Project 3: Unsupervised Learning & Dimensionality Reduction
CS7641

Anita Rao
arao338@gatech.edu

1 DATASETS

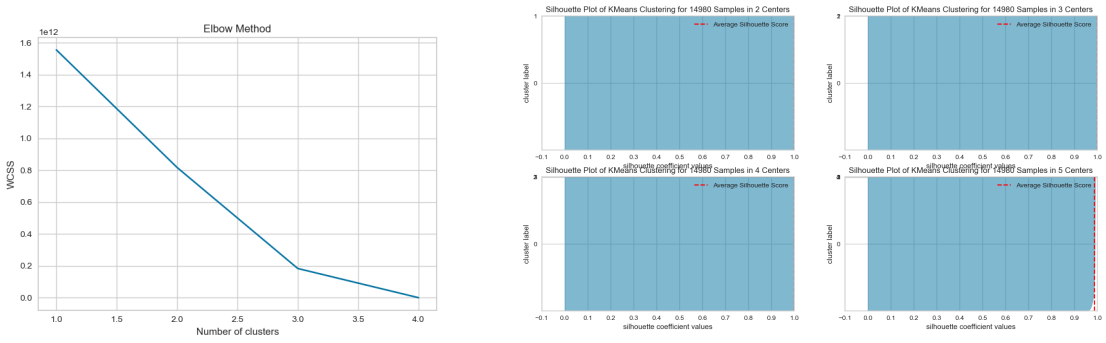
The two datasets used are the EEG Eye State (14980 samples), which takes a continuous EEG measurement and classifies the eye-state as "Open" or "Closed", and the Flight Delay dataset (26969 samples), which considers airline information and classifies flights as "Delayed" or "Not Delayed". These two datasets are interesting because they are non-trivial from a machine learning standpoint. While the datasets are labeled, we can use them for unsupervised learning (UL) and identify any logical mappings between the data clusters and their target labels.

2 EXPERIMENT 1: CLUSTERING ALGORITHMS

We will look at K-Means and Gaussian Mixture (using Expectation Maximization technique) clustering algorithms. K-Means groups data into K clusters, while GM finds the probability that the data belongs to the various K clusters. GM assumes clusters come from normal distributions, while K-means assumes clusters are spherical in shape.

1.1 Determine K

The Silhouette method identifies a silhouette score from -1 to 1, which characterizes *both* the intra-cluster density and inter-cluster spacing based on the number of K clusters. In the cases



(a) *(b)*

Figure 1—(a) EEG Dataset: Determine K with Elbow Method
(b) EEG Dataset: Determine K with Silhouette Method

where silhouette score is identical, elbow method was used to determine K. The elbow method uses inertia to find the K that maximizes intra-cluster *density*.

In Figure 1a for the EEG dataset above, we see that the elbow method yields K=3, while the silhouette method in 1b has the same score for multiple values of K. Therefore we will proceed with K=3 as the optimal number of clusters.

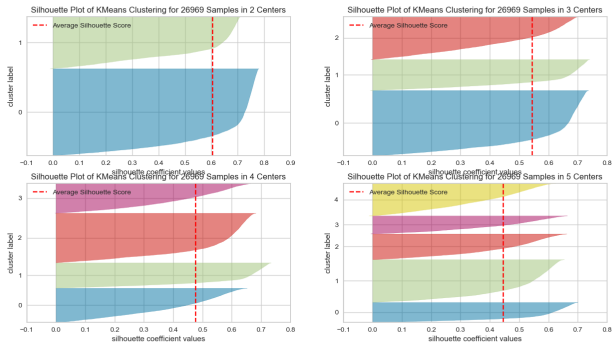
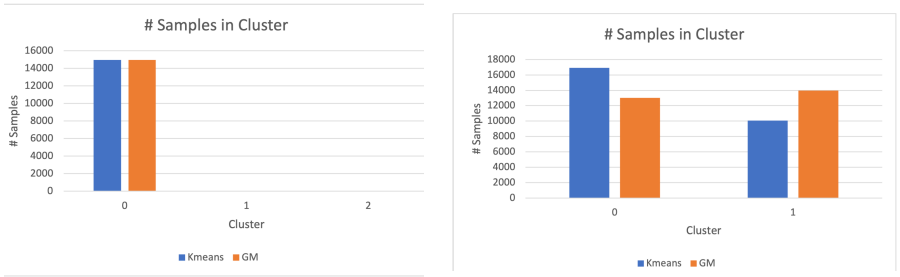


