Unsupervised Learning and Dimensionality Reduction

Doan Le

1 DATASETS

The same two datasets from assignment one were used in this assignment. Their description is repeated here along with some additional details.

The employee attrition dataset consists of 1,470 examples, each corresponding to an employee in the healthcare industry. The features include education level, position, pay, overtime work, and many others. Some of these features have numerical, continuous values while some are ordinal or categorical data. To encode the data, nominal columns with ordinal meaning were converted into integer values, maintaining the same order. Categorical data with no clear numerical relationship were one-hot encoded into binary columns. The resulting dataset consisted of 49 features and one label. The large number of features creates concern from a machine learning standpoint due to the curse of dimensionaility. This was observed in assignment one, and in this assignment it will be interesting to see how dimensionality reduction may affect performance.

The red wine quality dataset consists of 1,600 examples describing characteristics of red wine and a rating of their quality. The data consist of 11 characteristics with real, continuous values, describing various characteristics of the wine including acidity and alcohol content. The focus of the data is on objective, quantifiable measures. These real-valued attributes can make distance metrics used in k-Means and distributions used in EM more meaningful than the binary measures used throughout the employee attrition dataset.

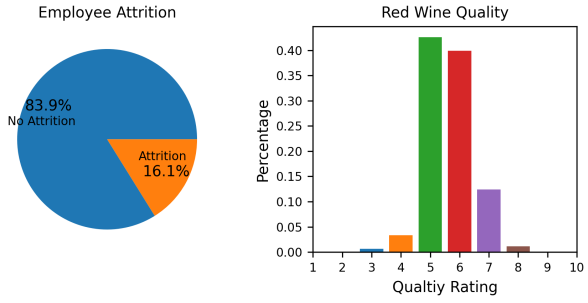

Figure 1 – Label distribution for datasets. The employee attrition dataset is imbalanced towards no attrition. The red wine quality is almost even distributed between scores less than five and scores greater than five.

In the employee attrition dataset, the target label was a binary value representing whether the employee left their organization or not. The class labels were imbalanced with 84% of the employees remaining at their organization and 16% leaving. In the red wine quality dataset, rating is on a scale of 1-10, where ten is the highest quality and one is the lowest. This label was encoded into a binary value (0 for ratings less than or equal to five and one for ratings greater than five) representing good or bad quality. This allows the problem to be solved as a binary classification and conveniently gives an almost balanced dataset. As such, the data is treated as balanced for the rest of the experiment. The class balances of the two datasets are shown in Figure 1. There is no noise in either of the datasets. All samples with the same attributes have the same class label.

2 METHODS

In addition to the encoding mentioned above, all data is preprocessed to standardize features with zero mean and unit variance. When scoring classifier performance, accuracy is used for the wine quality data and f1-macro is used for the employee attrition data.

2.1 Clustering

k-Means clustering and expectation maximization (EM). were applied to raw features and reduced features in both datasets. The number of clusters was chosen by doing a silhouette analysis on different numbers of clusters. The optimal number of clusters was chosen based on a combination of mean silhouette score, the shape of the clusters, and the size of the clusters. This method was employed to catch situations where one cluster with very few points and very high silhouette scores caused the mean silhouette score to be misleadingly high.

The resulting clusters were evaluated by calculating the homogeneity score, taking into account ground-truth labels. Homogeneity score as used rather than completeness because there were often more clusters than labels, so it should be expected that samples with the same ground-truth label can be spread out over multiple clusters. Therefore, it is more desirable for clusters to contain samples with similar ground-truth labels (homogeneity).

2.2 Dimensionality Reduction

The four dimensionality reduction techniques include principal component analysis (PCA), independent component analysis (ICA), random projection (RP), and decision tree (DT) feature importance. For PCA, the number of components was chosen based on the amount of explained variance, keeping only compo-

nents with higher variances. The cutoff point was determined using an elbow method, ignoring components once the explained variance began to approach a plateau. For ICA, the reduction was run separately on different numbers of independent components. The number of components was chosen by an elbow analysis aimed at maximizing mean kurtosis.

