Supervised Learning

CS7641 Assignment 3

Unsupervised Learning and Dimensionality Reduction

Hanna Sohn
Computer Science
Georgia Institute of Technology
Atlanta GA 30332
hsohn37@gatech.edu

***Abstract*—This assignment aims to explore unsupervised learning algorithms in clustering and linear or non-linear dimensionality reduction for two datasets and conduct several experiments to dive into the effects of unsupervised learning techniques including the combinations of their techniques and neural network algorithms. The experiments consist of five parts. The first one is for clustering algorithms. The second one is for dimensionality reduction. The third one is the combination of the first experiment and the second experiment: applying the clustering algorithms on the projected datasets via dimensionality reduction. The fourth and fifth one is for neural network using the datasets using dimensionality reduction and the combination of dimensionality reduction and clustering. The clustering algorithms used in the experiments are Expectation maximization and one of clustering algorithms. The linear dimensionality reduction algorithms are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections. One non-linear dimensionality reduction algorithm is added using Manifold Learning Algorithms. Through the experiments, we can see how the clustering and dimensionality reduction techniques work, what they behave under different situations, and how they improve performance in unsupervised learning environments.**

Keywords—unsupervised learning, clustering, dimensionality reduction, neural network, Expectation Maximization, Clustering, PCA, ICA, Randomized Projections, Manifold Learning

I. Introduction

Unsupervised learning is a type of machine learning algorithm "that looks for patterns in a dataset without pre-existing labels" [1] Unsupervised learning models are used to find similarities among unlabeled data using clustering and find relationships between data in groups. When the number of features in a dataset is high, dimensionality reduction technique can help reduce the number of inputs, which leads more convenient when handling data while keeping the data integrity. In this assignment, we will show that clustering and dimensionality reduction can help produce the similar results as the original datasets even though the method eliminate some features, and they even enhance the performance with combinations of some techniques. We also see when the techniques are useful under which situations and how they work. There are three kinds of unsupervised learning algorithms used here: clustering algorithms, linear dimensionality reductions, and a non-linear dimensionality reduction.

There are two datasets used in the experiments which are the same as the ones in the assignment 1. They are all labeled data and we exclude the target values when needed in the experiments. More detailed explanation will be described in the following section.

The experiments are divided into five sections. The first one and the second one are to apply the two clustering algorithms and four dimensionality reduction algorithms for each dataset. The third one is to apply the two clustering algorithms on the projected datasets using dimensionality reduction. The fourth one is to use a neural network technique on the dimension-reduced dataset and look into the impact to compare the original results with. The final one is the combination of the neural network, clustering, and dimensionality reduction and analyze the performance.

The primary library packages are Scikit-Learn for the machine learning functions, PyTorch for neural network, and Yellow Brick for visualizing clustering.

II. Datasets

One of the datasets used in the assignment is the dataset to predict students' dropout and academic success. The number of feathers is 36, including GPA, majors, and socioeconomic conditions. [2] There are three target values: dropout, enrolled, and success. The 'enrolled' value is removed since the target value does not indicate the current situation to determine

students' academic success. The number of instances is 3630, and the number of dropouts is 1421 while the number of successes is 2209.

The other dataset is the one for AIDS Clinical Trials Group Study 175. It has 23 features for healthcare statistics and categorical information of AIDS patients. The number of instances is 2139. The target value is binary where 1 means failure and 0 means censoring. The number of failures is 521 while the number of successes is 1618.

The two datasets are labeled data and the target values are sometimes not used in the experiments under unsupervised learning environments. They are from UC Irvine Machine Learning Repository.

III. Experiments

There are five steps to conduct the experiments in the assignment. In each step, we measure the performance of the algorithms in the aspects of time complexity and the prediction accuracy. For the accuracy, Decision Tree Classifier is used to compare the results of the original data and the projected datasets via the experiments. As we brief the experiments in the previous section, we start to detail the experiments as followings.