Figure 2—Flight Dataset: Determine K with Silhouette Method

In Figure 2 for the Flight Dataset above, we see that the silhouette score is maximized at 0.6 for K=2, therefore we will proceed with K=2 as the optimal number of clusters.

1.2 Cluster description

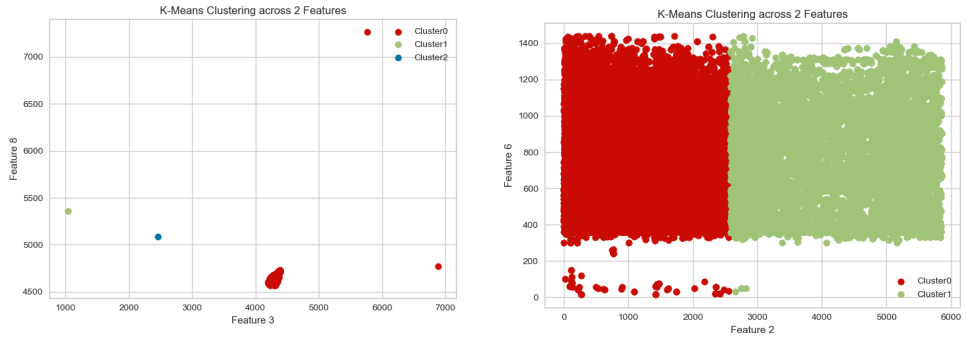


(a) ***(b)***

Figure 3—***(a)*** EEG Dataset: Samples per cluster
(b) Flight Dataset: Samples per cluster

Figure 3a shows almost *all* samples fall into cluster 0 for EEG dataset. 3b shows samples fall into both clusters without a natural alignment by label for Flight dataset. This is true for both K-Means and GM models. The visuals in Figure 4 below also demonstrate this. 4a shows the majority of points gathered in red Cluster0, with a few outliers, for EEG dataset. 4b shows the

points separated nicely into Cluster0 and Cluster1 for Flight dataset, but these clusters do not necessarily align with the true target labels.



(a) *(b)*

Figure 4—(a) EEG Dataset: Cluster separation for Feature 3 by Feature 8
(b) Flight Dataset: Cluster separation for Feature 2 by Feature 6

1.3 Analysis

It is interesting that almost all samples fall into a single cluster for the EEG dataset. This makes sense at this stage, as the EEG values obtained from 14 sensors on the scalp have extremely similar values without obvious differentiation. This makes separating the samples into distinct clusters difficult. With dimensionality reduction, we can transform the data to a lower dimensional space that can allow for more distinct clustering. Let's explore the Flight Dataset as this had more meaningful clustering results. As we compare the Flight Dataset clusters to the classification labels from assignment #1, we see that they do not line up exactly with the labels.

Table 1—Count of Labels (0,1) in Clusters (0,1)

	KMeans - Cluster 0	**KMeans - Cluster 1**	**GM - Cluster 0**	**GM - Cluster 1**
True Label 0	9028	5906	7027	7907
True Label 1	7884	4151	5984	6051

Table 1 shows that we do not have an exact or natural alignment of the true label values (0,1) with the cluster numbers (0,1). The fact that delayed flights (1) and non-delayed flights (0) appear in both clusters without an obvious alignment indicate that the clusters signify some other differentiation between samples that is not necessarily delayed vs non-delayed flights.

