ML Spring 2020 Project 3; Unsupervised Learning

Adam Pierce; apierce35
adam.pierce@gatech.edu

1 OVERVIEW

This project explores the algorithms to accomplish clustering and dimensionality reduction. In it, I present a set of algorithms in and how they perform when used individually, together, and with a neural network learner. Two clustering algorithms will be presented, along with four dimensionality reduction algorithms.

2 PROBLEMS

2.1 Adult

The 'Adult' dataset remains the same from project 1. This dataset is comprised of demographic data from the US Census, intended to be used to estimate if an individual's income is over 50k annually. It is a binary classification problem which comprises both real and categorical input features.

2.2 Baby

The 'Baby' dataset also remains the same from project 1. This dataset is comprised of 21 input features related to fetal health. This is three-class classification problem intended to identify the health of the fetus. It is also a combination of real and discrete input features.

3 CLUSTERING

3.1 K-Means

For k-means, the most important parameter is k, the number of clusters. To pick an appropriate k for these problems, I performed silhouette and elbow analyses on both datasets, shown in figure 1, below.

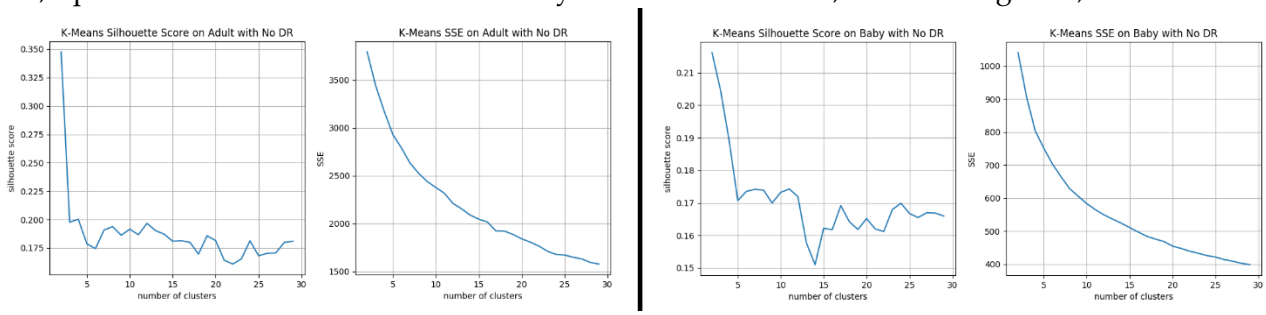

Figure 1: Silhouette and Elbow Analyses for K-Means Clustering

From the above, despite the SSE being high, I viewed the silhouette score to be so much better, two clusters would be best for Adult. For Baby, two was also selected for a similar reason. In neither case did the SSE plot show a distinct elbow. And in both cases there are intuitive hypotheses for two; Adult is a binary classification problem, and Baby could be viewed as binary if the suspect and pathologic groups were combined.

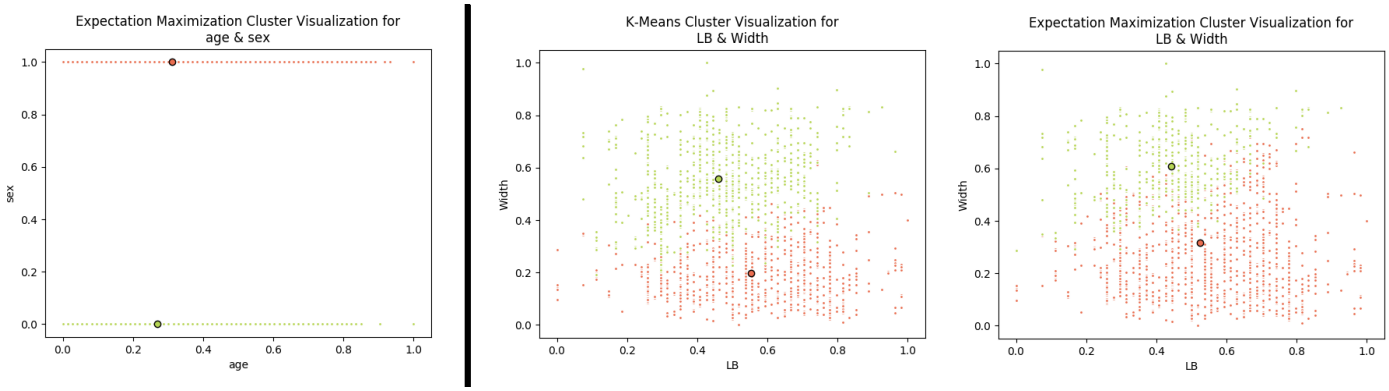

Figure 2: Clusters on Adult (left) and Baby (right)

3.2 Expectation Maximization (EM)

Though the same analysis for above was performed on EM, it is not shown due to the answers ending up the same, two clusters was found to be best for both, and for similar reasons to the above. In lieu of the analysis, figure 2 illustrates the two clusters that were found.