A. Step 1

In Step 1, we apply two clustering algorithms to the two datasets. The two clustering algorithms are Expectation Maximization (EM) and Clustering Algorithm of the choice, where the GaussianMixture(GMMs) is chosen as EM and KMeans is chosen as a clustering algorithm. Scikit-Learn package is used for implementing. The choice of the number of clustering is tricky. To select the right numbers of clustering, we use the Silhouette scores for all algorithms. Moreover, to gather more information on the choice, the Silhouette Visualizer and Intercluster Distance from yellow brick are used in KMeans while the Distance between train and test GMMs for model generalization evaluation and clustering performance evaluation.

B. Step 2

The four dimensionality reduction algorithms are used on the two datasets. The first three algorithms are linear while the other is non-linear. To implement Step 2, the classes we use for linear ones are PCA as PCA, FastICA as ICA, and SparseRandomProjection as Randomized Projections. The class for non-linear dimensionality reduction is TruncatedSVD. To select the right number of components, the explained variance ratio and cumulative explained variance ratio are used in PCA and Non-linear dimensionality reduction. Since the explained variance ratio indicates how the principal components cover the data variance and the cumulative explained variance explain the total variance using the principal component, we take a look at the curve of the variance values to choose the least value of the clustering to have the high performance. For ICA and Randomized Projections, the cross validation and the test accuracy values are used to determine the right value of clustering. To visualize the results, we use the scatter plots to compare the relations of features.

C. Step 3

The combinations of Step 1 and Step 2 are utilized for two datasets. The first thing to do for Step 3 is dimensionality reduction to reduce the number of features. Next, the clustering techniques are used on the reduced datasets. The proper numbers of components of the dimensionality reduction are the ones selected as the best in Step2. Like Step 1, we test the range of clustering numbers and choose the proper ones using Silhouette scores, the Silhouette Visualizer and Intercluster Distance, and the Distance between train and test GMMs. The total number of experiments in step 3 is 16 and four total demonstrations between the two datasets are detailed.

D. Step 4

Neural Network in Assignment 1 is rerun in this and the following steps. The neural network used here is NN class from Pytorch. Torch.optim is also used as an optimizer object. There are two hidden layers and the rectified linear unit (ReLU) function as an activation function. The first dimension of the NN should be changed according to the results of dimensionality reduction. Only one dataset is used: students' dropout and academic success. Instead of the original dataset, the projected dataset via dimensionality reduction in Step 2 is used to compare the performance. The performance metrics are measured using the training time and the accuracy rate of cross validation score and test accuracy. To do this, we need to know the original performance data in Assignment 1 to compare the performance. Among the results, two total demonstrations are detailed.

E. Step 5

Using the same neural network in Step 4, we use the same dataset, the students' dropout and academic success, used in Step 3, which are the results of clustering and dimensionality reduction. To do so, we use the proper numbers of clustering and components in Step 3. Since we apply the clustering algorithms, we can predict the y values via the clustering processes. Instead of the original target values, we feed the predicted target values into the neural network and compare the results using the cross validation scores and test accuracy, plus the time consumption in training.

IV. Results

A. Step 1

For Students' dropout dataset, the silhouette score in KMeans is dropped when it is 3. Moreover, the kmeans intercluster distance map shows that the circles are duplicated when it is 3. Therefore, the ideal number of cluster is 2. In EM, the silhouette score in EM dramatically decrease when it starts to be 3, and the distance between train and test gmm's also significantly increases when it gets to be 3. Therefore, the proper number of cluster is 2 in EM.

For AIDS dataset, the silhouette score in KMeans when the number of clusters is 2 or 3 and the results of the kmeans intercluster distance map shows the similar results when the value is 2 or 3. Therefore 2 or 3 are similar in KMeans. However, in EM, the results are conflicts because the silhouette score is the highest when the value is 3 while the distance between train and test GMM's is much higher when it becomes 3. But the difference of silhouette between 2 and 3 is smaller

than the difference between train and test GMM's, 3 is chosen there.