Figure 5 shows K-Means and GM models with covariance type 'sphere' and 'tied' outperformed the other GM models. It appears that spherical clusters (common across K-Means and GM-Sphere) perform well on this data. The performance of these algorithms can be improved through smarter preprocessing of the data, such as normalization and applying dimensionality

reduction. For the EEG dataset in which almost all samples fell into one cluster, these techniques will be especially important to determine differentiation, if any, between samples.

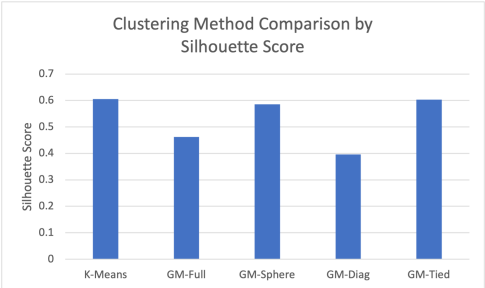
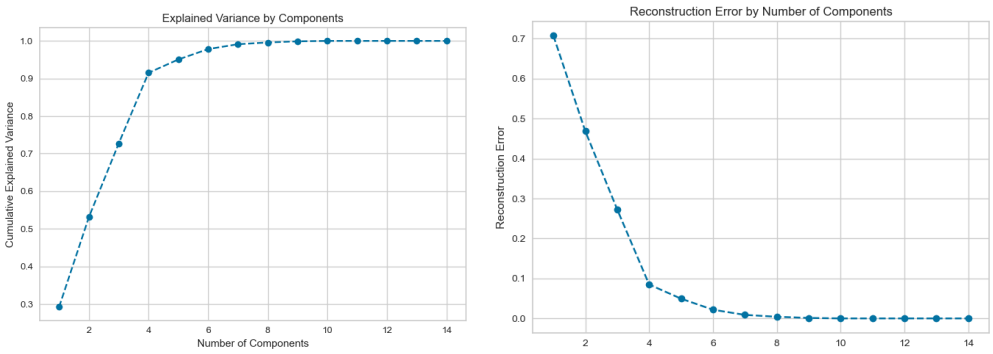


Figure 5—Flight Dataset: Comparison of Clustering Algorithms by Silhouette Score

3 EXPERIMENT 2: DIMENSIONALITY REDUCTION

Dimensionality reduction projects an existing set of features onto a lower-dimensional space resulting in ideally fewer features and reducing complexity of the clustering problem. PCA, ICA, Gaussian Random Projection (RP) and Backward Elimination (BE) techniques were used. PCA maximizes variance of the features and reconstructs the data with fewest *principal* components, while ICA transforms the features such that the new components are as mutually independent as possible. RP is useful if the original number of features is quite large and there are time constraints, which was not the case for the 2 datasets. BE is the most intuitive, as the same feature space was maintained but certain features were eliminated using KNN.

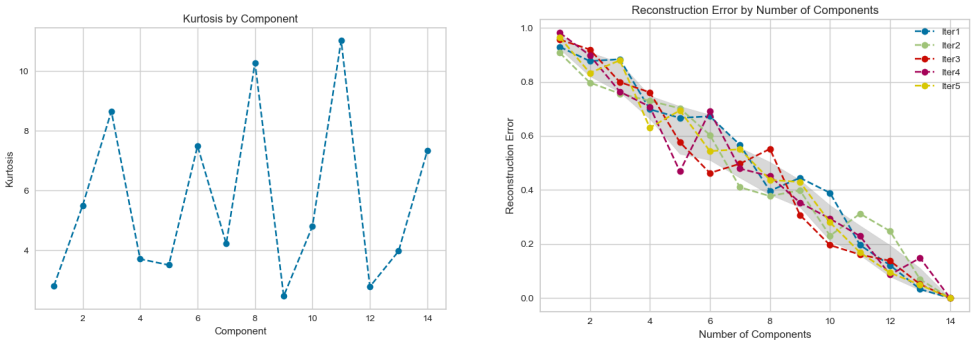
For EEG dataset, Figure 6a shows distribution of PCA eigenvalues in which 4 components explain >80% of the data variance. 6b shows 4 components sufficiently reconstruct the data.



