Assignment 3

Name: Amr Ali
gtID: 903850635

All code and data present at: https://github.com/aali343/as3

Introduction

In this assignment, we'll be exploring several unsupervised learning techniques, as well as algorithms for dimensionality reduction.

Datasets

For more convenience in evaluating performance in a comparative manner, we will be using the same datasets used in the previous assignments.

Titanic Dataset

The goal of this classification problem is to select a classification algorithm that can correctly classify whether a passenger aboard the Titanic would survive or not, given pieces of information (or features) for each person.

Below are a few samples from the dataset:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S

Table 1. Samples from the Titanic Dataset

Breast Cancer Wisconsin Dataset

The goal of this classification problem is to select a classification algorithm that can correctly classify whether a sample scan of the nucleus of a cancerous cell is malignant or benign, given a set of characteristics of the nucleus.

Below are a few samples (and only 6 features out of 32) from the dataset:

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
842302	M	17.99	10.38	122.8	1001	0.1184
842517	M	20.57	17.77	132.9	1326	0.08474
84300903	M	19.69	21.25	130	1203	0.1096

Table 2. Samples from the Breast Cancer Wisconsin Dataset

For preprocessing, we do some feature-wise normalization on the columns of the dataset. In addition, we select only prominent features based on the correlation matrix produced from the dataset. The chosen features are *radius_mean, texture_mean, smoothness_mean, compactness_mean, concavity_mean.*

Clustering

The goal of this branch of algorithms in the unsupervised learning domain is to be able to split an unlabelled dataset into a set of clusters. In this assignment, we'll be exploring 2 iterative methods for clusterings: K-means and Expectation-Maximization.

K-means

The goal of this algorithm is to split the dataset into **k** clusters, where it calculates the distance between each data point and the center of its associated cluster; the metric used for that evaluation is the "within cluster sum of errors". K-means is an iterative process where the centroids are adjusted per iteration to minimize the distance; the stopping criteria is having the distance below a specified threshold. The distance function we use for the upcoming experiments is the Euclidean Distance function.

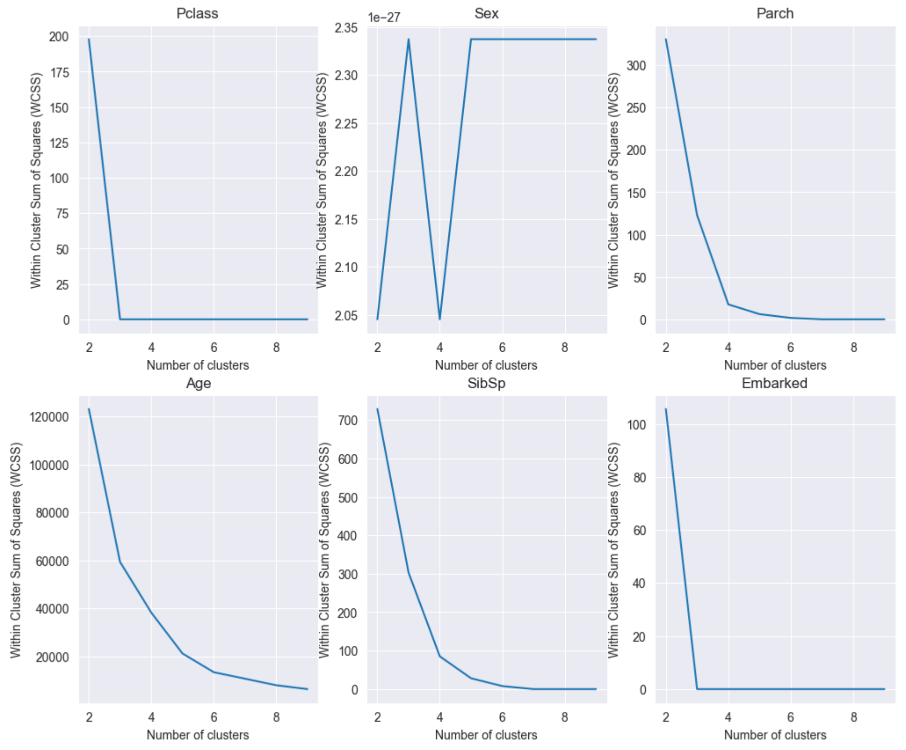
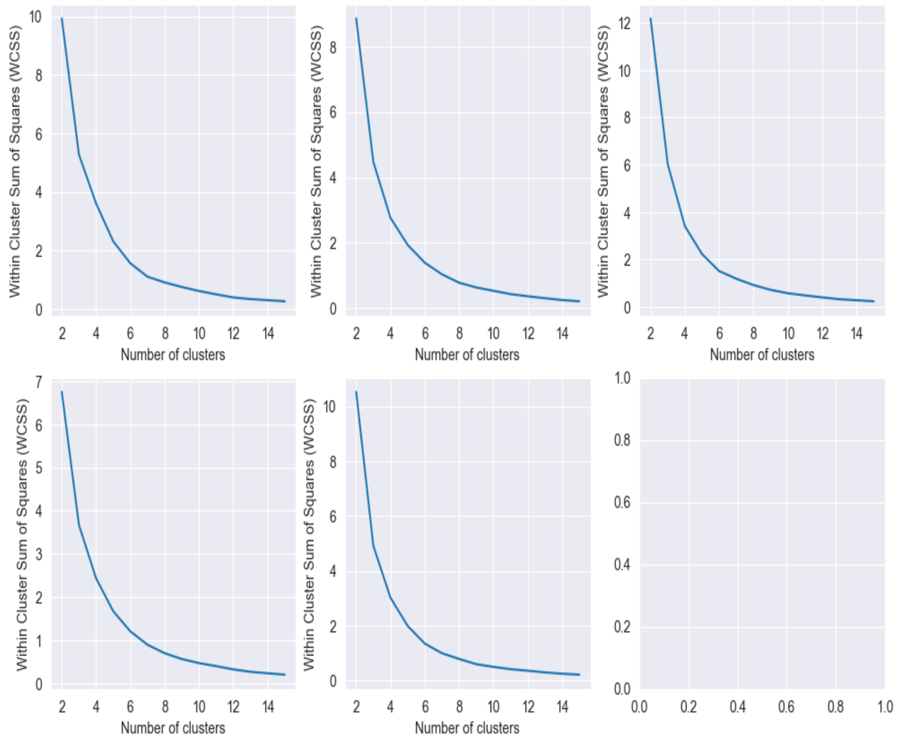
Titanic Dataset	Breast Cancer Dataset
	

Table 1. WCSS given all features of each dataset

While we know that this is a Binary Classification problem, we will undergo the assumption that we do not know the nature of this problem. Thus, we will be in search of the suitable value of **K.** To do so, we use the elbow method on the WCSS to select the number of clusters.Table 1 shows the WCSS corresponding to a range of tested values of K (between 2 and 10).
From this we select the appropriate values of K for the Titanic and Breast cancer datasets to be **4** and **6,** respectively.
Furthermore, we use these values of K to visualize the clusters for each dataset.