For RP, a sparse projection matrix was generated for different numbers of components. The reconstruction error for each configuration was calculated and plotted in order to find the number of configurations that minimized reconstruction error. If there was no clear minimum or elbow, the number of components was chosen arbitrarily as the minimum number of components used in other methods in order to make meaningful comparisons. Using this approach, the number of dimensions is at least reduced enough to hopefully see some changes in clustering and neural network performance.

For decision tree feature importance, a decision tree was fit to the training data and labels with a minimum impurity decrease of 0.01 to keep the tree simple and prioritize relevant features. The feature importance for each attribute was extracted from the fitted tree, and these values were plotted in descending order. The graphs were analyzed qualitatively to find features with the most significant feature importance. The cutoff point was determined qualitatively with an elbow analysis.

The resulting features of each method were evaluated based on their performance using a dummy decision tree classifier. The new sample features were used to fit a decision tree with a specified minimum impurity decrease of 0.01 at each node. This requirement is not a rigorously tuned parameter but rather prevents some overfitting in order to compare scores between each method. Five-fold cross validation is used to generate training and validation scores. The fit time and tree size (number of nodes) is also compared.

2.3 Neural Networks

The employee attrition dataset was chosen for further analysis with neural networks. Four different neural networks were tuned and fitted to the four dimensionality reduction outputs. A grid search and cross-validation plots were used to find optimal hyperparameters for hidden layer size and learning rate. Cross-validation used a stratified five-fold sets and f1-macro score. The remaining parameters were left at sklearn's defaults. Loss curves, wall clock time, training score, and testing score were used to evaluate performance.

Finally, the clustering output on raw features were added to the raw features as additional dimensions. This generated two new input sets, one for k-Means clustering and one for EM. The tuning, fitting, and evaluation process was repeated for these two sets.

3 CLUSTERING ON RAW DATA

3.1 Employee Attrition

Running k-Means clustering on the employee attrition features, it was determined that using nine clusters would produce the most even and defined clusters (Figure 2). The smaller 6^{th} cluster and broader 0^{th} and 1^{st} clusters shows that there is still some variety in the shapes of the clusters formed; however, these issues were less apparent in the 9-cluster configuration than any other.

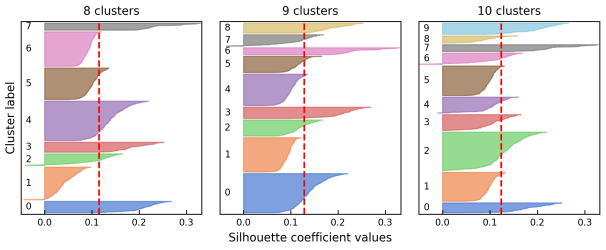

Figure 2 – The silhouette analysis of k-Means clustering on raw employee attrition features shows that there is a peak in silhouette using nine clusters. Although there are some small and concentrated clusters, the overall cluster sizes and silhouettes are more evenly distributed than using one more or one less cluster.

Doing a similar analysis on EM, the silhouette scores for EM were not as high as k-Means. Since EM is a soft clustering method, it follows that clusters can be more spread out, and outliers can lower silhouette scores.

Due to the large amount of binary features in the employee attrition dataset, TSNE was run with a perplexity of five to generate the two features shown in

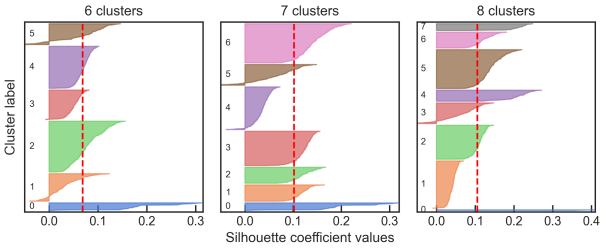

Figure 3 – The silhouette analysis found that seven clusters described by Gaussian's of full covariance best clustered the raw features of the employee attrition dataset. Although the mean silhouette score is higher at eight (left) clusters, the 0^{th} cluster in this case is not impactful.