TABLE I. NUMBER OF CLUSTERS IN CLUSTERING ALGORITHMS

# of clusters	***Dropout***	***AIDS***
KMeans	2	2 or 3
EM	2	2

B. Step 2

First, for Students' dropout dataset, PCA's explained variance ratio is significantly dropped when it is 3 or more while the cumulative explained variance ratio is much higher when it is 2 already. Therefore, 2 is chosen for PCA. However, in ICA, the best number of components is 14, and the value is 9 in Projected Projections. The reason that the values are higher than the other algorithms are due to the selection method, which is using the highest cross validation and test scores in DecisionTreeClassifier. The accuracy scores tend to be higher when the number of components is high, which predicts the target more correctly. For AIDS dataset, the trend of number of clusters is similar as the dropout datasets. The difference is that the curve of explained variances of PCA and Manifold algorithm in AIDS is smoother than Dropout because AIDS seems to need more clusters to cover the variance. Most of the number of clusters in AIDS are generally higher than Dropout.

TABLE II. NUMBER OF COMPONENTS IN DIMENSIONALITY REDUCTION

# of components	***Dropout***	***AIDS***
PCA	2	3
ICA	14	13
Projected Projections	9	10
Manifold Algorithm	2	4

C. Step 3

For two datasets, the number of clusters using projected data with dimensionality reduction are similar. This is interesting because we have different values of components in dimensionality reduction (2 of PCA and 14 of ICA in dropout dataset), but it does not seem to influence the number of clusters. In AIDS dataset, it seems to have a little more value of clusters in EM.

TABLE III. NUMBER OF CLUSTERS IN COMBINATION(CLUSTERING AND DIMENSIONALITY REDUCTION)

# of clusters	***Dropout***	***AIDS***
KMeans + PCA	2	2
KMeans + ICA	2	2 or 3
KMeans + Randomized Projections	3	2 or 3
KMeans + Manifold Algorithm	2	3
EM + PCA	2	3
EM + ICA	2	3
EM + Randomized Projections	2	2 or 3
EM + Manifold Algorithm	3	3

D. Step 4

The neural network using the projected dataset using dimensionality reduction shows the similar training time though the numbers of components are from 2 to 14. This might be due to the total amount of data is relatively small and the number of instances is more important than the number of features in neural network. But there is an important difference in accuracy in that ICA and randomized projections have much higher performance in cross validation score and test accuracy. We can compare the number of components in dimensionality reduction. When we have more components in dimensionality reduction, we can keep more data integrity which reflects the difference of accuracy.

TABLE IV. NUMBER OF COMPONENTS IN DIMENSIONALITY REDUCTION IN DROPOUT DATASET

Performance	***# of components***	***Train Time***	***Cross Validation Score***	***Test Accuracy***
NN + PCA	2	63.27	0.62	0.60
NN + ICA	14	55.99	0.86	0.86
NN + Randomized Projections	9	58.85	0.87	0.86
NN + Manifold Algorithm	2	57.90	0.57	0.60

E. Step 5

In Step 5, the datasets are transformed using the combination of dimensionality reduction and clustering. For dimensionality reduction, randomized projection is selected. The sigficant difference in step 5 is that the experiment shows much higher performance comparing with the Step 4. One of the reasons might be that the target values used in the neural network is not the original data, but the predicted values of clustering techniques which might influence the accuracy performance.

TABLE V. NUMBER OF COMPONENTS IN DIMENSIONALITY REDUCTION IN DROPOUT DATASET

Performance	***# of clusters***	***Train Time***	***Cross Validation Score***	***Test Accuracy***
NN + KMeans + Randomized Projections	2	56.58	0.93	0.90
NN + EM (Randomized Projections)	2	59.30	0.97	0.96

V. Analysis

A. *Step 3*

KMeans as a clustering method, and PCA or Manifold Algorithm as dimensionality reduction are selected in both of datasets, Dropout and AIDS, in Step 3.