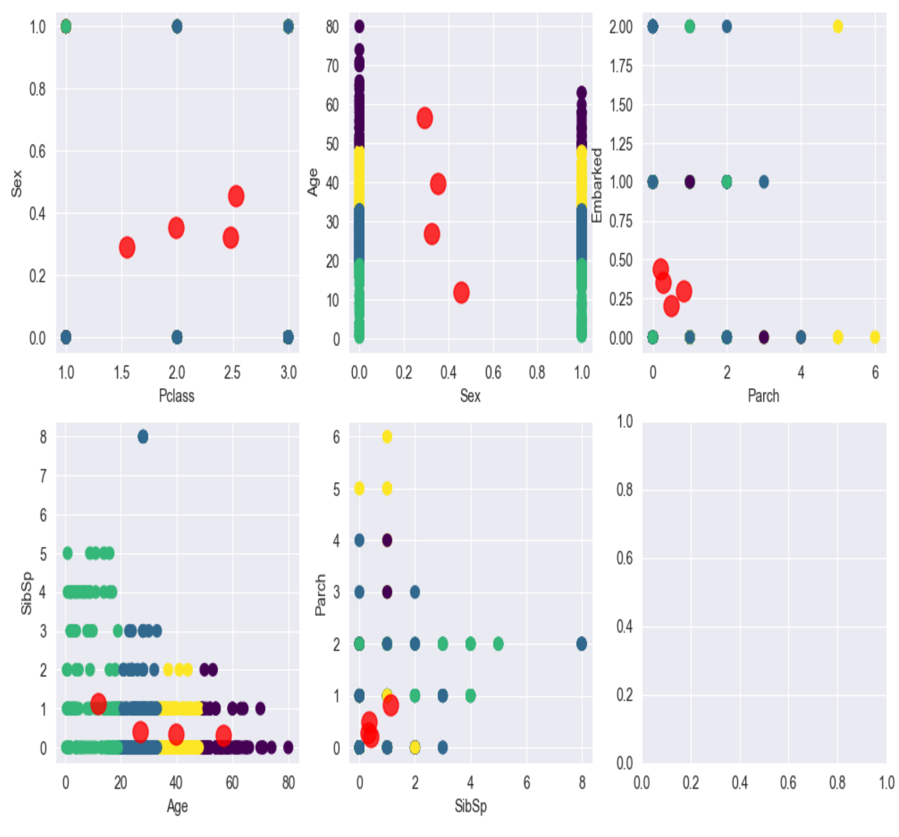
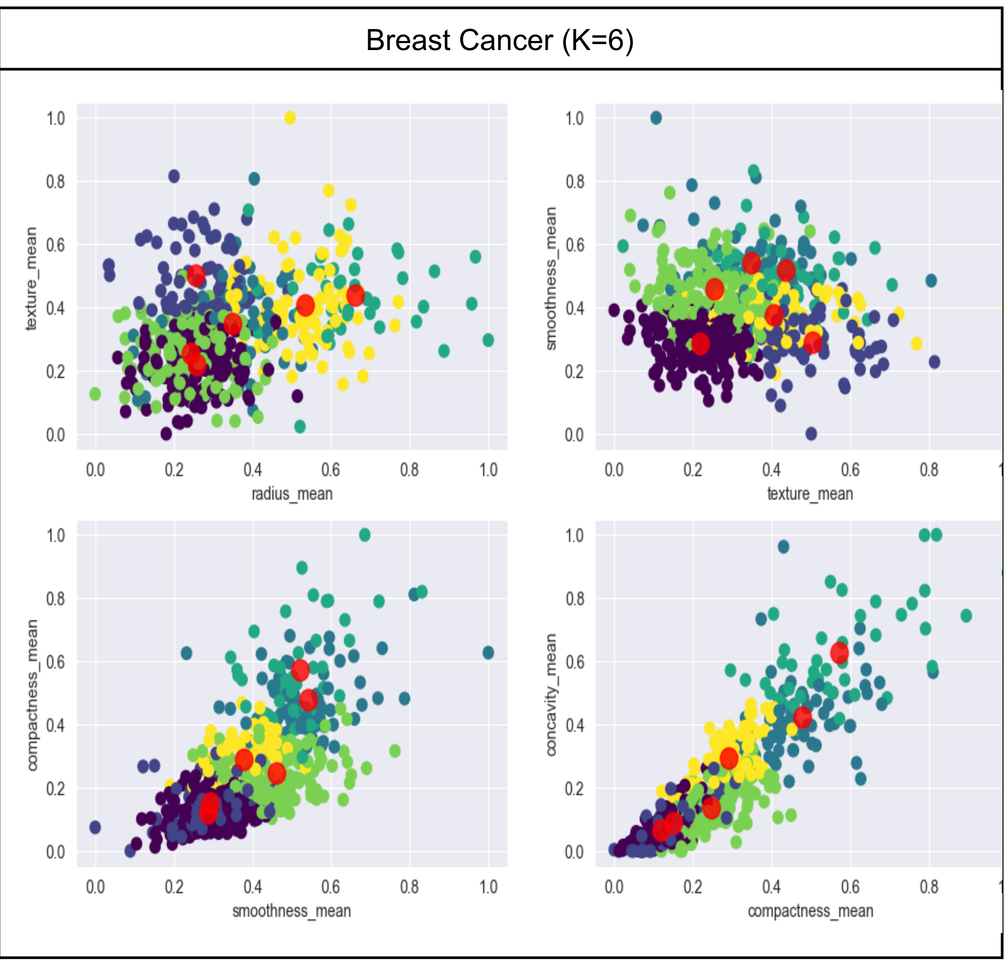
Titanic (K=4)	Breast Cancer (K=6)
	

Table 2. Feature X Feature visualization of clusters for both datasets

Furthermore, we visualize the centroids for each dataset. For the titanic dataset, we notice the separability of the clusters is highly dependent on the type of features. We notice that the Age plots produce the most representable plots given the projected 4 clusters of K-means, while the rest of the features did not have accurate clusters. This is an important thing to keep note of as we explore dimensionality reduction techniques later on in the report.
For the Breast Cancer dataset, we notice that all feature plots have more or less the same representative characteristics regarding the 6 clusters produced by the algorithm.

Expectation-Maximization

This algorithm converges towards finding K distributions such that the likelihood of the data points is maximized. In order to select the suitable number of clusters, we plot the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) curves and use the elbow method to select K. Given the plots in Table 4, for both datasets, we select **K=3**.

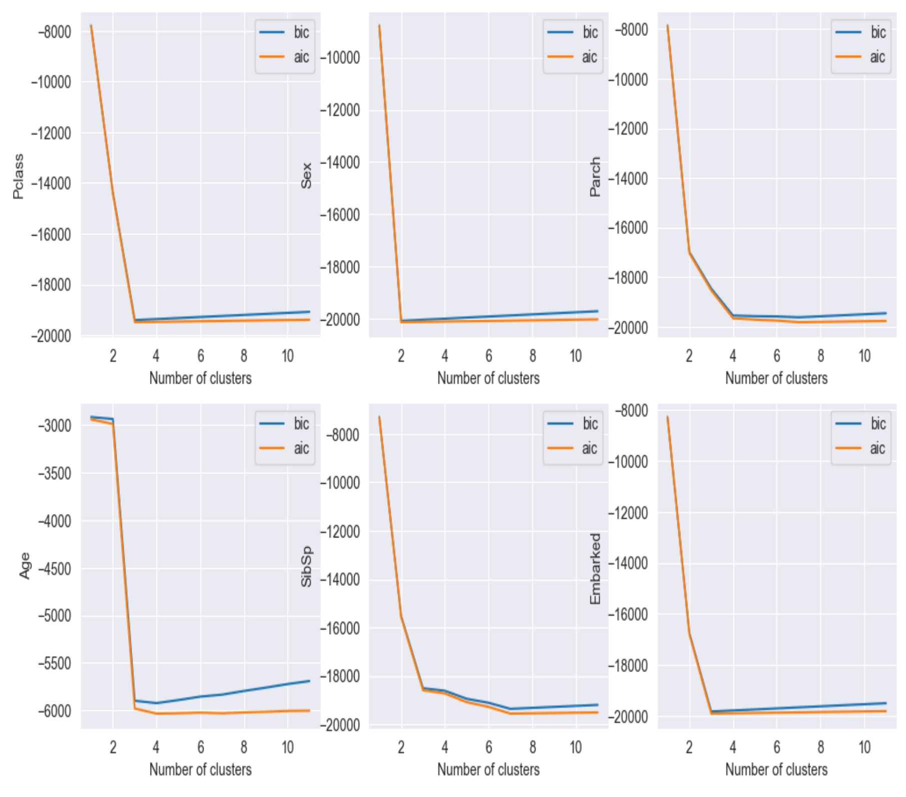
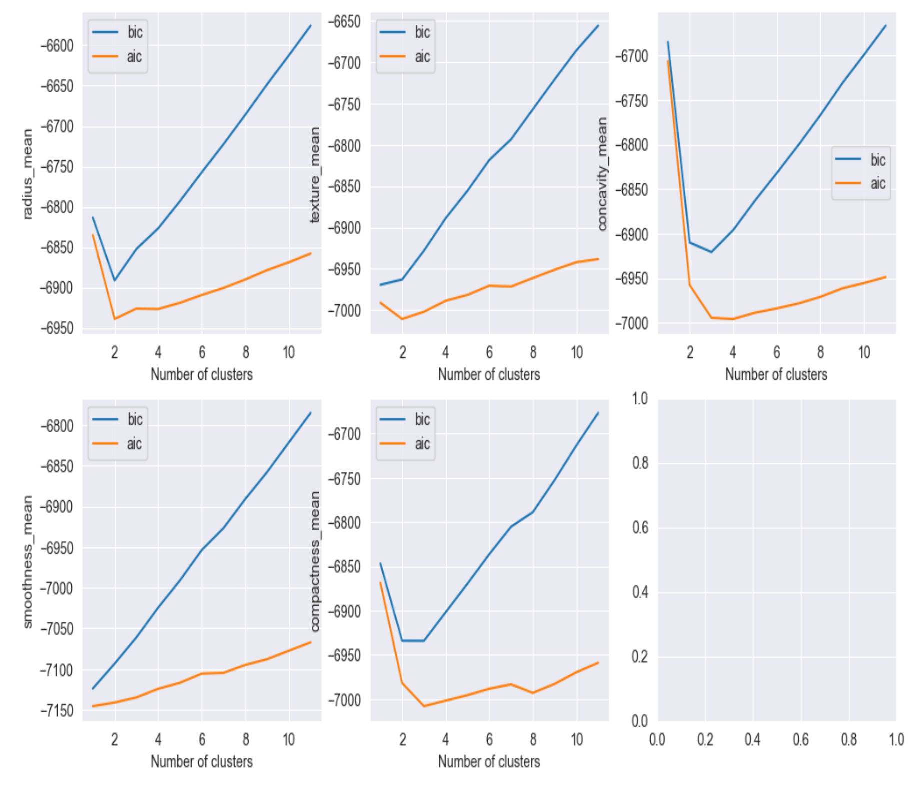
Titanic	Breast Cancer
	