(a) ***(b)***

Figure 6—(a) EEG Dataset: PCA Explained Variance by Components
(b) EEG Dataset: PCA Reconstruction Error by Components

ICA components with low kurtosis (indicating gaussian noise) were identified in Figure 7a. Taking the mean kurtosis over the components resulted in a monotonically increasing plot, so components with low kurtosis were discarded.

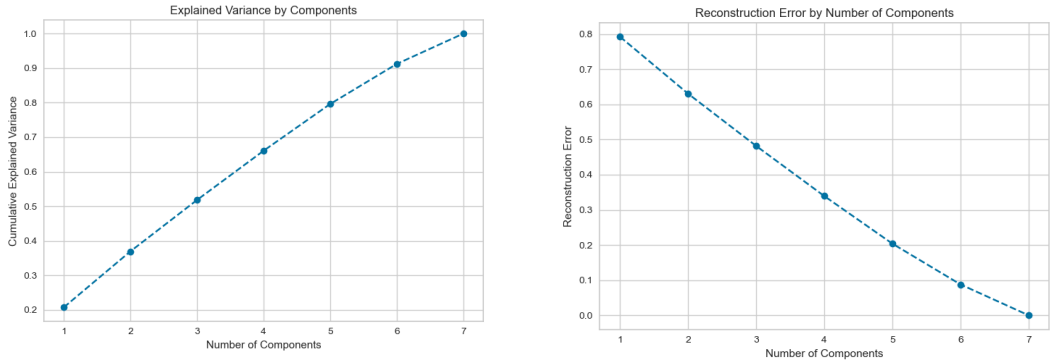


(a) ***(b)***

Figure 7—(a) EEG Dataset: ICA Kurtosis by Components
(b) EEG Dataset: RP Reconstruction Error by Components

Figure 7b shows for RP, 12 components result in 20% reconstruction error. There was clearly variation between runs, but ultimately they trended in a similar direction.

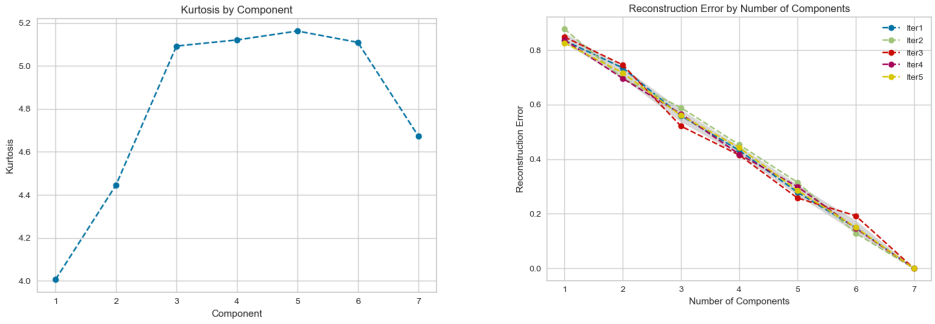
To determine if ICA projection axes were "meaningful" I projected the identity matrix onto the new feature space (Isbell, 1998). No particular features stood out, and because the features correlate to continuous values obtained by EEG electrodes it is difficult to determine if the ICA axes translate to any real-world takeaway. The components that were kept in ICA were 3, 6, 8, 11, reducing the number of features from 14 to 4.



(a) ***(b)***

Figure 8—(a) Flight Dataset: PCA Explained Variance by Components
(b) Flight Dataset: PCA Reconstruction Error by Components

For Flight dataset, Figure 8a shows distribution of PCA eigenvalues where 5 components explain >80% of variance, and 8b shows sufficient reconstruction of the data. There is not a significant decrease in components here, with each component seemingly contributing equally to the variance.



