Kunal Shah - CS7641- Machine Learning

Project 3: Unsupervised Learning

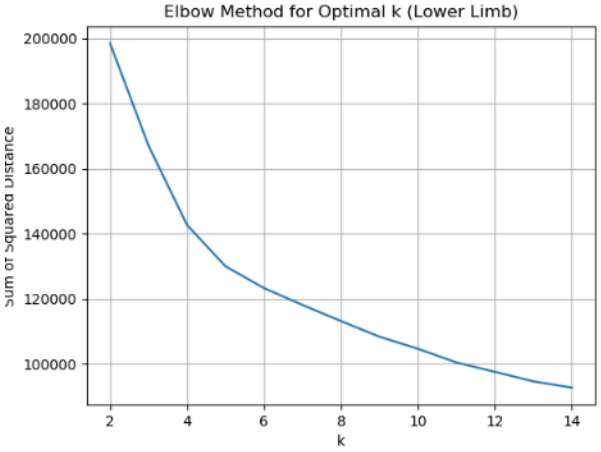
Introduction:
This project explored various unsupervised learning methods. The first two were clustering algorithms, K-means (KM) and Expectation Maximization (EM), while the last 4 were dimensionality reduction (DR) techniques including PCA, ICA, Randomized Projections (RP), and Linear Discriminant Analysis (LDA). Several experiments were run with different combinations of the DR and clustering techniques, ultimately creating inputs for a neural network classifier.

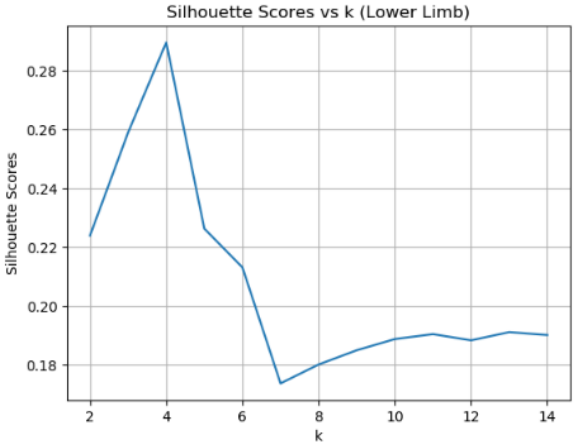
Brief Dataset Overview:
The Lower Limb dataset explored was a collected at the Shirley Ryan AbilityLab using a powered prosthetic leg on 2 different able-bodied users. There were 132 features and 2386 total examples. The labels for this data consist of 8 possible classes which correspond to different parts of the gait cycle (heel contact, mid-stance, toe off, and mid-swing) and different ambulation modes. These 8 classes are Level Walking (LW), Toe Off (TO), Ramp Descent (RD), Stair Descent (SD), Standing Toe Off (STTO), Standing Heel Contact (STHC), Mid-Swing (MSW), and Mid-Stance (MST). The dataset is imbalanced as the LW and TO classes compose over half of all examples collected and one third of all examples are TO.