Table 4. BIC and AIC curves for the EM algorithm on both datasets

Dimensionality Reduction (DR)

The goal of DR is to be able to project our original dataset features onto another space where we could maintain enough information on the dataset with possibly less amount of features. In this section we will explore four DR algorithms: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections (RP), Linear Discriminant Analysis (LDA). For each algorithm, we will assess how the DR is performed, the metric used for feature selection and effectiveness of each component on the classification accuracy given a clustering method.

PCA

This is a statistical technique that is used to linearly transform the data, preserving the maximum amount of information, where the projection aims to maximize the retained variance. This is an eigen problem where the principal components are the eigenvectors and the retained variance is resembled by the eigen values.

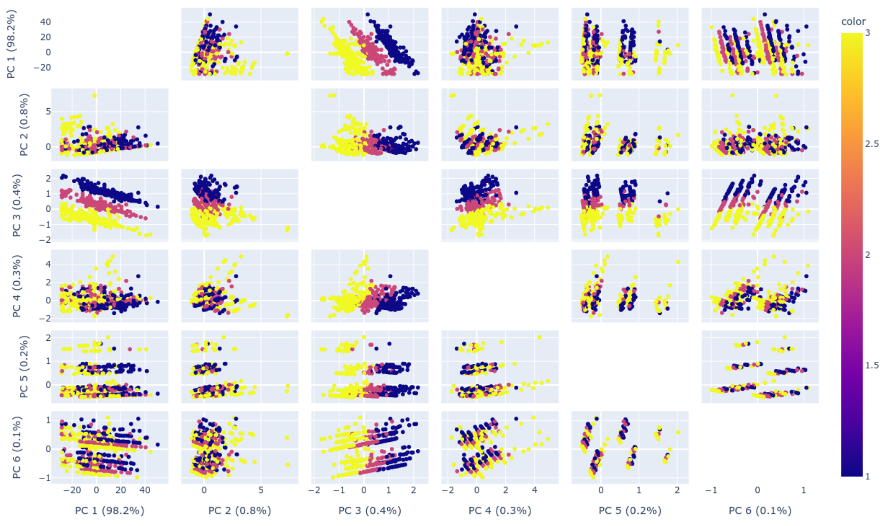
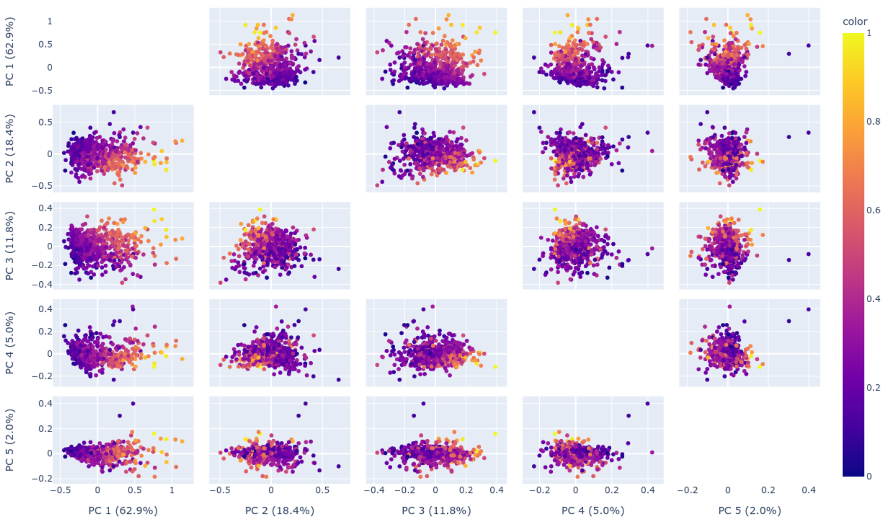
Titanic	Breast Cancer
	

Table 5. PCA projections for both datasets

The figures in Table 5 show the projections of the principal components, we can see the directions maximizing the variance. Selecting the principal components here is straightforward: we select the PCs with the largest retained variance. For the Titanic dataset the top 2 PCs retain 62% + 18% = 80% of the total variance. While, for the Breast Cancer dataset, the top 1 PC retains 98% of the variance. This means that out of 5 features from the dataset, one transformed feature would be sufficient in describing the dataset. We will later verify this with cross validation accuracy metrics produced.

ICA

This algorithm aims to separate multiple features into additive sub components which are statistically independent from each other. We calculate the kurtosis across each feature and select the highest as our chosen features.

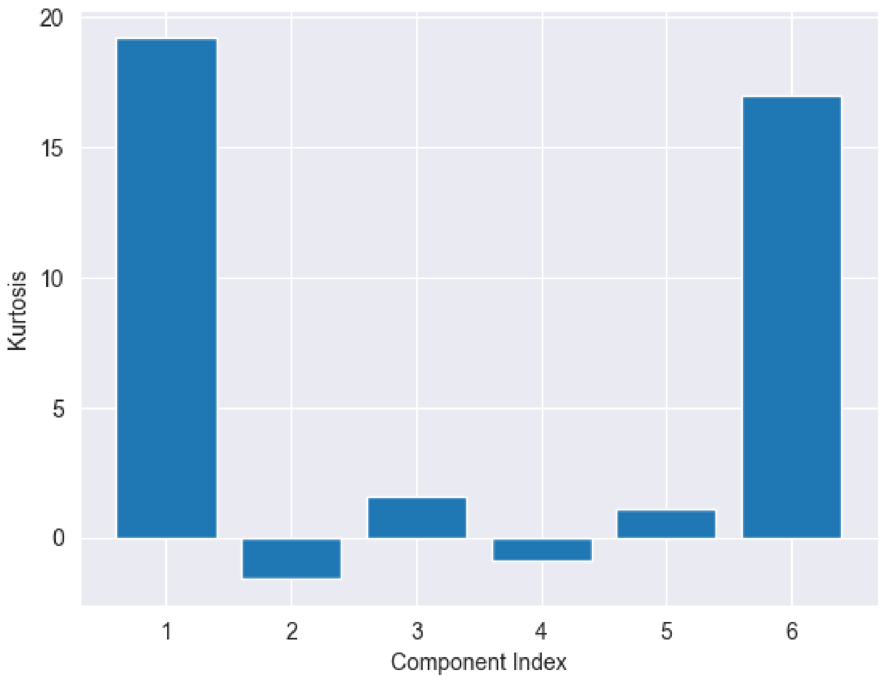
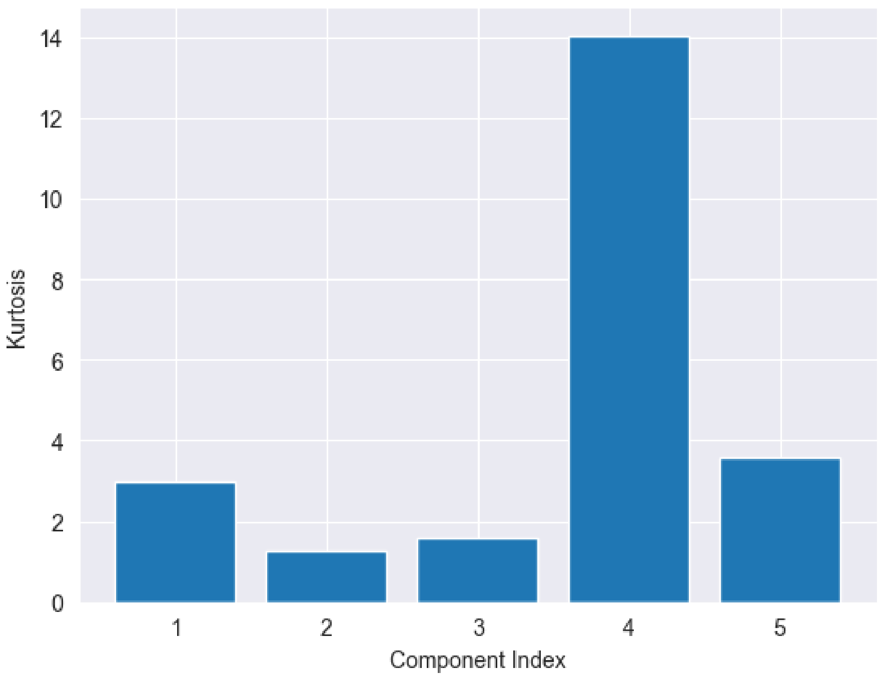
Titanic	Breast Cancer
	