At first, in dropout dataset, the graphs of silhouette scores in PCA and Manifold Algorithm are remarkably similar in that the value plunges when it gets to 3 and increases from 4 to 5. However, the values are a little different in terms of the number of clusters. The silhouette plot of KMeans shows the highest value when the number of clusters is 2, but the values are all positive even when the number of clusters is 6. However, when we investigate the KMeans intercluster distance map, we can see that the circles are duplicated even when the value is 3. Therefore, the value of clusters is 2 in both PCA and Manifold Algorithm.

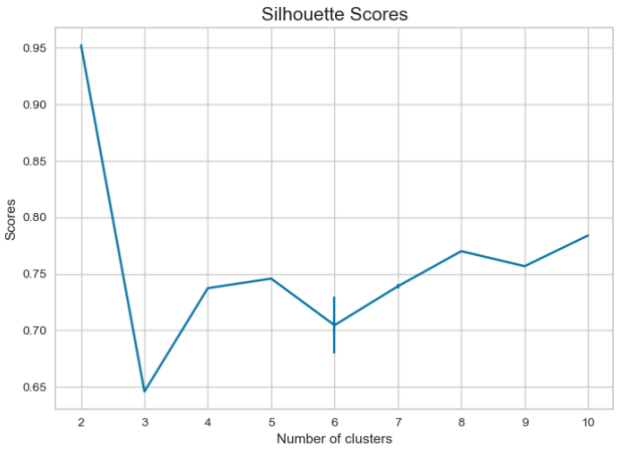

Fig. 1. The silhouette scores in terms of the number of clusters in KMeans with the dimensionality reduction using PCA in Dropout dataset

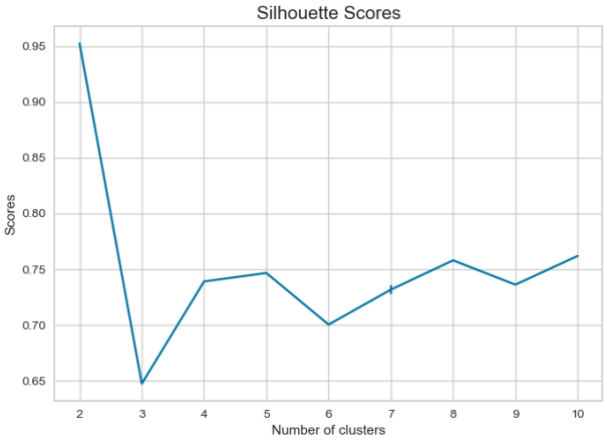

Fig. 2. The silhouette scores in terms of the number of clusters in KMeans with the dimensionality reduction using Manifold Algorithm in Dropout dataset