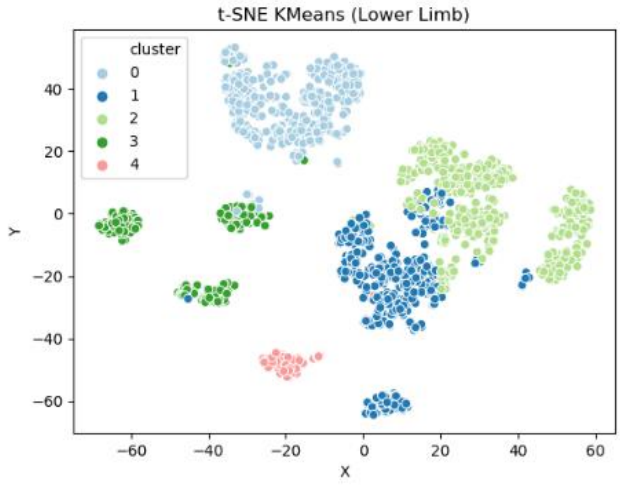
Experiment 1: Clustering

For K Means, the elbow method and silhouette scores were used to determine k. For EM, AIC/BIC and silhouette scores were used to determine the number of components. For all clustering experiments, t-SNE was used to visualize the clusters. This is a dimensionality reduction technique that is commonly used for visualizing high dimensional datasets. The datasets were reduced to 2 dimensions from the original high-dimensional space. While the plots can be used as visual aids, t-SNE is non-linear and probabilistic so it is not always consistent, and the t-SNE also does not preserve distances nor density. t-SNE also can produce shapes in the charts that are not always a true representation of the data, which will be seen later in the analysis.

KMeans:
For D1, a k value of 5 was selected based on the elbow plot. Using t-SNE, the 5 clusters can be visualized below. 4 clusters were also used (not shown) as the silhouette plot was maximal at k=4. From the t-SNE plot, 4 clusters looked the same but combined clusters 1 and 2 into one large cluster.



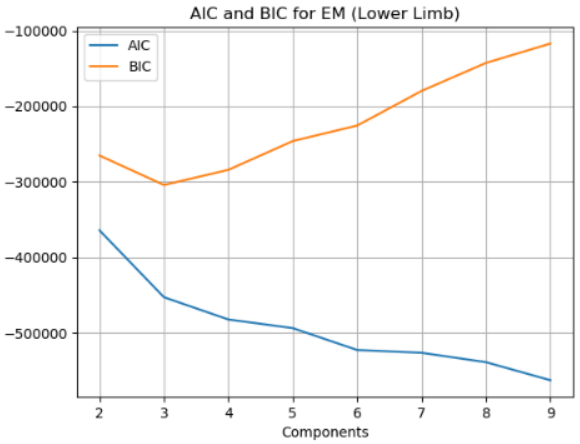


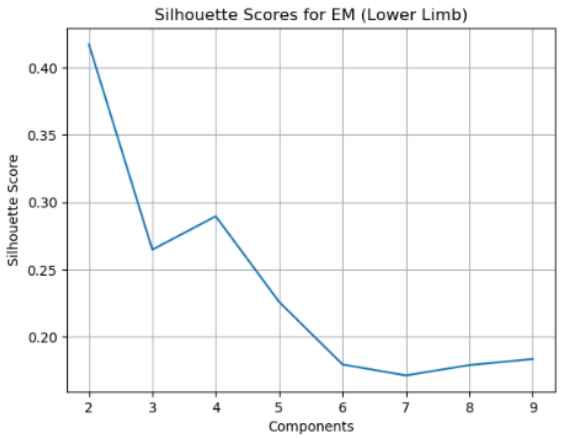


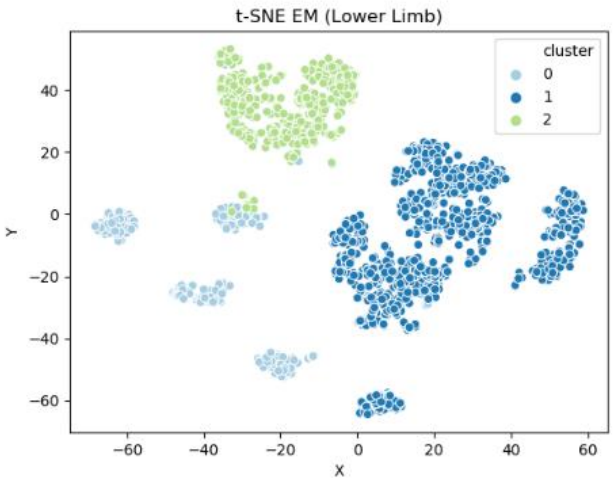
EM:

For EM, 3 components were selected. The AIC/BIC scores were very negative. With increasing model complexity, BIC also increases, while with increasing likelihood, BIC decreases. This means that number of components corresponding to the lowest BIC value would be the desired number of clusters. While AIC was constantly decreasing, BIC had a minimum at 3 components. BIC was used to determine the number of components because it penalizes the number of components more strongly than AIC (preventing model complexity from increasing and reducing overfitting). From the t-SNE plot the 3 clusters are fairly well separated, although 'cluster 0' is quite sparse.