Table 6. Kurtosis of IDA on both datasets

Randomized Projections

We use the Gaussian Random Projection algorithm which ..”reduces dimensionality by projecting the original input space on a randomly generated matrix where components are drawn from the following a gaussian distribution” (sklearn). We use the reconstruction error to select the suitable number of principal components.

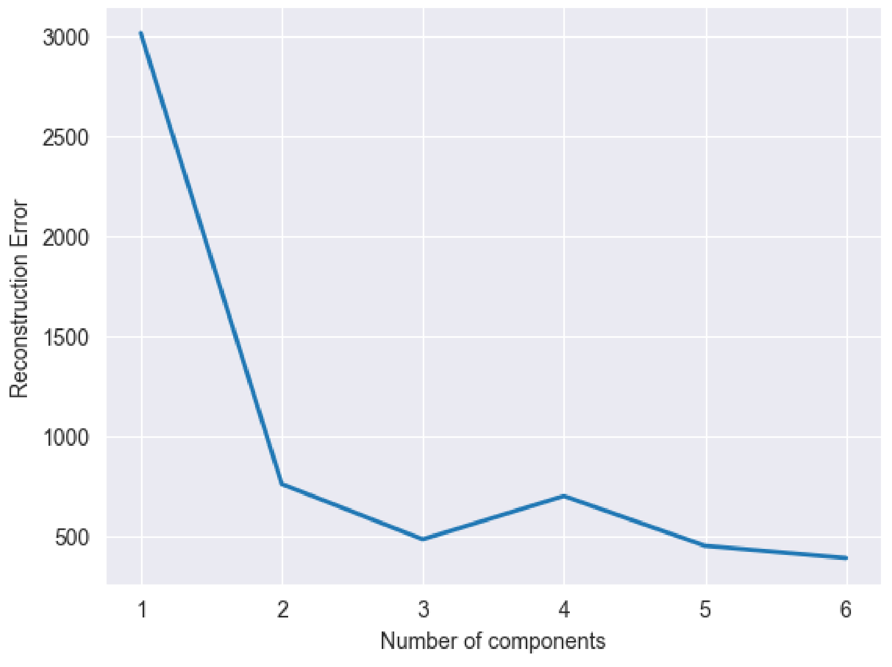
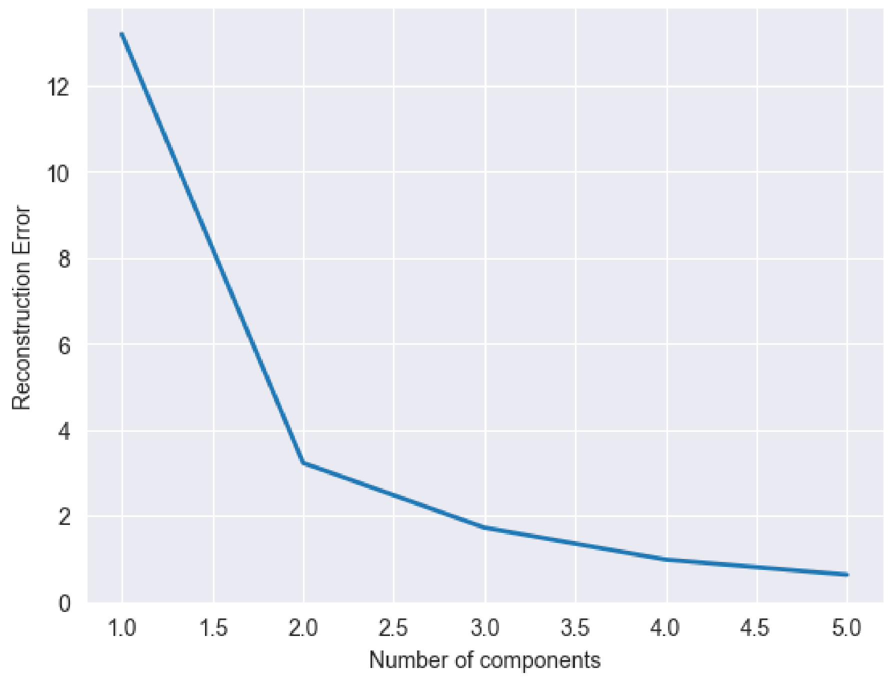
Titanic	Breast Cancer
	

Table 7. Reconstruction error of RP on both datasets

Based on the reconstruction error, we are confident that selecting 3 components for both datasets would produce sufficient

LDA

This algorithm is not solely a dimensionality reduction algorithm, but it finds directions of maximum class separability. It also uses the labels to be able to fit its function. As an opposite of PCA which produces the line which best represents the data, LDA produces the line which can best separate different classes.

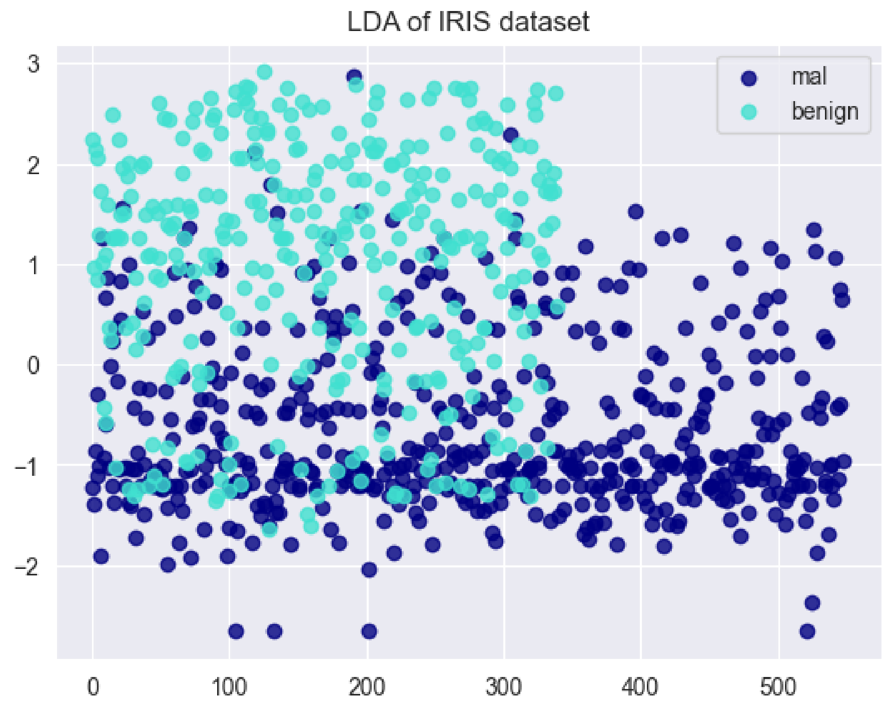
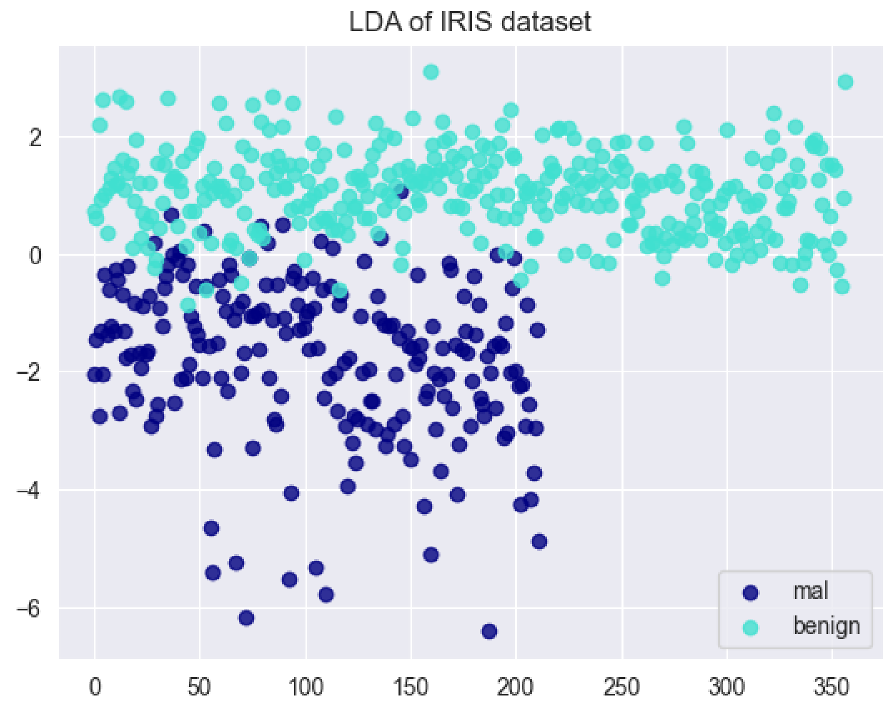
Titanic	Breast Cancer
	

