sum of Squares:  50.58
Accuracy:  0.95
*****************************************************
```

Fig 9: Final results for running **k-means clustering** on dimensionally reduced **BC data** after PCA, ICA, RCA, RFC (left to right) with appropriate number of clusters selected for each (13, 14, 13, 19 respectively)

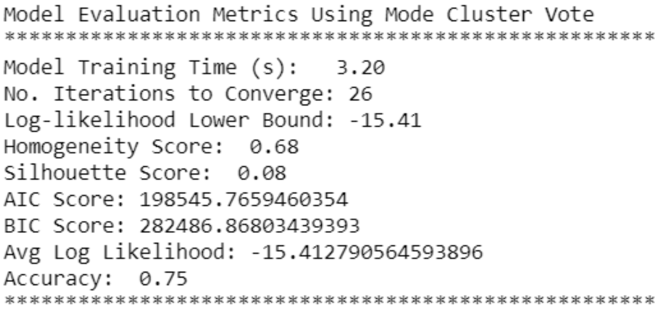


```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.20
No. Iterations to Converge: 11
Log-likelihood Lower Bound: -31.76
Homogeneity Score:  0.84
Silhouette Score:  0.25
AIC Score: 358384.0061668483
BIC Score: 363193.7192004419
Avg Log Likelihood: -31.755694498829918
Accuracy:  0.89
*****************************************************
```

```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.31
No. Iterations to Converge: 31
Log-likelihood Lower Bound: 33.20
Homogeneity Score:  0.83
Silhouette Score:  0.21
AIC Score: -371468.68777992664
BIC Score: -365783.27526987187
Avg Log Likelihood: 33.20130674198636
Accuracy:  0.86
*****************************************************
```

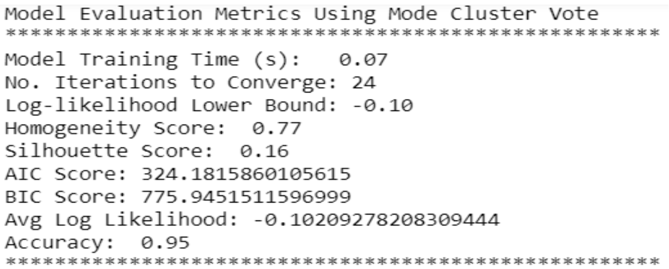


```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.28
No. Iterations to Converge: 26
Log-likelihood Lower Bound: -34.48
Homogeneity Score:  0.71
Silhouette Score:  0.13
AIC Score: 389080.9004198176
BIC Score: 394328.4631916418
Avg Log Likelihood: -34.474991140553165
Accuracy:  0.81
*****************************************************
```

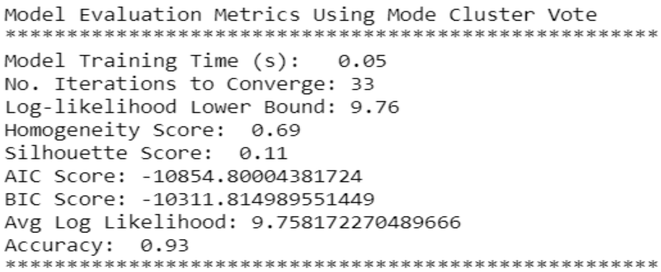


```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   3.20
No. Iterations to Converge: 26
Log-likelihood Lower Bound: -15.41
Homogeneity Score:  0.68
Silhouette Score:  0.08
AIC Score: 198545.7659460354
BIC Score: 282486.86803439393
Avg Log Likelihood: -15.412790564593896
Accuracy:  0.75
*****************************************************
```

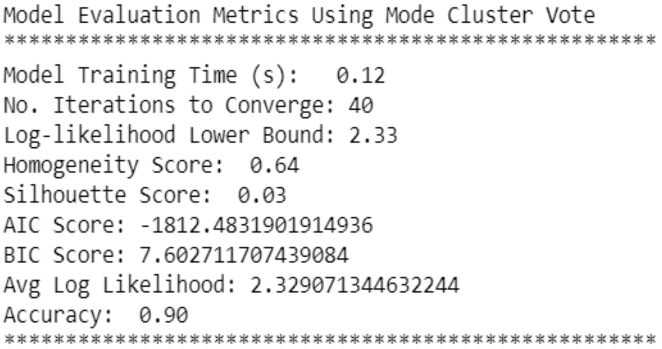
Fig 10: Final results for running **EM** on dimensionally reduced **Digits data** after PCA, ICA, RCA, RFC (left to right) with appropriate number of clusters selected for each (11, 13, 12, 19 respectively)