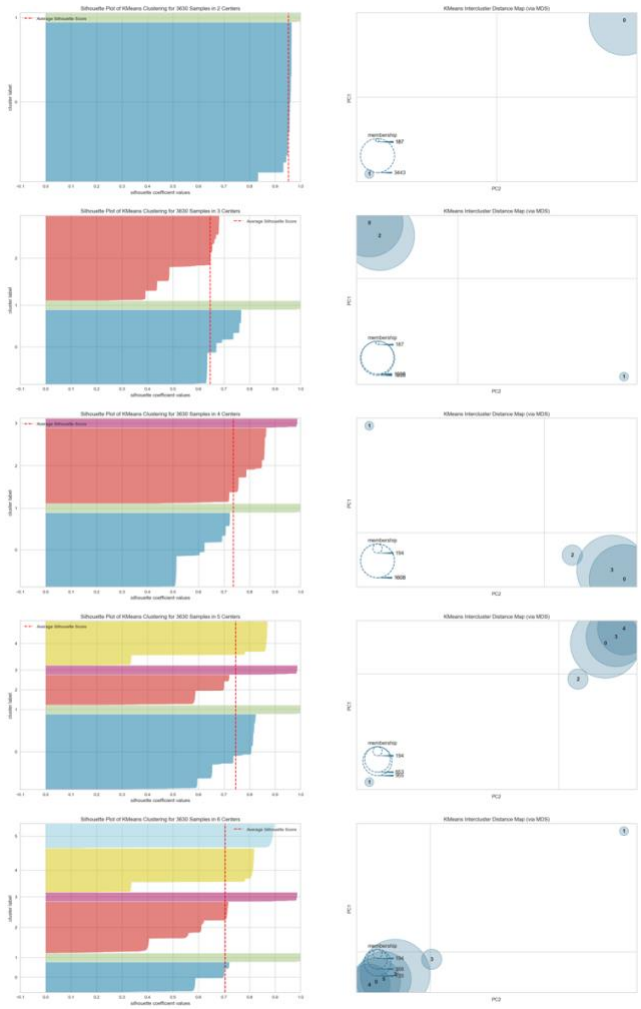

Fig. 3. The silhouette plot of KMeans clustering and KMeans intercluster distance map in terms of the number of clusters in KMeans with the dimensionality reduction using PCA in Dropout dataset. The values of clusters ranges from 2 to 6, from the top image to the bottom image.

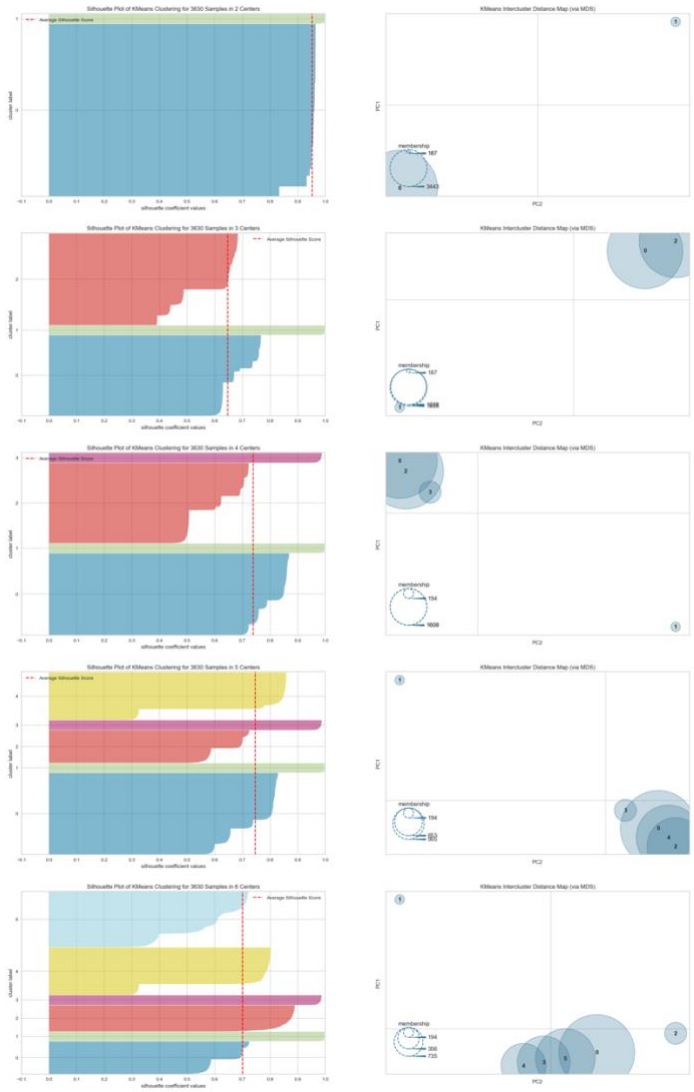

Fig. 4. The silhouette plot of KMeans clustering and KMeans intercluster distance map in terms of the number of clusters in KMeans with the dimensionality reduction using Manifold Algorithm in Dropout dataset. The values of clusters ranges from 2 to 6, from the top image to the bottom image.