Table 8. Visualization of LDA transformed data for both datasets

The visuals in Table 8 show that, while both are binary classification problems, LDA performs significantly better on the Breast Cancer dataset than on the Titanic Dataset. It shows the linear separability characteristics of both problems.

Results

Clustering results with Feature Selection

Given the first three algorithms: PCA, ICA and RP, we run both our clustering algorithms on every single component independently.

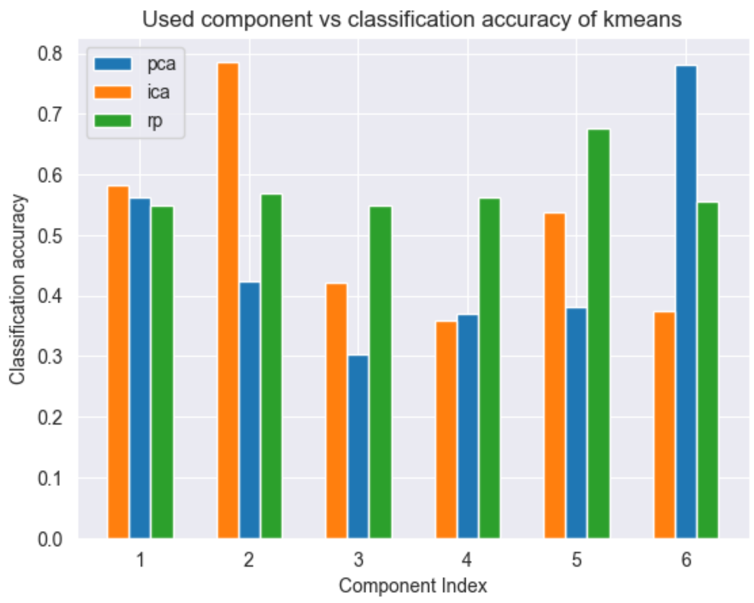
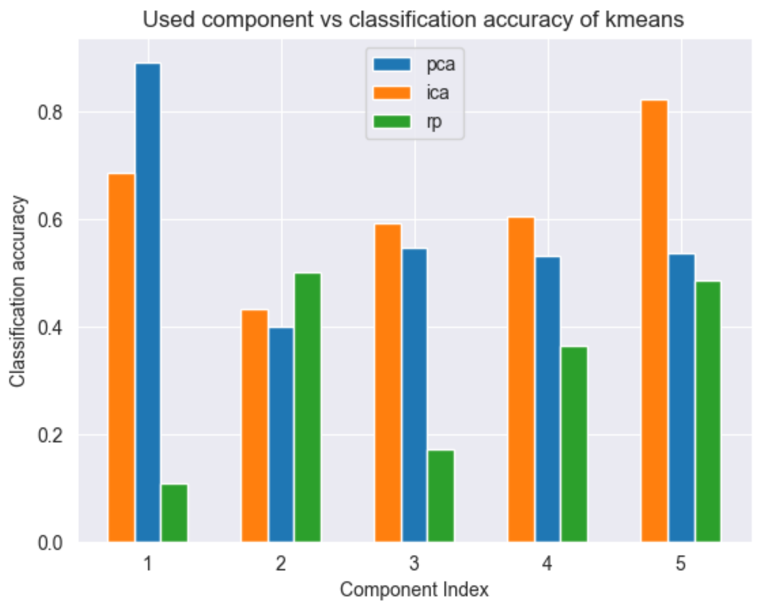
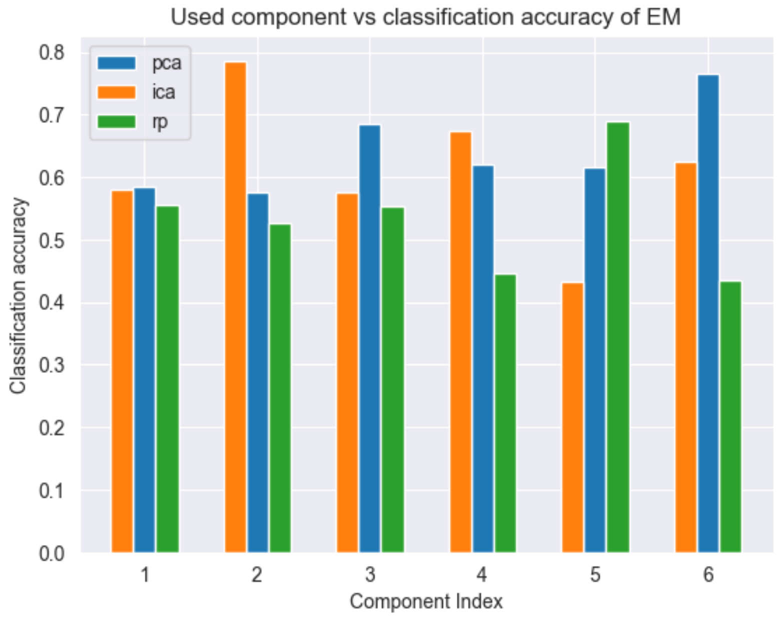
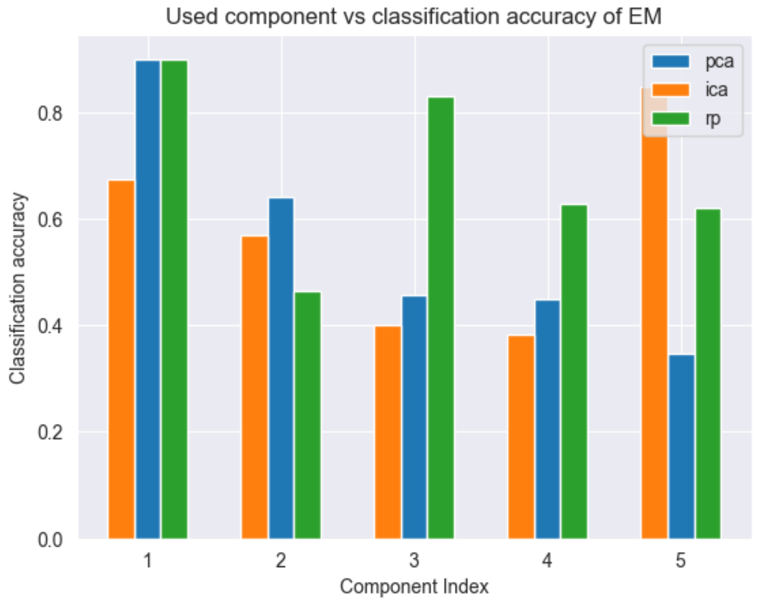
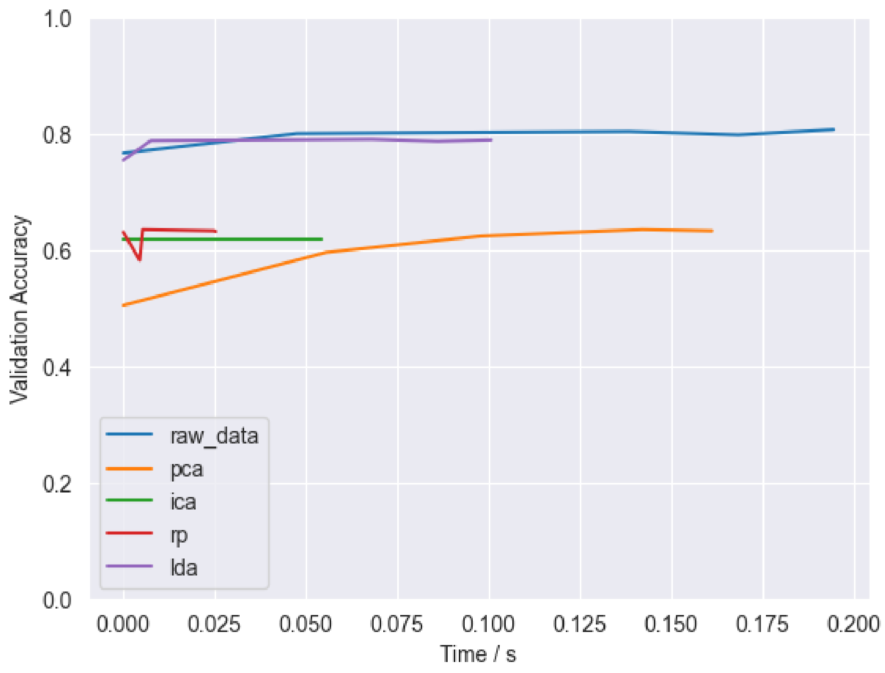
	Titanic	Breast Cancer
K-MEANS		
Expectation-Maximization		