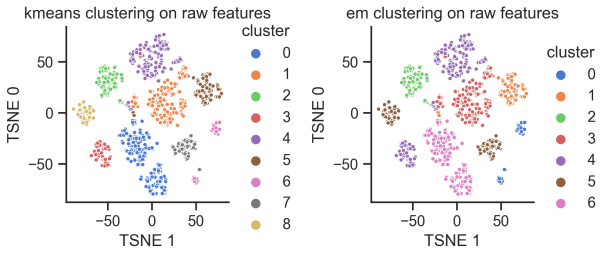

Figure 4 – TSNE plots reduce the 49 features of the employee attrition dataset into two dimensions. Clustering results for k-Means with nine components (left) and EM with seven components (right) are shown.

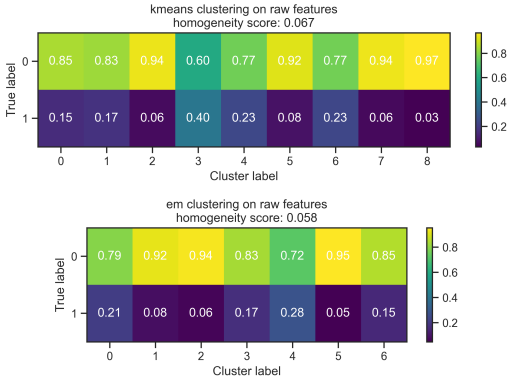

Figure 5 – For clustering on raw employee attrition data, the percentage of each cluster that belongs to a ground-truth label is shown.

Figure 4. Although the number of clusters are different, both k-Means and EM found similar groupings. The 5^{th} cluster of EM combines the 7^{th} and 8^{th} clusters from k-Means even though they do not appear to be close on the TSNE plot, hinting that this visualization, although useful, does not tell the full story. Additionally, the point(s) around $(50, -60)$ are classified differently. EM most likely recognizes it as an outlier where k-Means is assigning it to the center that is closest.

When evaluating the clusters formed by k-Means and EM, the differences were small but k-Means had a slightly better overall homogeneity score. All of the clusters formed consist mostly of samples with false labels (no attrition). Since the data is imbalanced towards no attrition, it appears that while the clusters formed may be well-defined in the unsupervised case, they do not partition the data in a way that is meaningful for the supervised case.

3.2 Red Wine Quality

For red wine quality, the raw features could be clustered in two components for k-Means (Figure 6) and three components in EM (Figure 7). As in the em-

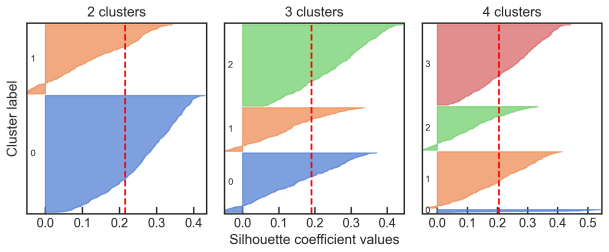

Figure 6 – The silhouette analysis on k-Means clustering of the raw red wine quality features show that limiting the number of clusters to three keeps the clusters themselves evenly distributed and fairly dense.

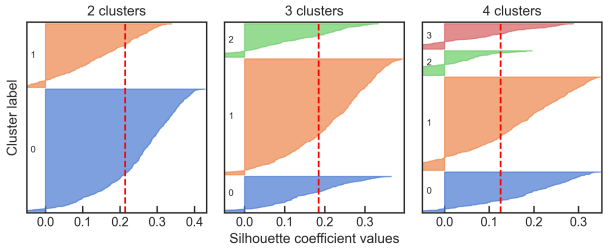

Figure 7 – The silhouette analysis on EM clustering of the raw red wine quality features show that creating more than two clusters introduces more error (negative silhouette scores) and uneven clusters. The ideal configuration was two components of tied variance.

ployee attrition dataset, lower silhouette scores can be expecting using the soft clustering method. Unlike the employee attrition dataset, tied variances produced better silhouettes than full variances. Using full variances, there were more samples with negative silhouette scores. By restricting the covariance matrix, it's likely that EM was able to find clusters with greater separation rather one dense cluster and one sparse cluster overlaid on top of each other.