For the AIDS dataset, PCA and Manifold Algorithm also shows similar results: similar silhouette scores, silhouette maps and KMeans intercluster distance maps. As the value of cluster increases, the silhouette score is the highest when the number of cluster is 3, but when they are 4 or 5, the value is higher than the other cases. However, when the number is over 6, the silhouette score plunges. In Figure7 and 8, we can see the same results as above. The average silhouette values are generally similar when the number of clusters are from 2 to 6. However, the circles are duplicated in the intercluster maps when the number of clusters is 6. Therefore, we can assume that the number of clusters are proper when the number is from 3 to 5, but the best value is 3.

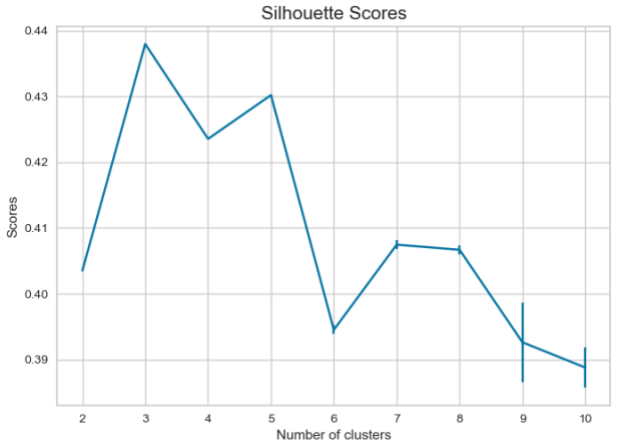

Fig. 5. The silhouette scores in terms of the number of clusters in KMeans with the dimensionality reduction using PCA in AIDS dataset

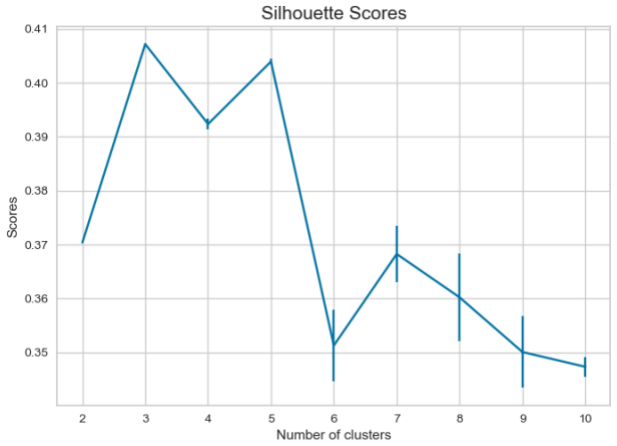

Fig. 6. The silhouette scores in terms of the number of clusters in KMeans with the dimensionality reduction using Manifold Algorithm in AIDS dataset