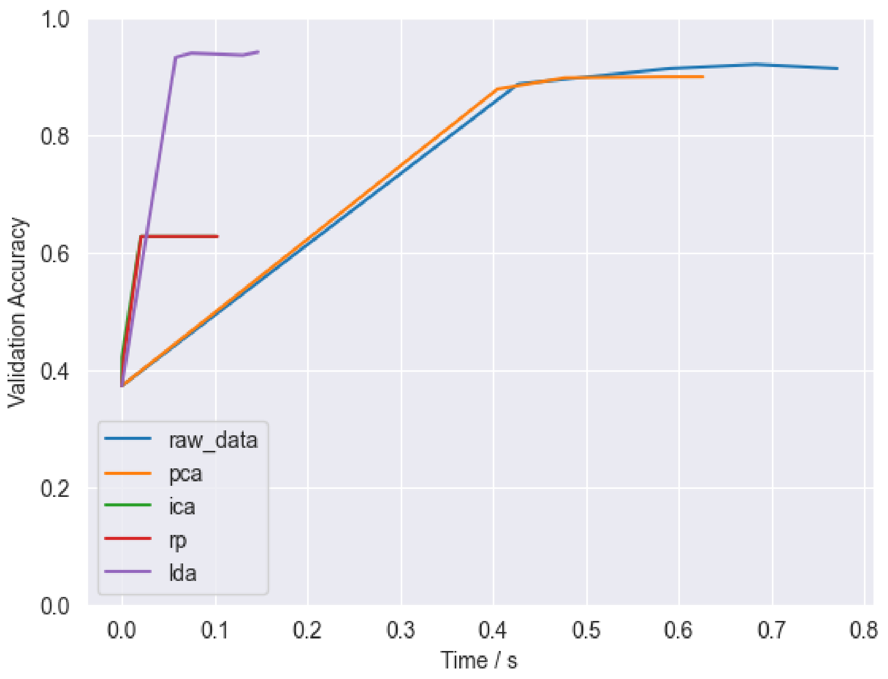
Table 9. Bar charts of the classification accuracy using each extracted component independently

For the experiments in Table 9, we set both of our clustering algorithms to look for K=2 clusters, to be able to evaluate accuracy with the ground truth labels. We can draw some interesting insights from these experiments:

- For PCA, we know that the first Principal Component retained 98% of the variance for the Breast Cancer problem, thus it is expected to see both clustering algorithms reach a peak of ~90% accuracy. We do not see the same behavior on the Titanic dataset as the variance retained is more spread across the first 3 components, making the use of just one single component not sufficient.
- We do some fluctuation in classification accuracy along the results of each principal component; we can presume this fluctuation is mainly influenced by the random initialization.

Neural Network fitting results with Feature Selection

<table>
<tr><th>Titanic</th><th>BreastCancer</th></tr>
<tr><td>


</td><td>


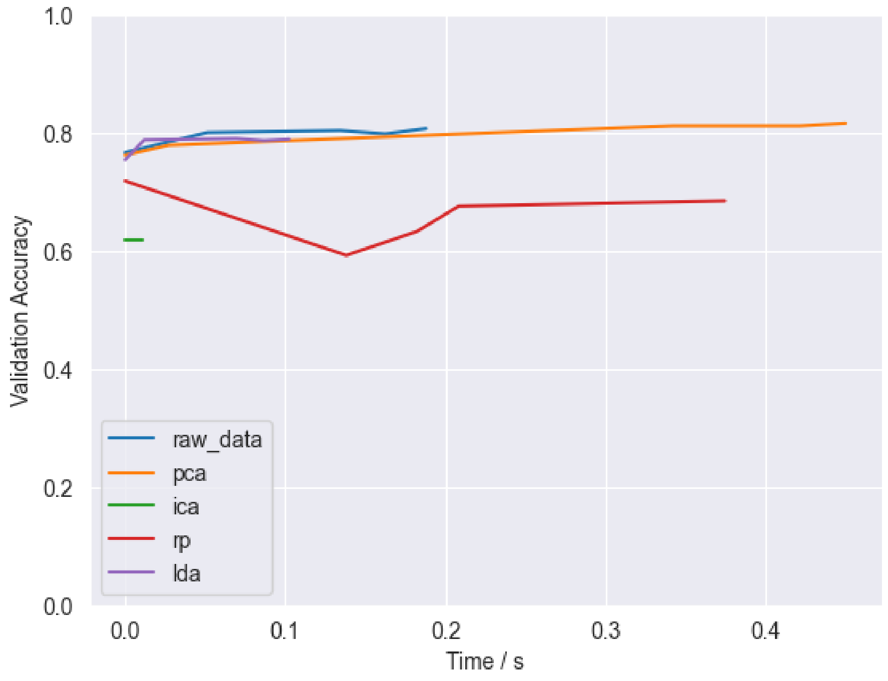
</td></tr>
</table>

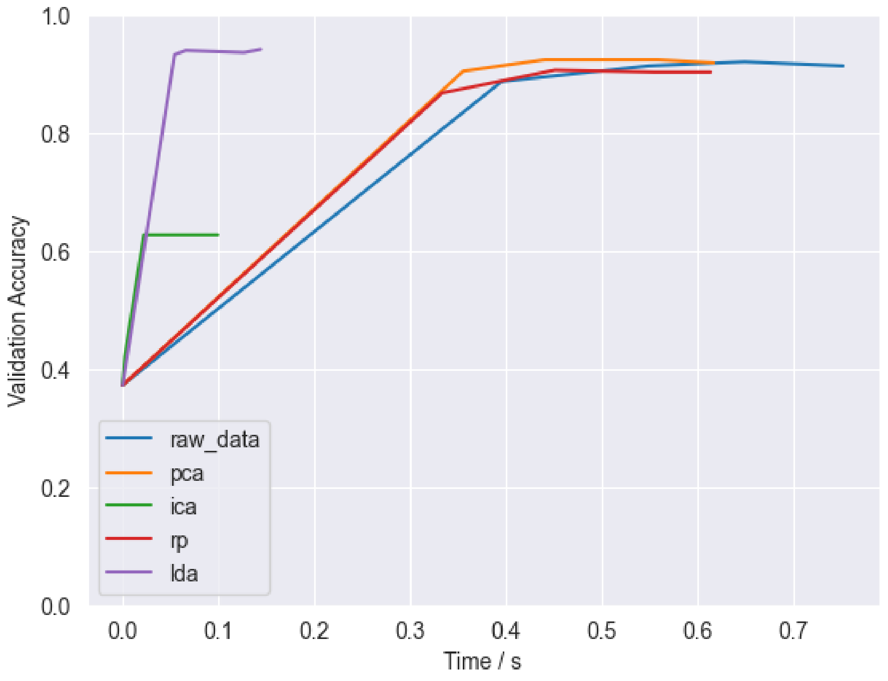
Table 10. Neural network cross validation accuracy based on transformed data with selected components