(a) *(b)*

Figure 9—(a) Flight Dataset: ICA Kurtosis by Components

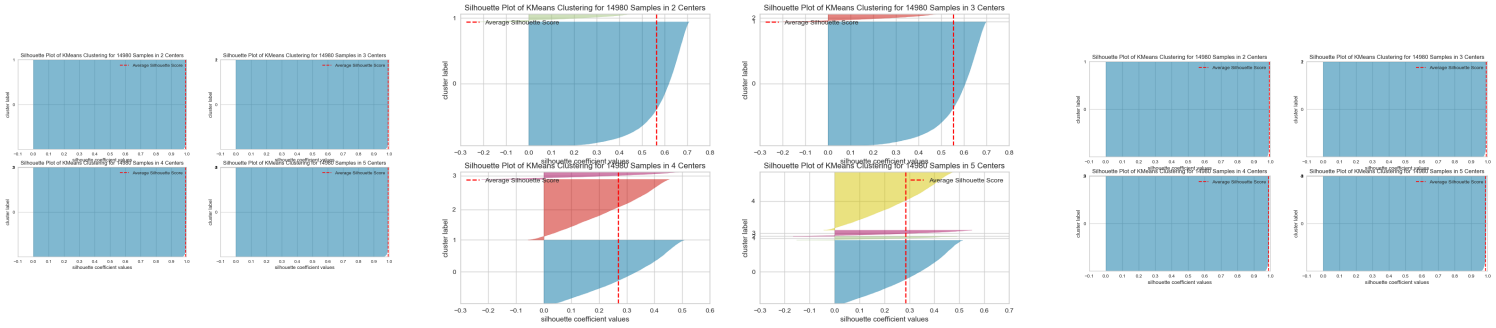
(b) Flight Dataset: RP Reconstruction Error by Components

Figure 9a shows ICA components 3, 4, 5 and 6 were maintained due to a higher kurtosis. Similar to the EEG dataset, projecting the identity matrix onto the new feature space did not result in ICA components with particularly significant weights for specific features, so no real-world takeaway could be made. 9b shows for RP, 6 components result in 20% reconstruction error, with a lower overall variance across runs than for the EEG dataset above.

The final dimensionality reduction technique used was Backward Elimination (BE), which uses 3-Nearest Neighbors to identify the "worst" feature that should be removed. K=3 yielded best results compared to larger values of K. For both datasets, this iterative feature selection was performed until one half of the original set of features remained which was the default behavior of the algorithm. The EEG dataset maintained features 2, 3, 6, 9, 10, 12, 14 which correspond to specific electrodes capturing EEG signals (difficult to interpret "meaningful" relationships to the original data without a physician). The Flight Dataset maintained features 1, 6, and 7, which meaningfully corresponds to the Airline, Departure Time, and Flight Duration features.

4 EXPERIMENT 3: DIMENSIONALITY REDUCTION + CLUSTERING

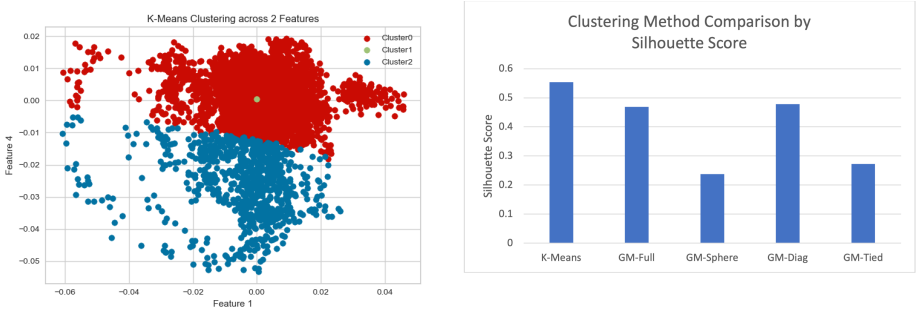
The optimal number of clusters for PCA, ICA, and BE were determined using silhouette scores (described in Experiment 1). RP yielded uninteresting findings compared to other algorithms as the dimensionality reduction was not significant, and is not shown here.