The clusters themselves have differences. In k-Means, the total sulfur dioxide seems to play a bigger role in defining clusters. In EM, the clusters seem to split more around alcohol content and citric acid. Comparing only citric acid and total sulfur dioxide doesn't show the full picture; however, it seems that with the tied variance approach the shape of the clusters are fairly similar. This could be attributed to using a silhouette analysis which is biased towards these dense, separated clusters.

During cluster evaluation, EM had a higher homogeneity score even though both methods produced very low scores. However, looking at the separation of ground truth labels themselves, k-Means did a better job of finding different clusters for different labels, which could end up being more useful in classification.

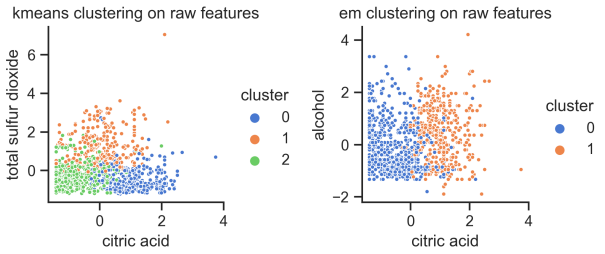

Figure 8 – On the red wine quality raw features, both k-Means and EM seem to distinguish between different citric acid levels. k-Means also makes a distinction between different sulfur dioxide levels.

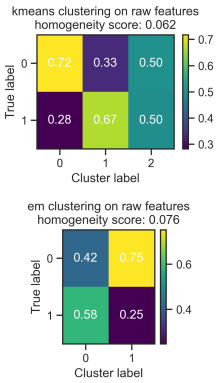

Figure 9 – For clustering on red wine quality data, the percentage of each cluster that belongs to a ground-truth label is shown.

4 DIMENSIONALITY REDUCTION

4.1 Employee Attrition

4.1.1 PCA

Looking at PCA on the employee attrition features, four components were selected based on the elbow analysis for explained variance (Figure 10).

When analyzing the projection matrix for PCA, the features in each principal component seem like they could be somewhat correlated. The 0^{th} component,

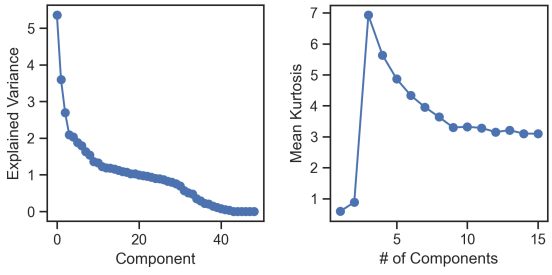

Figure 10 – For PCA on employee attrition (left), plotting explained variance and number of components shows that after four principal components, the explained variance begins to plateau. When performing ICA on the employee attrition dataset (right), the mean kurtosis across each component peaks at three features.

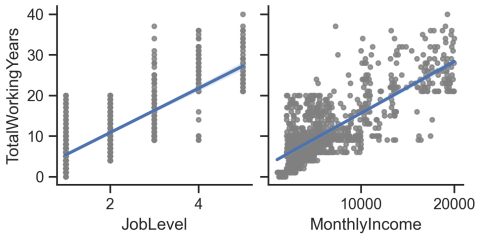

Figure 11 – The strongest features in the 0^{th} principal component appear to have some correlation with each other.

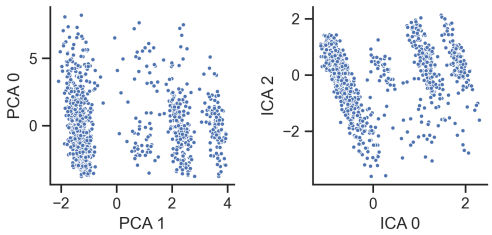

Figure 12 – Components found in PCA and ICA on the employee attrition dataset turn out to have similar shapes.

has a strong positive correlation between job level, monthly income, and total working years. These features are intuitively related, but Figure 11 empirically shows the correlations in the data. This makes sense because if features are tightly related, and one of them has a high variance, the others should as well. Therefore, PCA is useful at compressing these related features into one component. The same logic can be applied to other principal components as well.

4.1.2 ICA

ICA on the employee attrition data is most optimal using three components (Figure 10). In this case, there is a clear maximum without need for an elbow analysis.