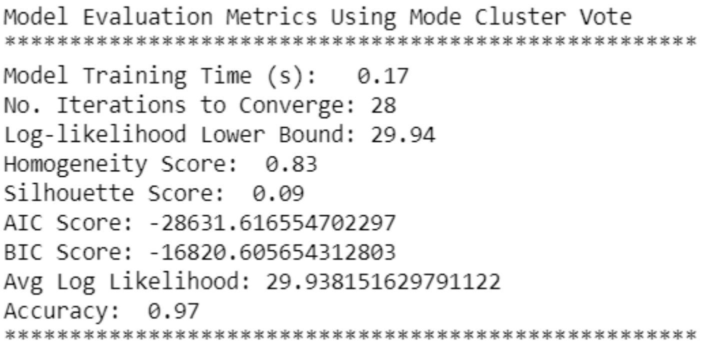
```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.07
No. Iterations to Converge: 24
Log-likelihood Lower Bound: -0.10
Homogeneity Score:  0.77
Silhouette Score:  0.16
AIC Score: 324.1815860105615
BIC Score: 775.9451511596999
Avg Log Likelihood: -0.10209278208309444
Accuracy:  0.95
*****************************************************
```



```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.05
No. Iterations to Converge: 33
Log-likelihood Lower Bound: 9.76
Homogeneity Score:  0.69
Silhouette Score:  0.11
AIC Score: -10854.80004381724
BIC Score: -10311.814989551449
Avg Log Likelihood: 9.758172270489666
Accuracy:  0.93
*****************************************************
```



```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.12
No. Iterations to Converge: 40
Log-likelihood Lower Bound: 2.33
Homogeneity Score:  0.64
Silhouette Score:  0.03
AIC Score: -1812.4831901914936
BIC Score: 7.602711707439084
Avg Log Likelihood: 2.329071344632244
Accuracy:  0.90
*****************************************************
```



```
Model Evaluation Metrics Using Mode Cluster Vote
*****************************************************
Model Training Time (s):   0.17
No. Iterations to Converge: 28
Log-likelihood Lower Bound: 29.94
Homogeneity Score:  0.83
Silhouette Score:  0.09
AIC Score: -28631.616554702297
BIC Score: -16820.605654312803
Avg Log Likelihood: 29.938151629791122
Accuracy:  0.97
*****************************************************
```

Fig 11: Final results for running **EM** on dimensionally reduced **BC data** after PCA, ICA, RCA, RFC (left to right) with appropriate number of clusters selected for each (7, 6, 15, 20 respectively)

Tabulated results and comparisons are shown below:

k-means Digits	**Full Data**	**PCA**	**ICA**	**RCA**	**RFC**
Model Training Time (s)	1.25	0.42	0.42	0.65	0.76