3.3 Additional Analysis

On Adult, the clustering turns out to be identical between K-Means and EM thanks to the separability of sex as a categorical input parameter scaled to the range of [0.0 1.0]. This is an artifact of the math this has nothing to do with the desired income labeling of income classes (>50k?) and more to do with the separability of this categorical feature in the input space.

On Baby, there is some distinction between the two clustering algorithms, as can be seen on the right in figure 2. The two parameters, 'LB' and 'Width' offer some distinction, particularly in the upper right quadrant of the graph, where the 'green' class dominates for K-means, but 'orange' has a bigger presence with EM.

4 DIMENSIONALITY REDUCTION

4.1 Principal Component Analysis (PCA)

Figure 3, below, illustrates the coefficient matrix for PCA for Adult on the left with the variance explained by each component aligned to the right so it can be read left to right, observing the input features that make up each component, and the explained variance of that component, aligned horizontally.

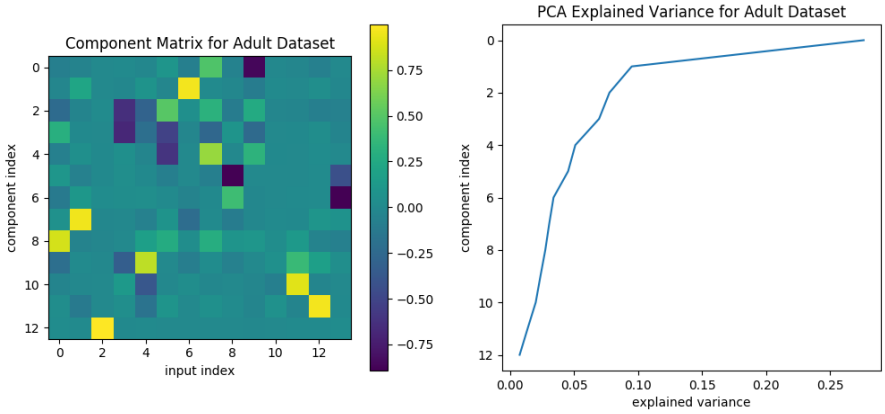

Figure 3: PCA Analyses on Adult

As can be seen, input features 7 and 9 feature prominently in the top 5 primary components, whereas 10-12 do not. From this explained variance chart, Use of 5 components was selected because that is where the explained variance crossed below 0.05.

A similar analysis was performed on Baby where 9 components was selected due to that being the 'elbow' of the explained variance curve. The coefficient matrix for Baby was far more complex due to more input features, as well as them being distributed pretty smoothly about the feature space.

4.2 Independent Component Analysis (ICA)

For ICA, a similar graph to the above figure was produced. Figure 4, below, illustrates the unmixing matrix for 8-component ICA on Baby on the left and the kurtosis of the resulting components on the right. A figure like this was produced for every number of components from 1 to n. This one was selected because it is the first to have a kurtosis of a component near 3, the kurtosis of a normal distribution. For this reason, 7 was selected because it was the highest number of components where they were all non-normal kurtosis.

A value of 7 was also selected for Adult, by the same logic and the same analysis.

Worth noting below, is that again some of the later features in the input space (index 15-19) do not feature strongly in any of the 8 independent components in the figure. This also holds for 7-component ICA.