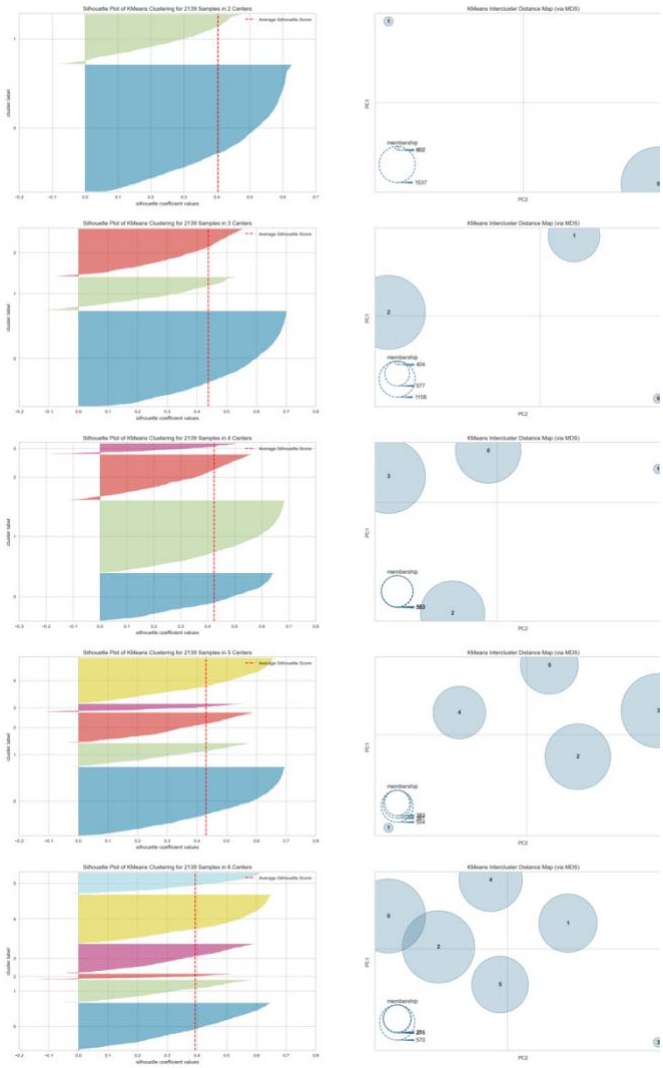

Fig. 7. The silhouette plot of KMeans clustering and KMeans intercluster distance map in terms of the number of clusters in KMeans with the dimensionality reduction using PCA in AIDS dataset. The values of clusters ranges from 2 to 6, from the top image to the bottom image.

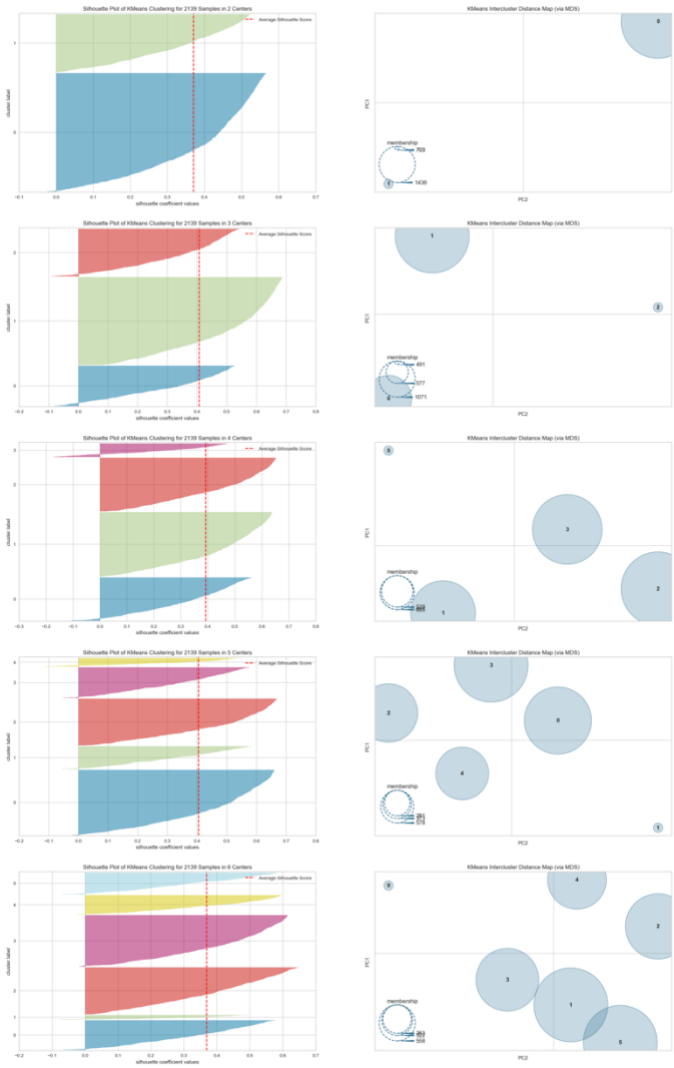

Fig. 8. The silhouette plot of KMeans clustering and KMeans intercluster distance map in terms of the number of clusters in KMeans with the dimensionality reduction using Manifold Algorithm in AIDS dataset. The values of clusters ranges from 2 to 6, from the top image to the bottom image.