# iterations	71	17	34	25	50
Homogeneity Score	0.86	0.78	0.74	0.61	0.74
Accuracy	0.92	0.86	0.81	0.73	0.83

k-means Breast Cancer	**Full Data**	**PCA**	**ICA**	**RCA**	**RFC**
Model Training Time (s)	0.20	0.14	0.19	0.23	0.27
# iterations	16	13	12	48	19
Homogeneity Score	0.80	0.80	0.61	0.52	0.83
Accuracy	0.95	0.95	0.90	0.85	0.95

EM Digits	**Full Data**	**PCA**	**ICA**	**RCA**	**RFC**
Model Training Time (s)	4.69	0.20	0.31	0.28	3.20
# iterations	45	11	31	26	26
Homogeneity Score	0.57	0.84	0.83	0.71	0.68

Accuracy	0.60	0.89	0.86	0.81	0.75

EM Breast Cancer	Full Data	PCA	ICA	RCA	RFC
Model Training Time (s)	0.08	0.07	0.05	0.12	0.17
# iterations	20	24	33	40	28
Homogeneity Score	0.75	0.77	0.69	0.64	0.83
Accuracy	0.92	0.95	0.93	0.90	0.97

Overall, we observe that PCA performed better for both the clustering algorithms for Digits data, and RFC performed better for both clustering algorithms for Breast Cancer data, mainly due to the nature of the problems. RFC usually takes the most time for Breast Cancer data but performs better. PCA takes the least time and gives the best clusters for Digits data. PCA performs best overall, since it requires least number of components for explaining the overall variance better in the data, and that is due to the nature of datasets at hand (repetitive graphs avoided)

Part 3: Dimensionality Reduction and Clustering for Phishing Dataset

For this part, we perform tasks d and e from our Experiment Methodology section, and avoid any repetitive graphs. After preprocessing the phishing data, we perform the required analysis on the same. Running **dimensionality reduction algorithms**, we get the below plots:

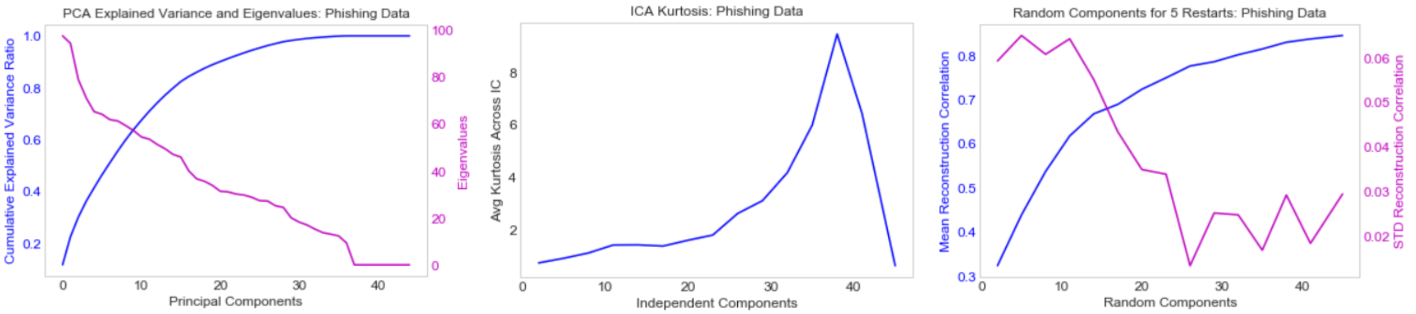


Fig 12: Running PCA, ICA, RCA on Phishing Data for optimal number of components for each case (15, 11, 13 respectively)

Hyperparameter tuning provided the optimal layer structure to be two layers with 50 and 25 units in the two hidden layers and a learning rate of 0.01. Below were the learning curves obtained along with results:

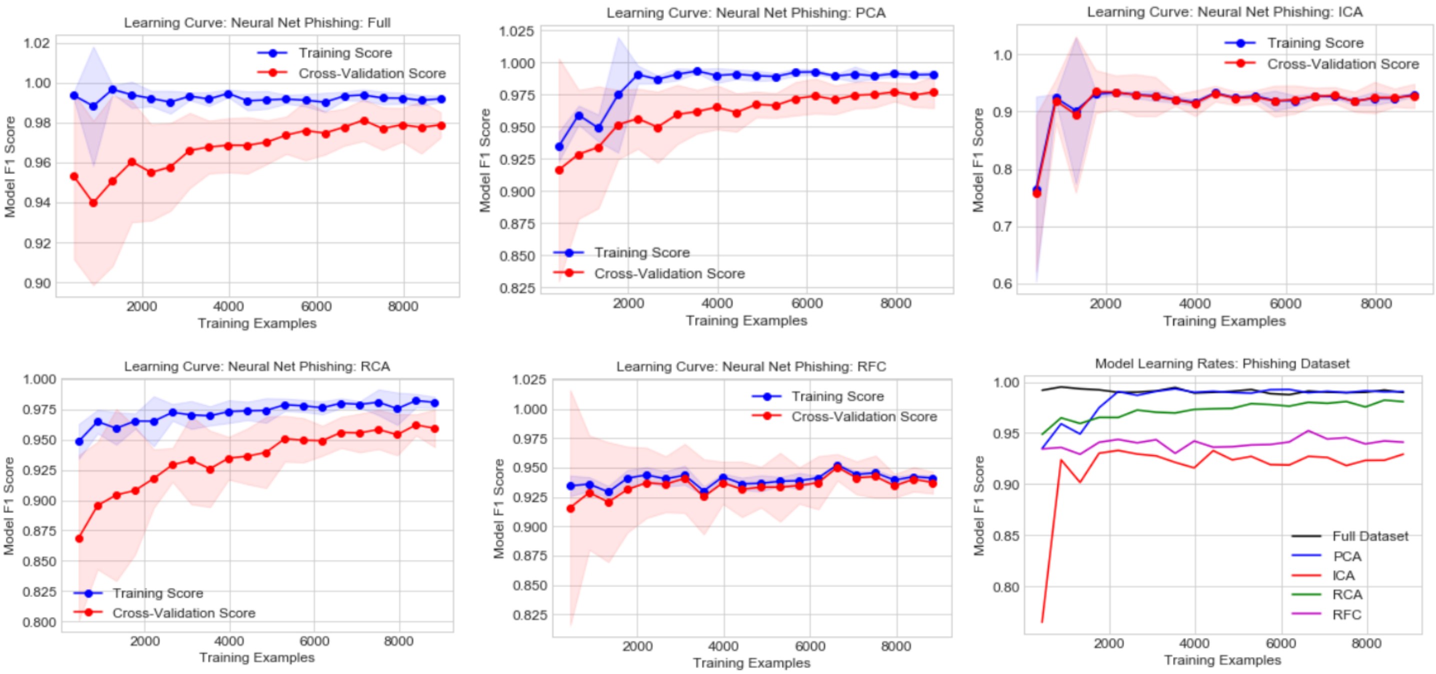


Fig 13: Neural Net performance for Full Data vs Dimensionally reduced data, and overall plot at the end

Metric	Full Data	PCA	ICA	RCA	RFC
Model Training Time (s)	7.79915	7.07616	1.70386	7.66016	4.44409
Model Prediction Time (s)	0.00322	0.00272	0.00220	0.00355	0.00217
F1 Score	0.97	0.96	0.93	0.95	0.94
Accuracy	0.97	0.96	0.93	0.94	0.93
AUC	0.96	0.96	0.92	0.94	0.92

Here, we can observe that ICA has the least training time, and provides reasonable F1 score, while the original dataset still has the maximum F1 score. PCA on the other hand takes lesser training time than Full dataset, and produces almost the same F1 score, since it can explain more than 80% cumulative variance (when reconstructed with 15 components only).

Running **clustering algorithms** and treating the cluster labels as the new features, we get the new dataset (original data + k-means and EM labels) on which we trained the Neural Net on. When clustering was run, for k-means the optimal clusters=10, training time was 1.46s, number of iterations=28, homogeneity score=0.61 and accuracy=0.91. For EM, it was no. of clusters=11, covariance type='full', training time=2.47s, homogeneity score=0.54 and accuracy=0.88. Thus, k-means labels are more inline with original labels. In addition (not asked in question), we also observed how dimensionality reduction algorithms performed with new data having the cluster labels as new features (both k-means and EM labels are being used as the new features). We have the following plots:

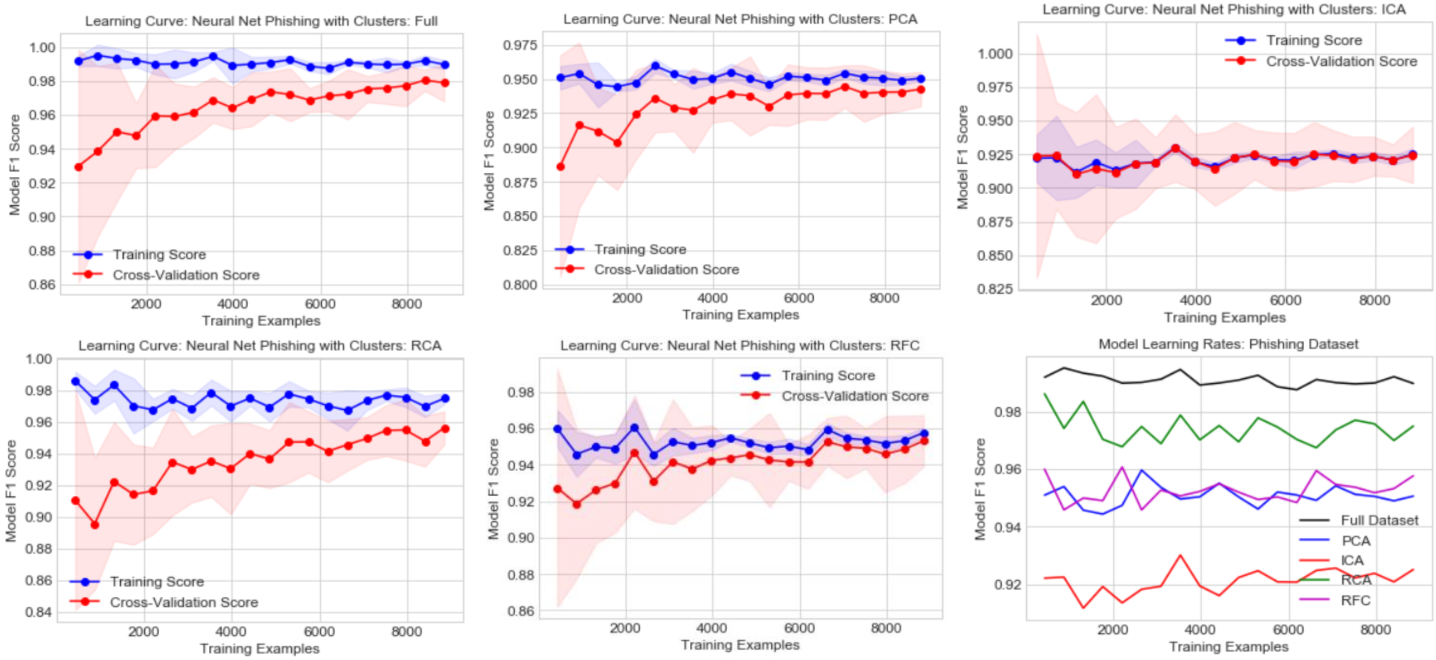


Fig 14: Neural Net performance for New Data vs Dimensionally reduced new data, and overall plot at the end

Metric	Original Data	New Data	PCA (New Data)	ICA (New Data)	RCA (New Data)	RFC (New Data)
Model Train Time (s)	7.79915	5.87178	4.97002	6.97211	5.90095	7.65430
Model Predict Time (s)	0.00322	0.00291	0.00187	0.00279	0.00416	0.00320
F1 Score	0.97	0.98	0.94	0.92	0.95	0.94