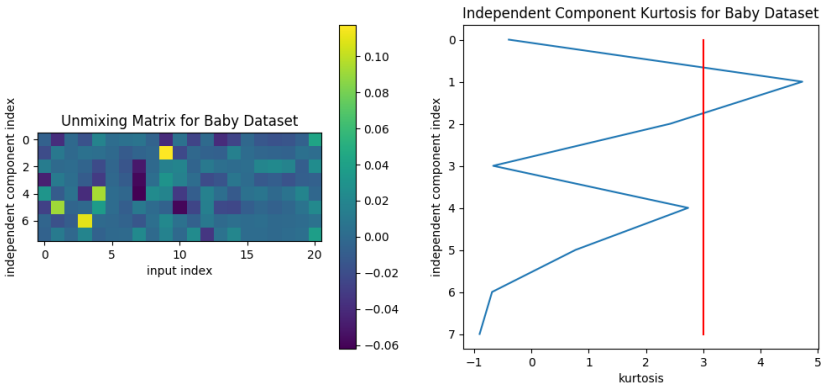

Figure 4: ICA Analyses on Baby

4.3 Randomized Projection (RP)

Randomized projection was tough to identify ideal projections with on both datasets. The reconstruction error for RP on Baby is shown in figure 5, below.

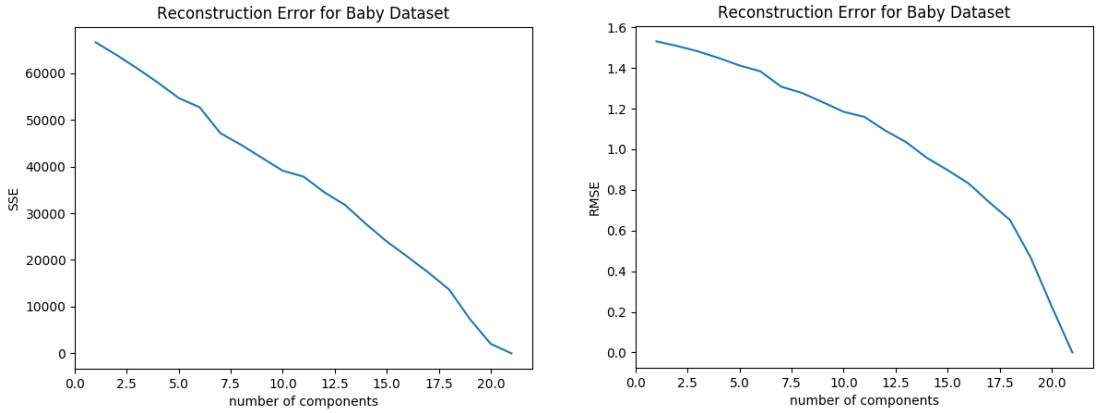

Figure 5: RP Reconstruction Error Analyses on Baby

This graph was produced by averaging reconstruction error over 10 runs of RP for Baby, for each of the illustrated numbers of components. My hope was to be able to identify some sort of elbow in the curve, however in all four cases (SSE/RMSE for Adult and Baby) the reconstruction error plots were either linear as on the left in figure 5, or curved the opposite direction (right), making an 'ideal' value difficult to identify. For that reason, the number of input features was simply halved for both datasets, so for Baby it was 10, and for Adult it was 6.

4.4 Factor Analysis (FA)

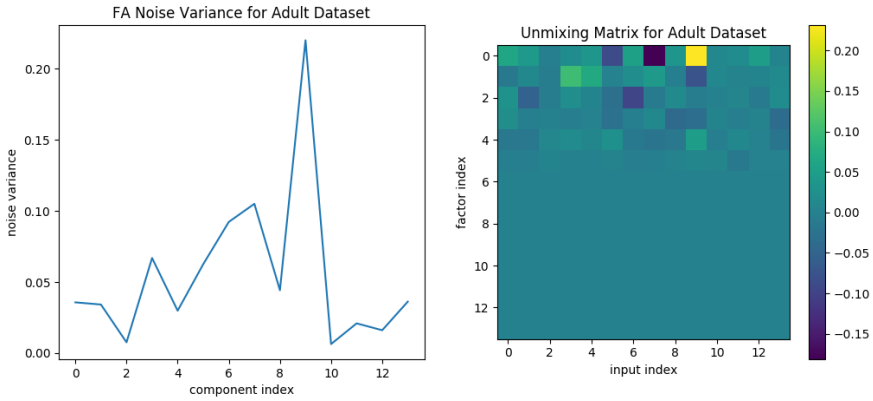

Figure 6: Factor Analysis Analyses on Adult

The final dimensionality reduction algorithm chosen was factor analysis (FA). This is kind of a combination of PCA and ICA, in that it intends to identify 'hidden sources' as in ICA, however instead of using non-normal distribution properties to identify independence, it focuses on the source eigenvectors of a singular value decomposition as is also used in PCA.