In D1, there are 8 possible classes, but the clustering methods only found 5 and 3, respectively. This shows that the features for certain ambulation modes may not be as distinct as currently perceived. One thing to note however is that this dataset is imbalanced, heavily favoring two classes. Looking at the t-SNE plots for D1, the argument could be made that each of the 5 small clusters on the left could be their own classes and the three in the upper right could round out the 8 possible classes. Creating a balanced dataset and rerunning the clustering algorithms may produce a graph with 8 clusters similarly shaped/sized.



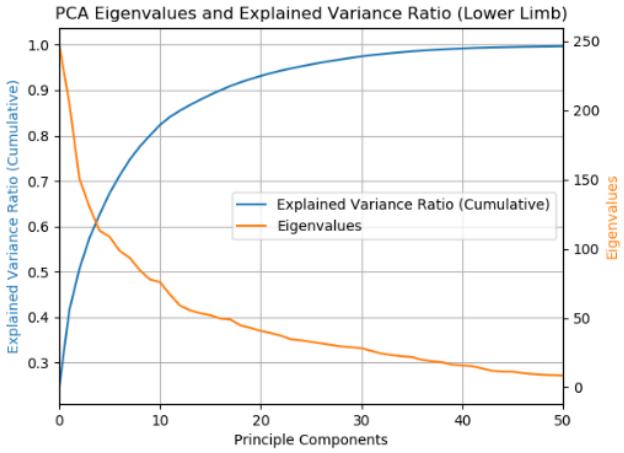


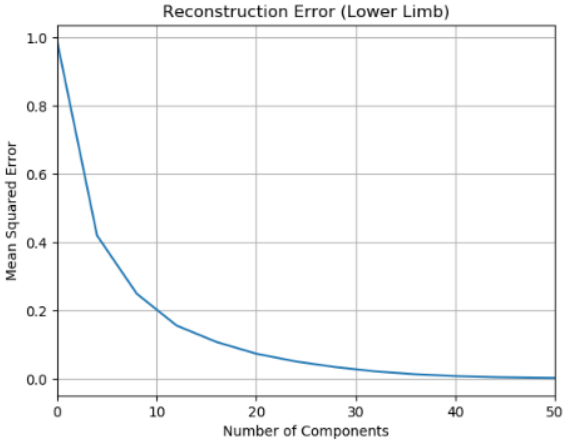


Experiment 2: Dimensionality Reduction

PCA:

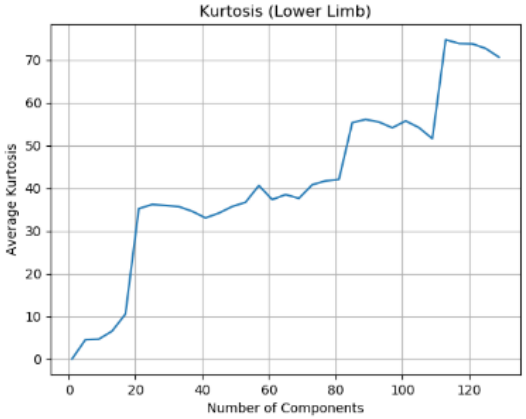
For PCA on D1, 25 components were selected as 25 components explained ~95% of the variance in the data. The reconstruction error for 25 components was around 10%.