When looking at component weights, there are many similarities between the ICA and PCA features. The 0^{th} ICA feature is a close match to the 1^{st} PCA feature with strong positive coefficients in *Department_Human Resources* and *JobRole_Humnan Resources*. Similarly, the 1^{st} ICA component closely matches the 2^{nd} PCA component, and the 2^{nd} ICA component is almost the exact opposite of the 0^{th} PCA component. Figure 12 shows the similarities between a PCA and ICA components. The ICA components appear to be these PCA components rotated and scaled differently.

These similarities are most likely due to the fact that ICA performs a whitening step in preprocessing. This whitening step uses singular value decomposition in a

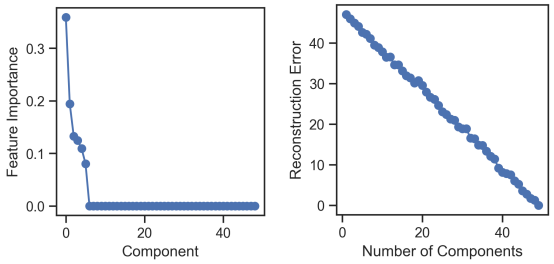

Figure 13 – The feature importance in a decision tree trained on the employee attrition features show that six features are most important (left). In random projection, there is no clear elbow or minimum (right).

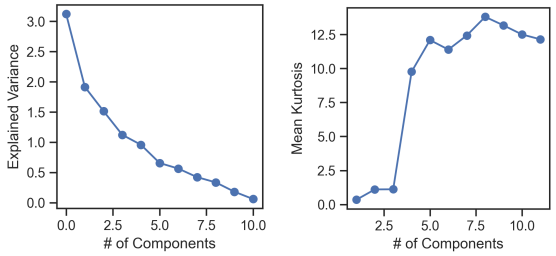

Figure 14 – On the red wine quality dataset, six PCA components are chosen based on a slight elbow on the explained variance curve (left). Eight ICA components are used due to the peak in mean kurtosis shown on the plot (right).

Method	No. Features	Fit Time (ms)	Training Score	Node Count
Raw	49	3.97	0.686	13
PCA	4	1.08	0.626	7
ICA	3	0.89	0.603	6
RP	3	0.88	0.250	2
DT	6	0.92	0.672	11

Table 1 – Comparison of dimensionality reduction methods for employee attrition on a decision tree classifier

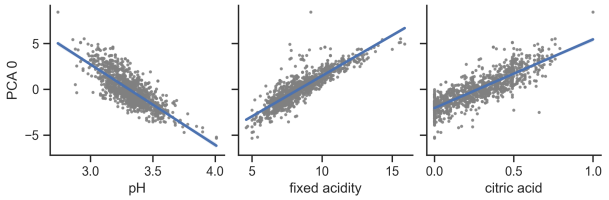

Figure 15 – The 0^{th} PCA component found correlated features in pH, fixed acidity, and citric acid.

similar manner to PCA. This is the mechanism used to reduce the number of dimensions. The whitened features are then treated as the observable features in the blind-source separation problem. The similarities between ICA and PCA components implies that ICA is unable to meaningfully separate the independent components. This can be due to errors in the ICA's underlying assumptions. Either the data cannot actually be explained by hidden independent variables or those variables have a Gaussian distribution.

4.1.3 RANDOM PROJECTION

With random projection, there is no clear peak or elbow in plots of reconstruction error (Figure 13). The reconstruction error is zero when using all features and increases linearly at a constant rate until all components are removed. Arbitrarily, three features were chosen to match ICA.

4.1.4 DECISION TREE

Since the minimum impurity decrease was fixed to avoid a large, overfit tree, it was easy to use the calculated feature importance to determine which features were used. In this case, the six features were *JobLevel, DailyRate, WorkLifeBalance, OverTime, YearsAtCompany,* and *StockOptionLevel*.