For the analyses shown in figure 6, above, the number of components to use in FA was selected as one minus the number of non-zero rows in the unmixing matrix. For Adult, as shown in the figure, that number was 5. For Baby, it was 8, by the same analysis and logic.

5 CLUSTERING ON REDUCED DIMENSION

5.1 Overall Approach

For this section, the number of components identified for each dimensionality reduction method in section 4 was kept, and each algorithm was run on both clustering methods for both datasets, for 16 different analyses. For each one, the same cluster size analysis as in section 3 was performed, followed by a complete pairwise component plot of the new clusters as in figure 2. This produced 900 plots like those in figure 2, when including the ones already discussed above with no dimensionality reduction!

Table 1: Cluster Counts for Dimensionality Reduction on Adult

Cluster Algo	No DR	PCA	ICA	FA	RP
K-Means	2	2	15	3	2
EM	2	2	27	3	2

Table 1, above, shows the final cluster counts for each dimensionality reduction algorithm on Adult. Table 2, below illustrates the same for Baby.

Table 2: Cluster Counts for Dimensionality Reduction on Baby

Cluster Algo	No DR	PCA	ICA	FA	RP
K-Means	2	2	26	8	7
EM	2	2	5	2	5

Worth noting from above, that with the exception of EM on Baby, ICA had by far the largest numbers for number of clusters needed. I believe this to be due to the nature of ICA tending toward non-normal data distributions which, may tend to group the data less. Additionally, PCA's cluster counts remain unchanged from the choice of 2 without dimensionality reduction. This is unsurprising due to the nature of PCA picking the input features of largest variance, which would tend to affect the dimensionality reduction algorithms the most as well.

5.2 PCA on Adult

The first interesting case of this analysis is PCA on Adult. As can be seen from figure 2, both clustering algorithms identified sex as the primary discriminator. Upon inspection of figure 3, input feature 9 is actually sex, so the primary component of PCA is dominated by sex again, as shown in figure 7, below, which illustrates all of the pairwise cluster comparisons between the first (0^{th} feature) and the remaining features of PCA.

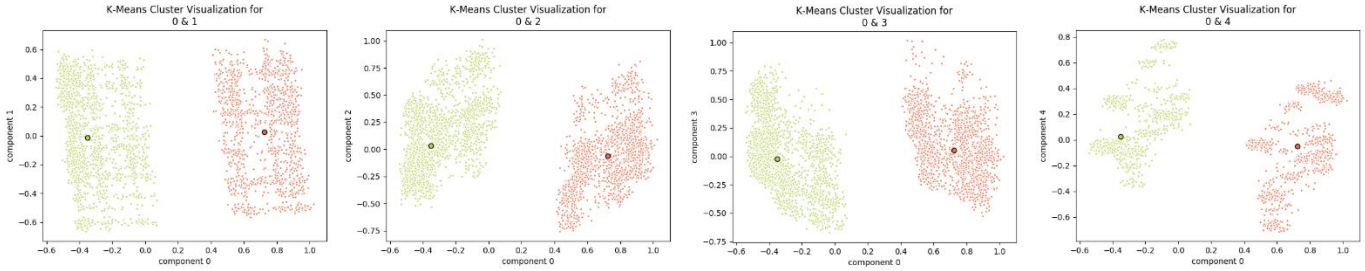

Figure 7: Pairwise Cluster Comparison of Primary Component to all others on Adult after K-Means

Figure 7 is identical for EM, again due to the separability of the feature space as discussed above. The clustering has little to do with the class labels for reasons also already discussed.

5.3 Randomized Projection on Adult

Figure 8, below illustrates one of the random projections of Adult into a 6-feature space, and then clustered by both EM and K-Means. In this projection, there does appear to be 'necking-down' between the two clusters indicating they may truly be independent in some way, perhaps thanks to the now oft-discussed sex input feature.