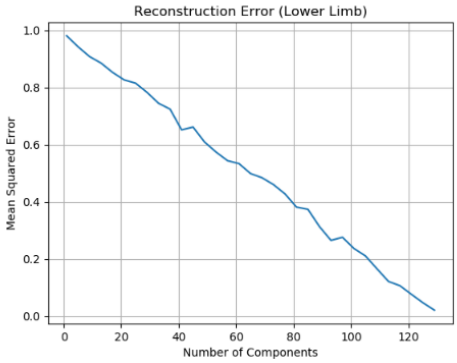
ICA:

For ICA, kurtosis was used to determine the nongaussianity of the dataset. Kurtosis is a measure of the sharpness or flatness of a distribution relative to a normal distribution. As ICA aims to create independent components, the higher non gaussian (farther from 0 in a positive or negative direction) behavior indicates higher level of independence. For D1, 113 components were selected because that corresponded to the highest kurtosis, although 20 components were also investigated (not shown) as that was the first 'peak' and reduced the feature by a greater degree. The plot of reconstruction error is in the appendix. The error at 15 components was 10% and at 113 components was virtually 0.

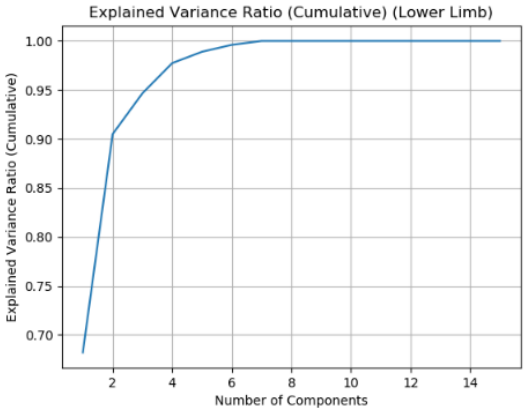


RP:

For RP, the reconstruction error for D1 was a linear decrease. Because of this, a threshold of 10% error was set, corresponding to 120 components.



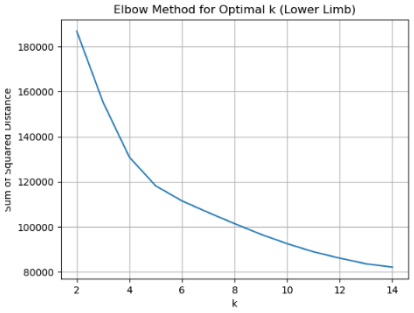
LDA:
Explained variance was used to evaluate LDA. For D1, 3 components were selected as they represented 95% of the variance in the data.

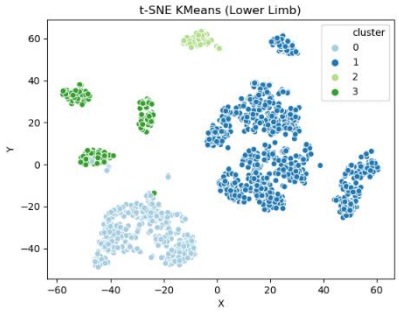


Experiment 3: Dimensionality Reduction + Clustering

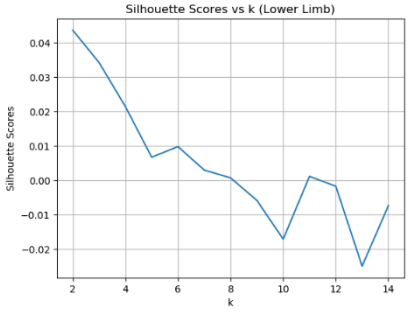
For this section, each combination has an elbow method or BIC/AIC plot, silhouette score plot, and t-SNE plot. For brevity, the majority of charts are in the appendix, leaving only the charts that were used to select the number of clusters. The values found from Experiment 2 were used for each of the DR techniques in this experiment.