(a) *(b)* *(c)*

Figure 10—(a) EEG Dataset: Determine K with Silhouette Method after PCA
(b) EEG Dataset: Determine K with Silhouette Method after ICA
(c) EEG Dataset: Determine K with Silhouette Method after BE

Figure 10 shows ICA (10b) resulted in smaller average silhouette scores than PCA (10a) and BE (10c), but upon further examination, PCA and BE exhibited the behavior in Experiment 1 with mostly all samples falling into one cluster. ICA yielded the most interesting results with K=3, as we selected mutually independent features that removed noise from the original dataset. Although K=3, the points separate into 2 major clusters (Cluster0 and Cluster2) with a single outlier in Cluster1.

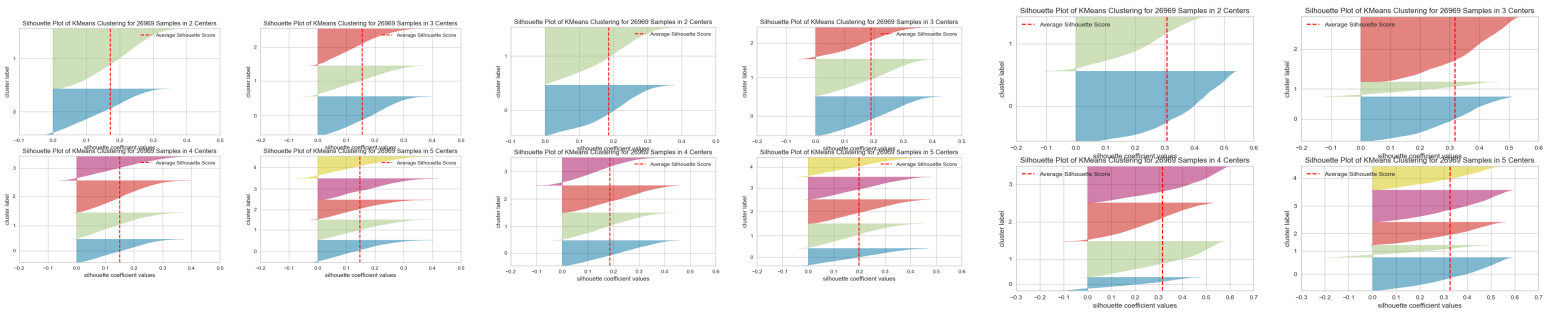


(a) *(b)*

Figure 11—(a) EEG Dataset: Cluster separation for Feature 1 by Feature 4 after ICA
(b) EEG Dataset: Comparison of Clustering Algorithms by Silhouette Score after ICA

Figure 11a shows 2 main clusters after ICA is performed on the EEG dataset. This is a vast improvement from Experiment 1, where all samples gathered into 1 cluster. Figure 11b compares the various clustering algorithms (with different covariance types used with GM), and K-Means outperforms the others in terms of silhouette score. This is perhaps because the EEG dataset does not resemble underlying gaussian distributions, which GM assumes.

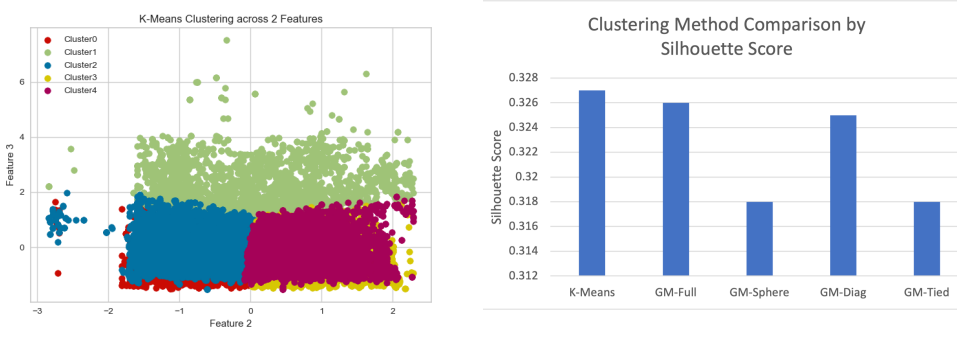
Figure 12 shows BE (12c) yielded the highest silhouette value at 0.33 with K=5, over PCA (12a) and ICA (12b). However, this score is lower than in Experiment 1, with score=0.6 at K=2, indicating that the inter-cluster spacing of the Flight dataset is less optimal after BP. It is likely that reducing the number of features was not an improvement over using all of the original features. This is further supported by PCA performed in Figure 9a, showing that all components contribute seemingly equally to the overall variance, so removing features may ultimately hurt clustering performance.