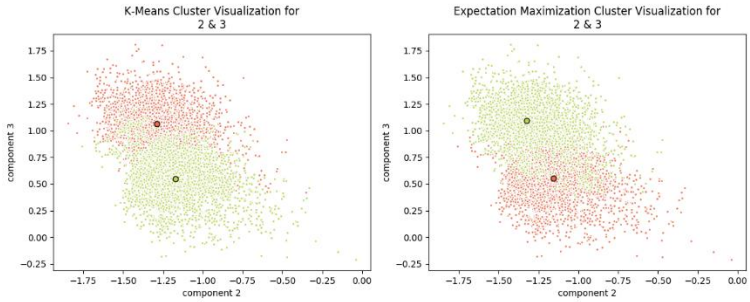

Figure 8: Pairwise Cluster Comparison of components 2 and 3 on Adult after RP

The other features show significantly more overlap, so I believe these are driving the cluster algorithms.

5.4 ICA on Baby with EM

This case is an interesting case-study of the complexity of clustering in high-dimensional spaces over a large number of features. In this example it is made slightly more digestible with ICA on EM only producing 6 independent components and identifying 5 clusters. Figure 9, below, illustrates some of this complexity.

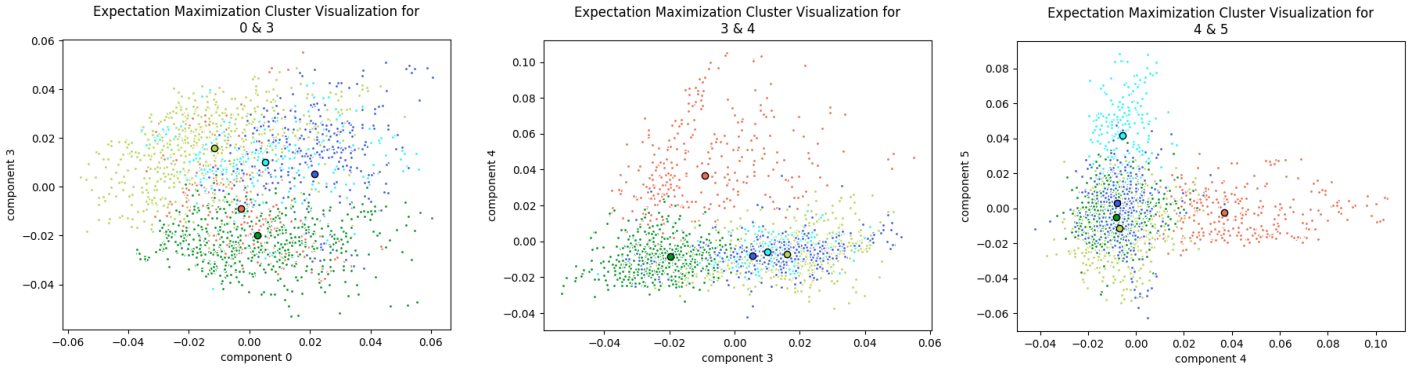

Figure 9: Pairwise Cluster Comparison of ICA reduced Baby data with EM Clustering

Starting with the plot on the left, the light green, blue, and green clusters means are fairly distant from one-another, with aqua and orange overlapping pretty significantly with one or more. However, when considering features 4 and 5 as in the two graphs to the right, one can see how orange and aqua found their own unique spaces to define, and especially in the far right graph, light green, blue, and green overlap almost exactly. Also of note, is that it took both dimensions in the left plot to distinguish light green, blue, and green from one-another, but orange is clearly the unique one defined only by component 4, and aqua is the unique one defined only by component 5. Such are the limitations of visualizing a 6-dimensional problem in 2 dimensions of space.

6 DIMENSIONALITY REDUCTION AND NEURAL NETWORKS

6.1 Prior Analysis

On project 1, the neural network learning was capable of getting 88.7% accuracy on the test set for Baby. On project 2, using randomized hill-climbing and simulated annealing, the network got 88.9% accuracy on the test set.

6.2 Principal Components Analysis (PCA)

For starters, a grid search was performed to identify a near-optimal configuration of the neural network learner. The resulting network had a single layer of 20 relu activated neurons and a learning rate of 0.01. Once tuned, the resulting network was able to get 89.3% accuracy on the test set for Baby.

6.3 Independent Component Analysis (ICA)

The same grid search was performed using ICA components as inputs. This network was also a single layer of 20 neurons with a learning rate of 0.01, but these used tanh activation functions. This network was able to get 84.3% accuracy on the test set.

6.4 Randomized Projection (RP)

A grid search was not performed using RP due to the random nature, it would be very computationally expensive to calculate a near-optimal model for 6-component RP on average. In its place, the ideal model from project 1 is maintained, with the random projections trained against it. The results turned out well, the mean accuracy was 86.6%, with all 10 runs occupying a range of just 85.0% to 88.4%.