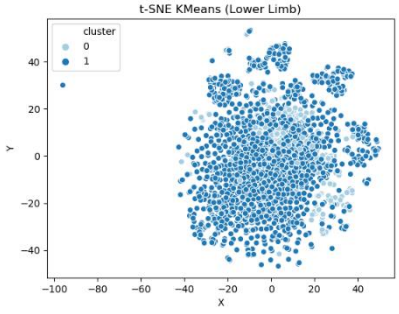
PCA + KMeans:
Using PCA prior to clustering suggested that k=4 was the best number of clusters. From the visualization, the clusters are separated although, again, the dataset imbalance could explain the varying sizes of the clusters. Compared to the KM plot in experiment 1, the distributions are almost identical (although rotated/mirrored). This makes sense because PCA finds the components with the most influence but does not necessarily transform the data, just relies on unimportant features less (which are hopefully just noise).



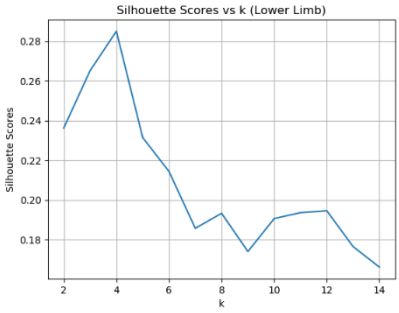


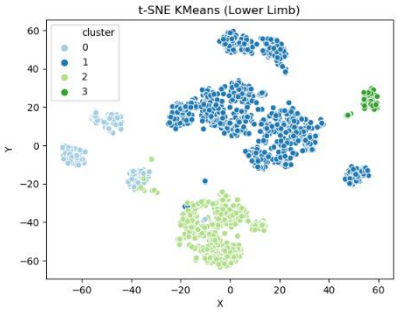
ICA + KMeans:
Doing ICA prior to clustering forces all of the data together. The silhouette score was highest at 2 so that was selected for k, but no matter the number of clusters, this dataset would not be very separable. ICA attempts to build statistically independent components, and from the mass created by the ICA algorithm, that would imply that dataset has features that are not statistically independent. This could be attributed to the fact that 6 features are extracted from 22 mechanical sensors channels. The features for each sensor may not be statistically independent. Additionally, many of the sensor channels come from the same channel (i.e. 6 axis load cell with Fx, Fy, Fz, Mx, My, Mz) which could also lead to some statistical coupling of features.



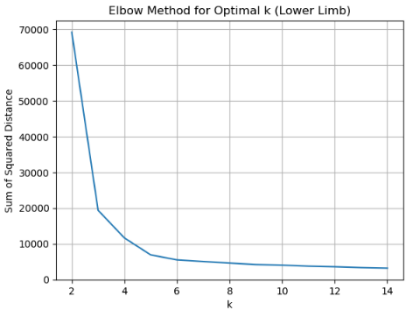


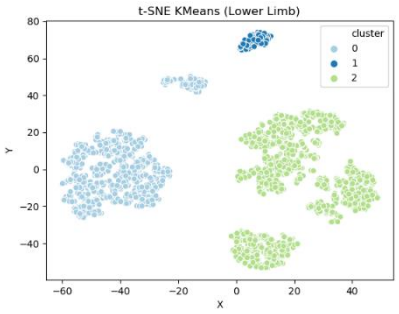
RP + KMeans:
RP has similar outcome as PCA. Both suggested a k=4 and have similar distributions to the original KM plot.



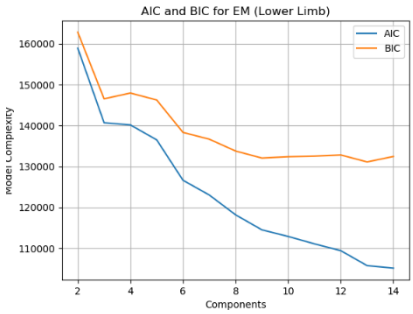


LDA + KMeans:
LDA was able to consolidate some of the smaller clusters seen in the original KM plot and k=3 was found from an elbow plot. This method did the best at clustering, with no outliers lying in regions of other clusters.