(a) *(b)* *(c)*

Figure 12 *— **(a)*** Flight Dataset: Determine K with Silhouette Method after PCA
(b) Flight Dataset: Determine K with Silhouette Method after ICA
(c) Flight Dataset: Determine K with Silhouette Method after BE

Figure 13a shows 5 clusters within the Flight dataset, mapping Feature 3 against Feature 2 after BE. 13b shows that K-Means, GM-Full (individual clusters can take any shape), and GM-Diag (cluster shapes are axis-aligned) outperform the others. The GM performance can be explained by the Flight dataset potentially exhibiting some underlying gaussian distributions.

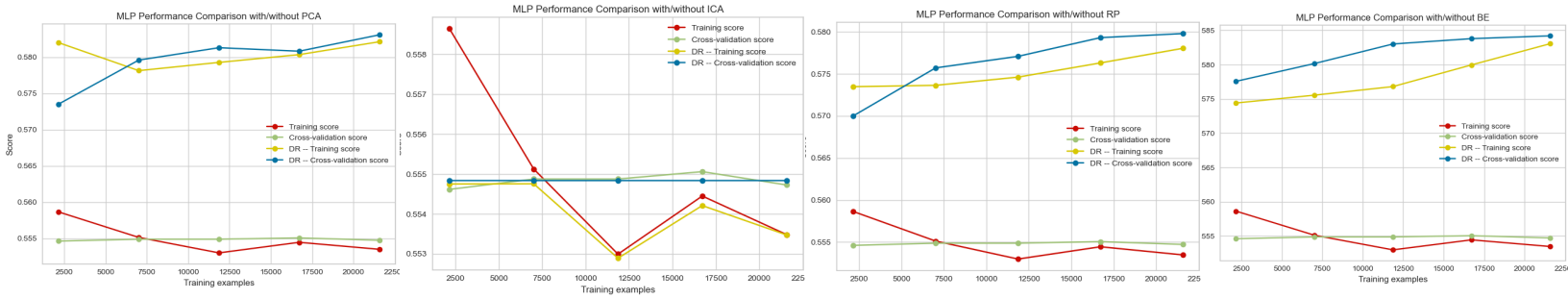


(a) *(b)*

Figure 13 *— **(a)*** EEG Dataset: Cluster separation for Feature 2 by Feature 3 after BE
(b) Flight Dataset: Comparison of Clustering Algorithms by Silhouette Score after BE

5 EXPERIMENT 4: DIMENSIONALITY REDUCTION + NEURAL NETWORK

In this experiment, we fed the transformed Flight data from Experiment 2 to the multilayer perceptron network constructed in Assignment #1. We compared this network performance to the original network that used raw data for training.



(a) *(b)* *(c)* *(d)*

Figure 14 *—(a)* MLP Performance Comparison with/without PCA
(b) MLP Performance Comparison with/without ICA
(c) MLP Performance Comparison with/without RP
(d) MLP Performance Comparison with/without BE

Dimensionality reduction yields better performance of the neural network than using raw data, as shown in Figure 14. Only ICA (14b) does not obviously perform better than the original neural network, perhaps if selected components do not sufficiently reconstruct the data.