6.5 Factor Analysis (FA)

A last grid search was performed using the 8 input features of the FA processed Baby dataset. The tuned network used 20 relu activated neurons in a single layer with a learning rate of 0.01. The resulting network was able to get 88.4% accuracy on the test set.

6.6 Additional Analysis

Overall, the best performing network was the PCA fed network with 89.3% accuracy, beating out not only any other dimensionality reduction algorithm, but also all other neural networks from any of the two prior projects. I believe this is due to the nature of PCA to identify the most important input features and allow the gradient descent of neural network tuning to focus on those.

I was disappointed with the performance of ICA which was the worst performing network out of this group of four. I believe this is thanks to the fact that ICA may be inappropriate to apply on Baby, based on my inspection of the pairwise cluster graphs and other dimensionality reduction analyses, it seems like Baby, far more than adult is a very 'smooth' dataset. This may also be driven by its use of exclusively numerical inputs (though not all are continuous) unlike Adult which uses many categorical inputs. Intuitively this would seem to make it more difficult for ICA to discern independent 'sources' for the data based on non-gaussian distribution.

Lastly FA showed a surprisingly strong performance, especially being similar in nature to ICA in that it intends to identify 'hidden sources' to the data, except based on variance rather than independence. This leads me down the same road as above with ICA except here the smoothness of Baby was not a detriment to the successful application of FA.

7 CLUSTERING AND NEURAL NETWORKS

7.1 K-Means

For this analysis, the labels resulting from the K-Means clustering were added as an additional input feature to the existing features for the neural network, and as above, a grid search was performed to find a near-optimal network configuration. The resulting network model used a single layer of 5 neurons with relu activations, and a learning rate of 0.01.

The resulting network tied with the PCA network with a best-yet test set accuracy if 89.3%.

7.2 Expectation Maximization (EM)

Similar to above, the labels from EM clustering were added to the input features of the neural network and a grid search for the optimal model was performed. This network used 10 neurons in a single layer with tanh activations and a learning rate of 0.01. It got a test-set accuracy of 89.0%.

7.3 Additional Analysis

Overall, it seemed adding clustering results to the input features provided the gradient descent algorithm enough of a 'hint' to provide some marginal improvement to the base learner, to the point of matching the PCA network's performance for K-Means clusters.

Out of curiosity, an additional neural network model was tuned and tested, this one using PCA components as inputs, with the K-Means clustered labels appended to it, however this model got the same 89.3% accuracy.

8 OVERALL ANALYSIS

8.1 Clustering

The two clustering algorithms being evaluated here showed some differences in their solutions, but also that given certain 'insurmountable' features in the input data, can and will actually come up with the same answers.

Expectation Maximization was by far the longer of the two clustering algorithms to run. This is likely due to its more statistical iterative nature, requiring not only more iterations but more operations during each iteration. Given that it is clear to me why I had heard more about K-Means clustering prior to this introduction, it seems to be a pretty good sweet-spot of 'expressiveness' in clustering as well as efficiency in computation time.

8.2 Dimensionality Reduction

Based on the results from the neural network experiment, it is clear why PCA is one of the most popular dimensionality reduction algorithms used in the world today. It offered the best neural network performance of all projects in this class, on top of enabling the use of fewer input features to the network, improving computation time, at only a relatively small pre-processing expense.

The other reduction techniques seem to be more situational in their application. One would likely have to have certain domain knowledge to indicate some 'hidden source' effects in the dataset to see ICA or FA compare or surpass PCA in performance improvement.

Randomized projection theoretically has the possibility to outperform any of the other three approaches discussed here, except that the existence of a better projection is not guaranteed, and the probability of finding it, intuitively, is vanishingly small. However, as above, there may be a dataset that does not lend itself to any of the prior approaches, and the likelihood could be greater. The computational expense of generate-and-test of RP with an ML algorithm could be valuable when the evaluation time of a ML algorithm is critical, and perhaps largely a function of the number of inputs, so reducing the number of features would be valuable.

9 REFERENCES

1. Various scikit-learn tutorial files were used as-is and modified:
   a) https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html
   b) https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html
   c) https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html#sphx-glr-auto-examples-cluster-plot-mini-batch-kmeans-py