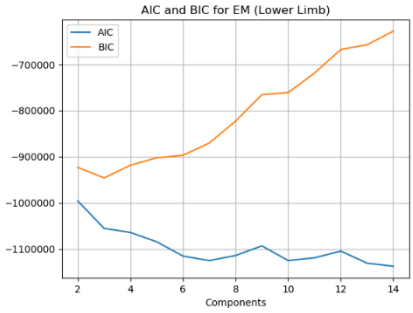




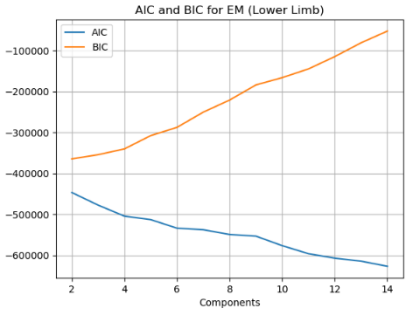
PCA + EM:
PCA with EM and KM perform similarly. EM chose to have one less cluster (3) and did a better job of clustering, with the 3 groups easily separable. The t-SNE plot is in the appendix, but it is the same as above, with clusters 2 and 3 combined.

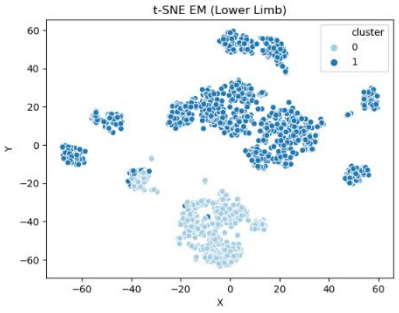


ICA + EM:
For ICA, 3 clusters were selected using AIC/BIC, but again the data was formed into a single cluster and was not separable.

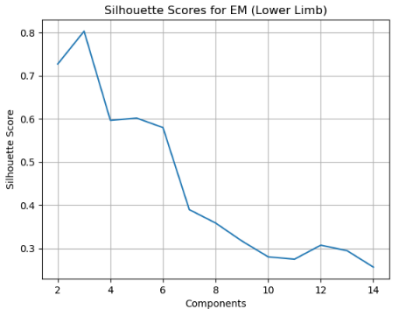


RP + EM:
The BIC for EM was increasing after 2 components so that was selected as the number of components. This method could benefit from selecting additional components, which may be able to capture the original 8 class labels, but further testing is needed.





LDA + EM:
KM and EM both selected 3 components to use and resulted in the exact same t-SNE plot.



For the last two experiments the Abalone Dataset was used.

Experiment 4: Dimensionality Reduction for Neural Network

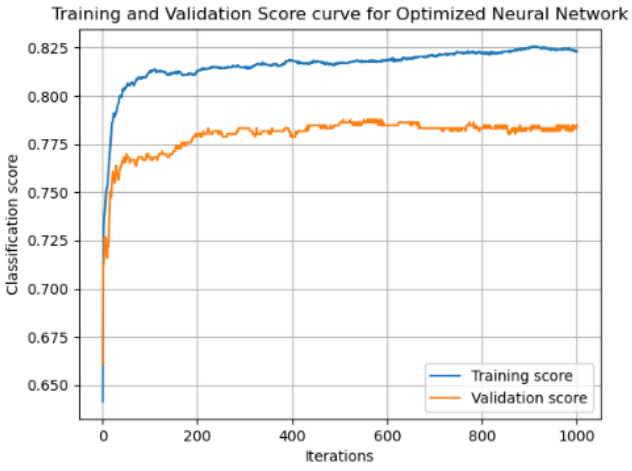
For this experiment, DR techniques were applied and then evaluated with a neural network (NN). Shown below is the NN with no DR applied. With no DR, the model suffers from relatively high bias and variance, with the validation score much lower than the testing score.

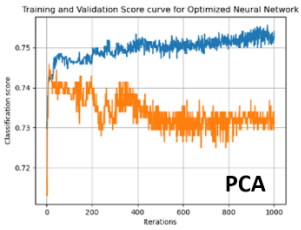
For the 4 images to the right of the training/validation curve for no DR, are the charts with DR. They are small, but the trend can still be observed. For PCA, the bias is increased but the variance decreased. For ICA, the bias is increased but the variance is similar. For RP, the bias is only slightly increased, and the variance stays similar. Finally, for LDA, the bias and variance do not change compared to no DR.