Accuracy	0.97	0.98	0.93	0.91	0.94	0.94
AUC	0.96	0.98	0.92	0.91	0.94	0.93

As we can see, adding the cluster labels as the new features and training Neural net on the same improved the F1 score of the model. This was because the predicted cluster labels were about 91% in line with actual labels via k-means clustering and about 88% in line with actual labels via EM. Also, PCA performs better overall with the new data, because of the inclusion of the two newly added cluster labels as features. Training time also reduces. ICA and RFC performed worse, because of the correlation between the two new features, but PCA could easily account for that. But, as we can see, RCA performs well on an average and is good at handling the correlations among features in the data. This is a great example of the scenarios where we should or should not keep the correlated features, and how does it affect in each case.

References

[1] Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi (2012) An Assessment of Features Related to Phishing Websites using an Automated Technique. In: International Conferece For Internet Technology And Secured Transactions. ICITST 2012 . IEEE, London, UK, pp. 492-497. ISBN 978-1-4673-5325-0

[2] Mohammad, Rami, Thabtah, Fadi Abdeljaber and McCluskey, T.L. (2014) Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications, 25 (2). pp. 443-458. ISSN 0941-0643

[3] Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi Abdeljaber (2014) Intelligent Rule based Phishing Websites Classification. IET Information Security, 8 (3). pp. 153-160. ISSN 1751-8709

[4] Dataset obtained from UCI and OpenML Repository:

- Handwritten Digits Data – OpenML
- Breast Cancer Wisconsin Data – OpenML
- Phishing Websites Data – UCI Repository, OpenML

[5] W. Nick Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In Biomedical Image Processing and Biomedical Visualization, volume 1905, pages 861–871. International Society for Optics and Photonics, July 1993.