B. Step 4

In Step 4 and Step5, we use the students' dropout datasets looking into the results of neural networks using the randomized projections and manifold algorithms as dimensionality reduction. Before that, let's see the results again. comparing the original data, the training time is reduced in both of the randomized projections and manifold algorithm because the dimensionality reduction reduces the number of features, which save the computation cost. However, when we consider that the features are reduced to 9 or 2 from 36, the training time does not show the significant difference. It is also interesting because the smaller number of features mean the smaller number of weights in the neural networks, saving the memory space and calculating time. It might be due to the strong power of computer capacity and the relatively lower number of instances. On the other hand, for randomized projections, the accuracy performance is much better than the manifold algorithm. I think it might indicate the importance of number of features which can contain the more information. On the other hand, the randomized projection

outperforms the original data, which shows that the dimensionality reduction cannot harm the data integrity enhancing the performance.

TABLE VI. NUMBER OF COMPONENTS IN DIMENSIONALITY REDUCTION IN DROPOUT DATASET

Performance	***# of components***	***Train Time***	***Cross Validation Score***	***Test Accuracy***
Original Data	-	65.47	0.75	0.89
NN + Randomized Projections	9	58.85	0.87	0.86
NN + Manifold Algorithm	2	57.90	0.57	0.60

C. Step 5

In Step 5, we can see find the powerful tools, which is the combination of dimensionality reduction and clustering algorithms in neural networks. The accuracy data in both of projected data outperform the original training, and the combination of EM is incredible when considering that the EM performance is not the best in the previous experiments. By reducing the dimensionality, we can reduce the noise of the data but pertain the primary characteristics in the data, which can improve the performance in time complexity and accuracy. Moreover, it can also help find the insights through clustering by grouping the similar data points. Especially in neural networks, dimensionality reduction can help neural networks learn features effectively when extracting significant features from the data.

TABLE VII. NUMBER OF COMPONENTS IN DIMENSIONALITY REDUCTION IN DROPOUT DATASET

Performance	***# of clusters***	***Train Time***	***Cross Validation Score***	***Test Accuracy***
Original Data	-	65.47	0.75	0.89
NN + KMeans + Randomized Projections	2	56.58	0.93	0.90
NN + EM + (Randomized Projections)	2	59.30	0.97	0.96

VI. CONCLUSION

In this assignment, we can find how the clustering and dimensionality algorithms work in unsupervised learning environment using two clustering algorithms and four dimensionality reduction algorithms. We apply those algorithms or the combination of algorithms to the two datasets, students' dropout and AIDS datasets. Even though they have their own characteristics in the datasets, it is interesting the different clustering algorithms indicate the right number of clustering, 2, which is the actual number of target data in each dataset.

The more interesting facts about the experiments are that dimensionality reduction does not harm the results of neural network, which reduce the number of features improving the computational performance. More than that, the combinations of dimensionality reduction and clustering algorithms enhance the performance of neural networks by helping the neural networks find the structures of the datasets much faster and more correctly.

Since selecting a good number of clustering or components is not easy, it would be helpful to find more methods in visualizing EM or other algorithms which Yellow Brick does not provides. Normalized datasets would be good to apply by managing the scales in features and reducing the performance of the gradient descent in neural networks.

REFERENCES

[1] "Unsupervised Learning: Definition, Techniques, and Applications." Unsupervised, unsupervised.com/resources/blogs/what-is-unsupervised-learning.

[2] "Predict Students Dropout and Academic Success." UCI Machine Learning Repository [https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success], University of California, Irvine, School of Information; Computer Sciences, 2021

[3] "AIDS Clinical Trials Group Study 175." UCI Machine Learning Repository, [https://archive.ics.uci.edu/dataset/890/aids+clinical+trials+group+study+175], University of California, Irvine, School of Information; Computer Sciences, 2021