For the experiments shown in Table 10, we run several training cycles for a neural network on our transformed data using our DR algorithms. We select the best set of features based on the analysis presented earlier. We are able to deduce some interesting conclusion given these plots:

- Raw data resembles the ideal scenario for the classification problem. It is expected that using the raw data would allow the model to achieve the highest accuracy.
- PCA serves as the safest way of performing DR as it aims to represent the data given set principal components. For Breast Cancer, the highest PC retained 98% of the variance, so it was sufficient to reach the same peak accuracy as the raw data experiment. However, for the Titanic Dataset, even using 3 PCs with a total of 90% retained variance could not provide the neural network with enough information to perform on this problem.
- ICA and RP selected features were not sufficient for the neural network to learn. This means that the threshold used for our Kurtosis and Reconstruction Error metrics may have needed to become more lenient.
- LDA, for the most part, worked on solving the binary classification problem as it had access to the ground truth labels. Therefore, as we apply the transformation on the data, the neural network should easily converge at an optimal accuracy.

In Table 11, we will run the same experiments, but without we will select all features from the transformed data. This is to see which DR algorithm is irreversibly lossy. We notice that PCA using all components reached the peak for the Titanic Dataset; however, the ICA and RP transformations seem to have produced insufficient information in the data overall.

<table>
<tr><th>Titanic</th><th>Breast Cancer</th></tr>
<tr><td>


</td><td>


</td></tr>
</table>

Table 11. NN results on transformed data using all features

Neural Network results given clustering output

In this section, we run both our clustering algorithms and use the output of attributions of the algorithm as input to our neural network. We set the value of K to be the same as the one produced in the analysis in the Clustering section.

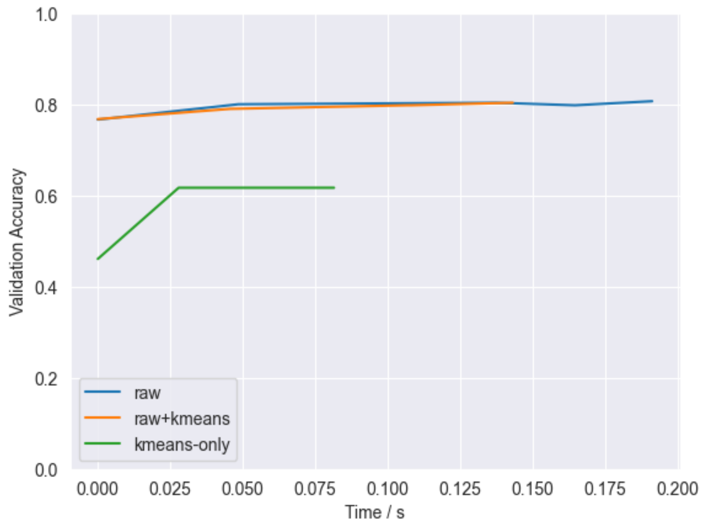
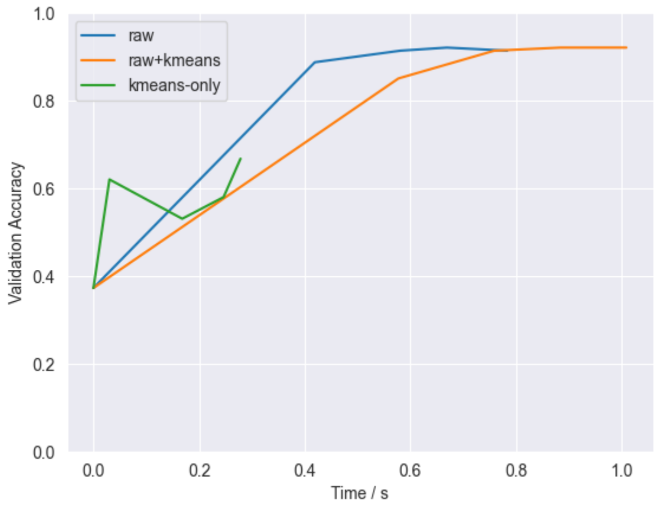
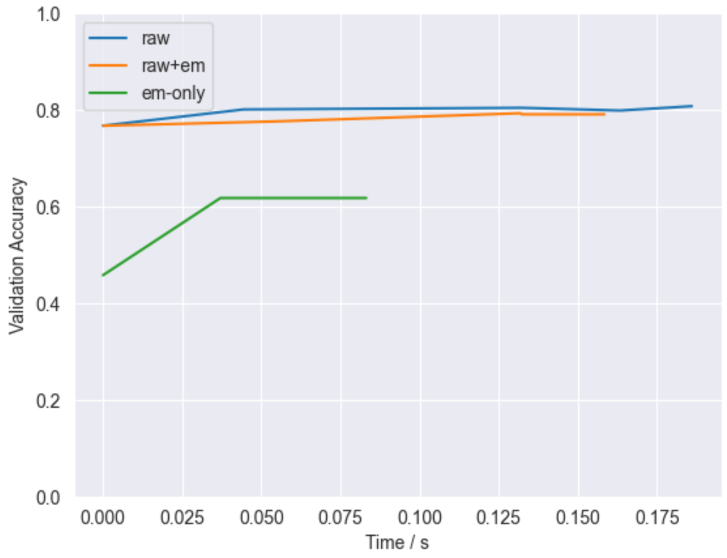
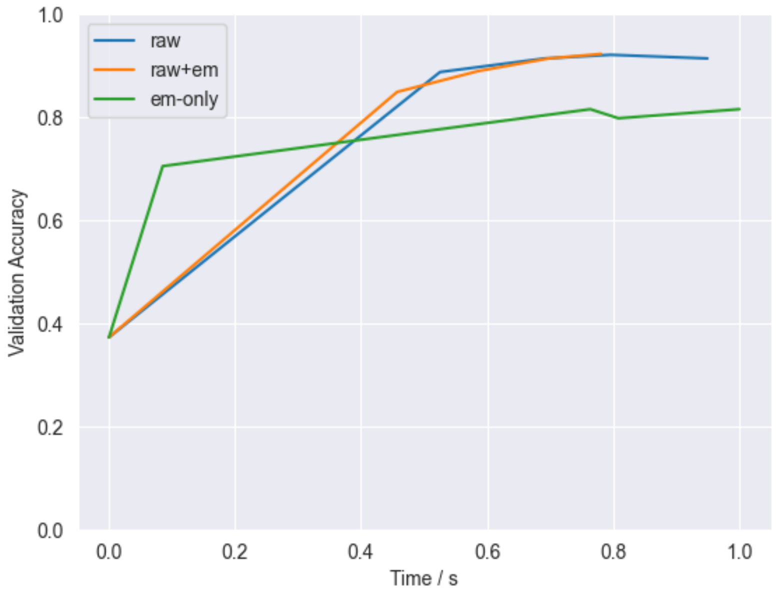
<table>
<tr><th>Algorithm</th><th>Titanic</th><th>Breast Cancer</th></tr>
<tr><td>K-means</td><td>
</td><td>
</td></tr>
<tr><td>EM</td><td>
</td><td>
</td></tr>
</table>

Table 12. NN results using clustering output as input features. Raw+EM and Raw+Kmeans is where we concatenate the clustering labels as an additional feature to the raw features of each dataset.

As seen from Table 12, using the clustering output as the sole feature to the neural network will only produce good convergence if:

1. The value of K matches the characteristics of the classification problem.
2. The clustering algorithm performs well given the features of the dataset.

The issue with the Titanic dataset is that its features are not easily correlated with the label, that is why it is expected to see the data produced from both clustering algorithms to be insufficient for the neural network.