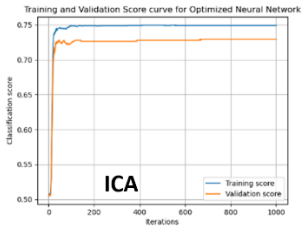
For PCA and ICA, the increased bias is expected as we are removing data given to the NN, which can be seen as a form of undersampling. We are not giving it all the features, so it is expected to learn and generalize from a smaller set of information. RP and LDA show that they can represent the full feature set with only the features selected by the DR algorithms which is encouraging and shows the success of DR.

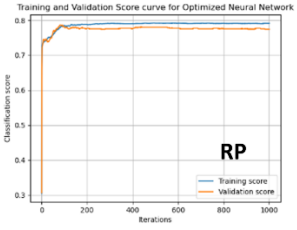
The accuracy may be similar or worse (confusion matrices shown in the appendix for the DR techniques), but that is not the only metric affected by DR. The loss of LDA was very similar to the original data and increased with the increased bias seen from the training/validation curves (LDA: lowest bias, lowest loss --> PCA/ICA: highest bias, highest loss. The time complexity of the NN also was investigated, with the results summarized in the table below. The training times (with the exception of RP) were almost half of the original NN.

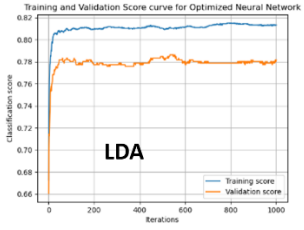
For the NNs for experiments 4 and 5, the same structure was used as Assignment 1 (hidden layers = (16,16)) but learning rate and alpha were tuned.

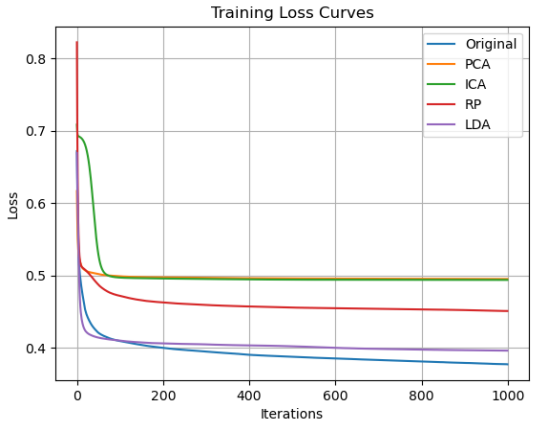






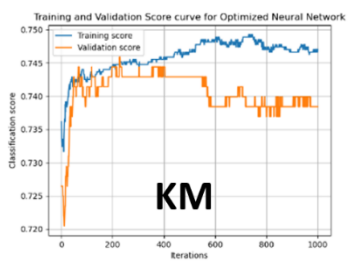


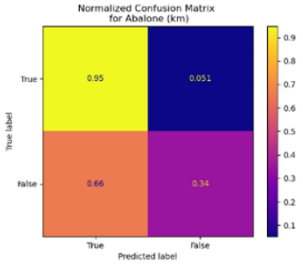


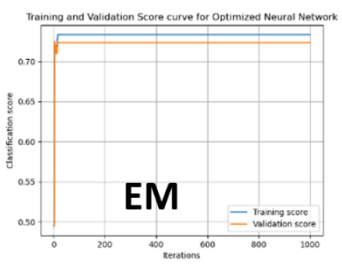


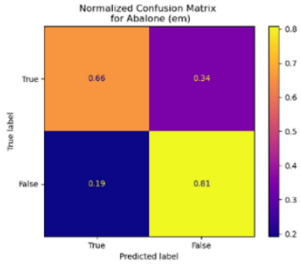
Experiment 5: Clustering for Neural Network