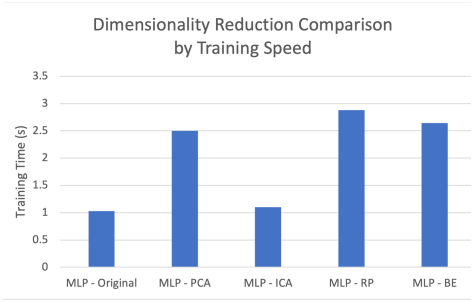
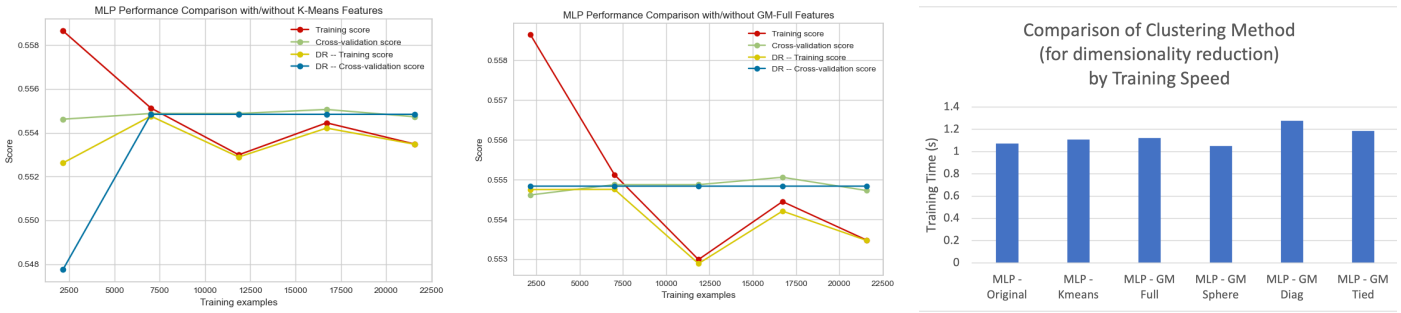


Figure 15 — Comparison of MLP with Dimensionality Reduction by Training Time

Figure 15 shows the various neural networks also differ in their training time, with the original neural net completing training the fastest. This is likely due to scaling the values, resulting in more distinct entries. ICA runs fairly fast as the values are very close to 0. The remaining methods, PCA, RP, and BE, transform the raw data into continuous values that include large positive and negative values -- this is likely why the training time is longer for these algorithms.

6 EXPERIMENT 5: CLUSTERING + NEURAL NETWORK

In this experiment, dimensionality reduction was performed by substituting clusters as features. Using the Flight dataset, we reused steps performed in Experiment 1 to determine K=2. K-Means and GM (with different covariance types) were performed, generating new features with a 1 if the sample belongs to that cluster and a 0 if it does not. The performance of the neural network using GM as features was the same across all GM types as captured in 16b.



(a) *(b)* *(c)*

Figure 16—(a) MLP Performance Comparison with/without K-Means
(b) MLP Performance Comparison with/without GM
(c) Comparison of MLP with Dimensionality Reduction by Training Time

Using clusters as features does not improve the performance of the neural network as shown in Figure 16a and 16b. However, it does not *worsen* model performance either compared to the original neural network. This indicates that the clustered data is "as good" a predictor of flight delays as the original dataset, but is not neatly aligned with the target values. Other forms of dimensionality reduction above *did* seem to improve performance, so it may be that using clusters as features is not ideal due to the nature of this dataset. 16c shows training times are similar across networks, since 0/1 feature values do not contribute greatly to computation time.

7 REFERENCES

1. Elena IKONOMOVSKA'S web page. (n.d.). Retrieved February 22, 2021, from http://kt.ijs.si/elena_ikonomovska/data.html

2. UCI machine Learning Repository: Eeg eye state data set. (n.d.). Retrieved February 22, 2021, from https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State

3. C. L. Isbell and P. Viola. (1998). Restructuring sparse high dimensional data for effective retrieval. *Advances in Neural Information Processing Systems* 11, 480–486.