On the other hand, for the Breast Cancer dataset, the features are highly correlated with the label, thus the main deciding factor on whether the produced output from either clustering algorithm would be considered as a useful feature vector for our neural network is the value of K. In the plots above, we had K=6 and K=4 for K-means and EM respectively.

If we were to use some simple prior information on the dataset, knowing that it is a binary classification problem, and set the value of K=2, we would have both clustering algorithms produce spot-on feature vectors that can be efficiently used to train our neural network, as shown in Table 13.

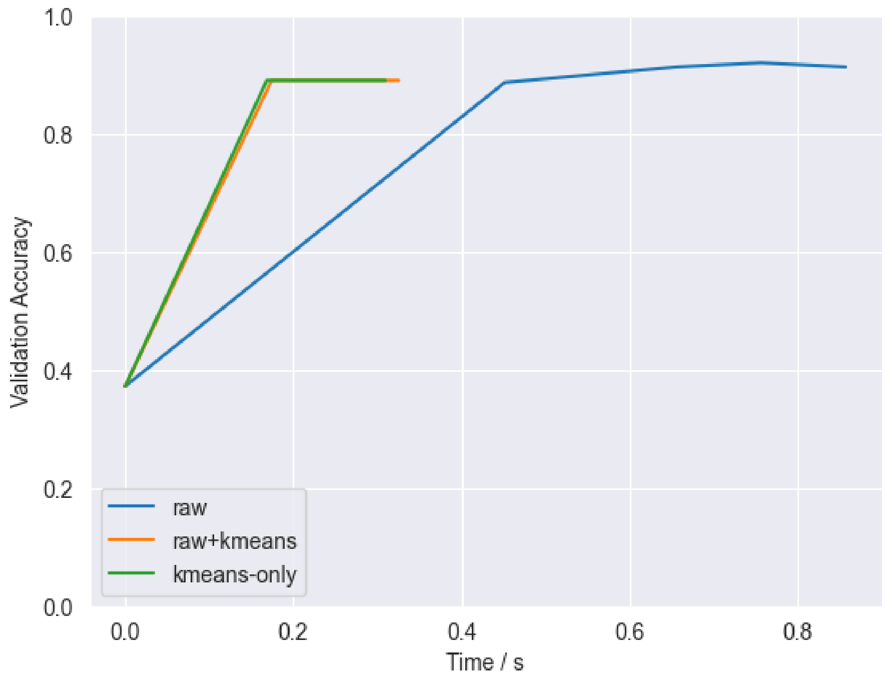
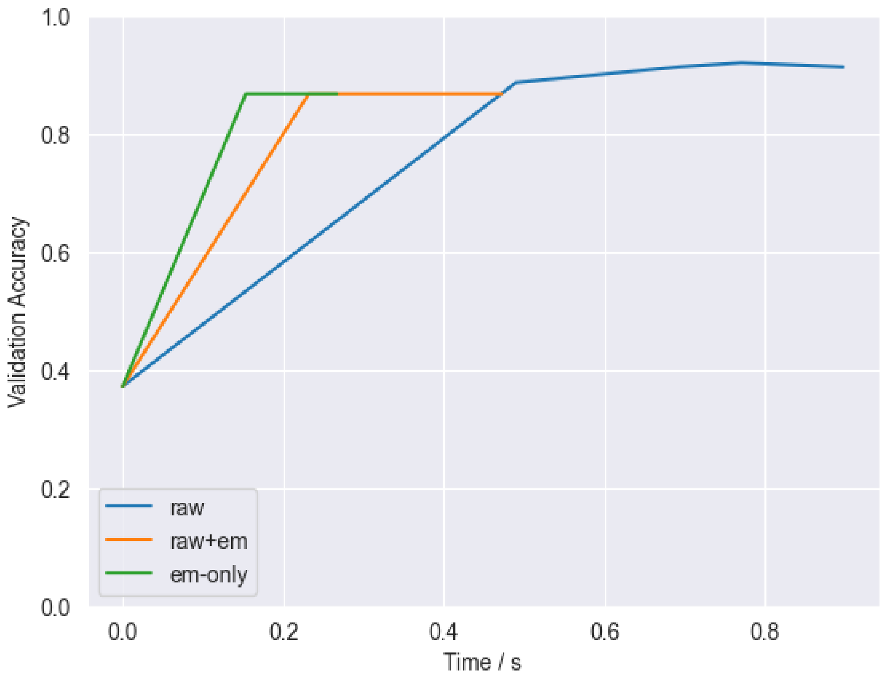
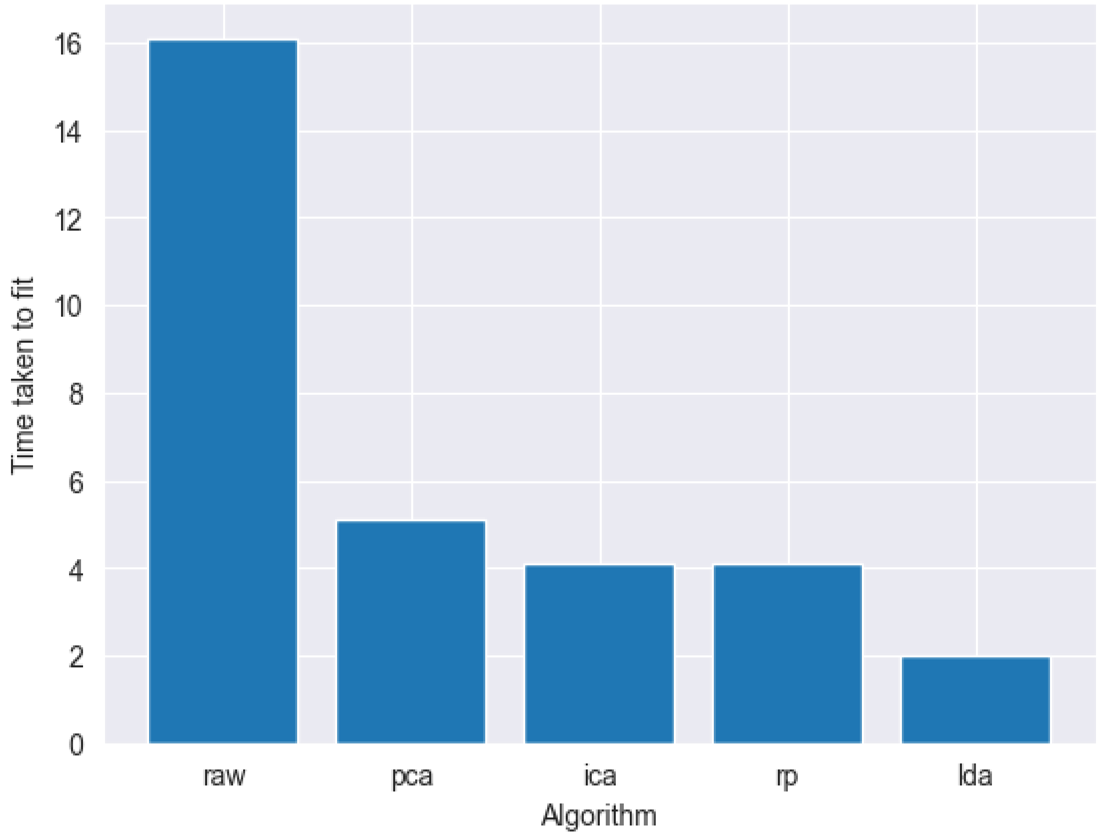
K-means	EM
	

Table 14. Results of our Neural Network on the Breast Cancer data with K=2

Wall Clock Time

One important goal of using dimensionality reduction is to reduce the complexity of our learning models. Less features should mean less complexity in the learning algorithm, as per the Curse of Dimensionality narrative. By selecting a subset of the features of the dataset, we are able to reduce the size of the neural network, reduce the time to fit and possibly shrink the memory footprint of the model. Figure 1 shows the wall clock time to fit the same neural network but on different datasets.



Conclusion

In this assignment, we had the opportunity to explore two large branches under the Unsupervised Learning field. We explored the behavior of 2 major clustering algorithms. We were able to see the difference between hard and soft clustering, where K-means assigns data points to neighboring centroids while EM takes a more probabilistic approach towards the same goal. We also explored the 4 dimensionality reduction techniques and metrics for selecting the extracted features or components. We solidify our analysis with results using a set of experimentation cycles using a neural network and the ground truth labels of our classification problems.

References:

https://scikit-learn.org/stable/modules/random_projection.html