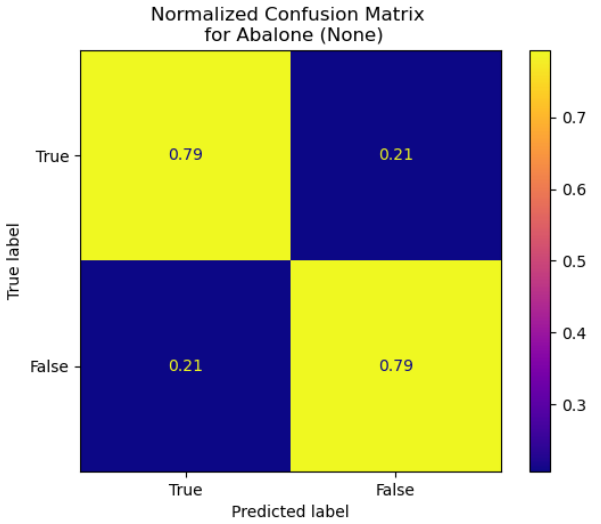
For this experiment, clustering techniques were applied to and the resulting clusters were passed to a NN as input. This experiment can help to see how much information is captured by the clustered created by KM and EM. The training/validation curve for the data with no clustering can be seen at the beginning of experiment 4. With KM, the bias increased, while the variance decreased. The overall accuracy of the neural network also decreased. For EM, there was less bias than for KM, but the bias still increased while the variance stayed decreased. The accuracy of this classifier slightly decreased as well. The increased bias and decreased accuracy, again, was expected as the clusters are only representations of the features space as opposed to the actual features themselves. The loss for both clustering methods was higher than the NN with no clustering. Shown next to the loss curves is a confusion matrix for the NN with no clustering for reference. Below is a table summarizing the training/prediction times and accuracies for the different algorithms. The decreased accuracy is offset by the training times that are ~40% faster to train.

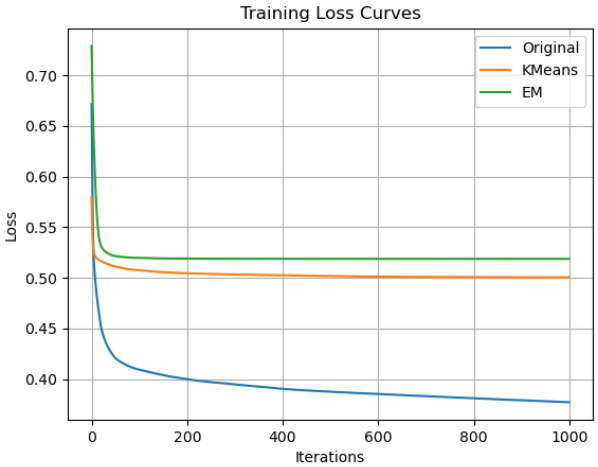












EXP4	Training Time	Prediction Time (Train)	Prediction Time (Test)	Accuracy (Train)	Accuracy (Test)
Original	2.307	0.00299	0	0.810	0.793
PCA	1.349	0.00296	0.00151	0.748	0.746
ICA	1.301	0.00299	0.000997	0.741	0.638
RP	2.656	0.00304	0.000993	0.786	0.757
LDA	1.505	0.00302	0.000997	0.804	0.658

EXP5	Training Time	Prediction Time (Train)	Prediction Time (Test)	Accuracy (Train)	Accuracy (Test)
Original	2.307	0.00299	0	0.810	0.793
Kmeans	1.520	0.00299	0.00103	0.743	0.639
EM	1.411	0.00299	0.000996	0.732	0.734

References:

[1] https://cmdlinetips.com/2019/07/dimensionality-reduction-with-tsne/

[2] Vanschoren, J. (n.d.). Abalone. Retrieved September 18, 2020, from https://www.openml.org/d/183