4.1.5 COMPARISON

When fitted to a decision tree classifier, the performance differed across the dimensionality reduction outputs (Table 1). The raw features took the longest to run, had the highest training score, and formed the biggest tree. Using PCA and ICA reduces the training score; however, the difference may be acceptable considering the improvements in fit time and tree size. Random projection achieves the same fit time improvement as ICA; however, the training score is unacceptably lower. The decision tree struggled to find good splits on the three random features. The decision tree method saw the best training score and a better fit time than PCA, which uses less dimensions; however, this evaluation method is clearly biased to perform well here.

4.2 Red Wine Quality

4.2.1 PCA

Plotting the explained variance showed a more gradual decrease than the employee attrition dataset (Figure 14). The elbow was identified at the 5^{th} component (6 components total).

Again, PCA was able to find features that were correlated. The 0^{th} PCA component has strong positive coefficients for fixed acidity and citric acid, and it has a strong negative coefficient for pH (Figure 15). Fixed acidity measures the presence of acids that stay in their liquid form at ambient conditions, which includes citric acid. pH also decreases as acidity increases. PCA was able to find these relationships and describe them in one feature.

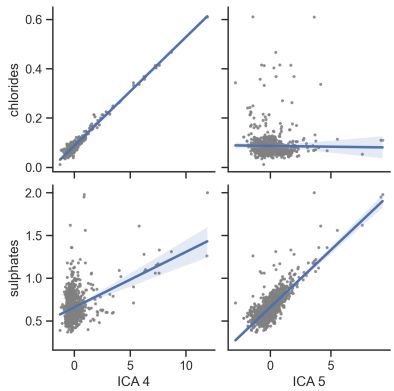

Figure 16 – Independent components found during ICA on red wine quality can be seen affecting different observed features to different extents.

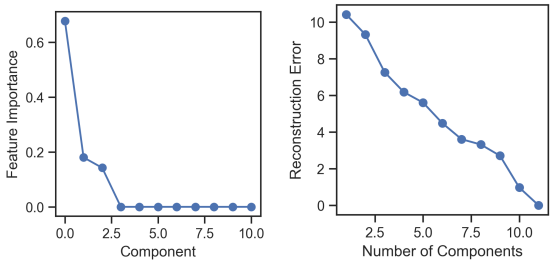

Figure 17 – On the red wine quality dataset, three features are used on a decision tree(left). Random projection does not have a clear minimum or elbow (right).

4.2.2 ICA

When varying the number of components in ICA, a peak in mean kurtosis occurs using 8 components (Figure 14). In this case, the components are not as similar to the ones found in PCA, indicating that the whitened features are a more distributed mix of the independent components. This is also supported by a higher value for mean kurtosis than in the employee attrition dataset.

This mixing can be seen in the 4^{th} and 5^{th} independent components (Figure 16). The chlorides feature seems to be strongly affected by the 4^{th} ICA component but farther away from the 5^{th}. On the other hand, sulphates appear to be affected by both with a stronger signal from the 5^{th}.

4.2.3 RANDOM PROJECTION

The reconstruction error shows no clear minimum or elbows (Figure 17). Therefore, six components were chosen in order to be comparable to ICA and PCA methods.

Method	No. Features	Fit Time (ms)	Training Score	Node Count
Raw	11	2.71	0.717	9
PCA	6	1.96	0.709	7
ICA	7	2.53	0.738	7
RP	6	2.14	0.671	6
DT	3	1.74	0.722	8

Table 2 – Comparison of dimensionality reduction methods for red wine quality on a decision tree classifier

4.2.4 DECISION TREE

Using a decision tree, the important features are the only ones that are used. Figure 17 shows that three features are sufficient. These three features are *alcohol*, *sulphates*, and *volatile acidity*.

4.2.5 COMPARISON

Using the dummy decision tree classifier, Table 2 shows comparsions between dimensionality reduction outputs. ICA interestingly has the highest training score, even higher than using raw features and using a decision tree to generate the features. This indicates that ICA was very effective and its underlying assumptions most likely hold true. The underlying hidden variables that were found not only match the observed variables but also have some effect on the ground truth labels.

PCA showed a reduction in fit time, training score, and node count, which is consistent with its performance on the employee attrition dataset. Random project performs the poorest, but the difference is not as drastic as in the employee attrition example since not as many features are taken away.

5 CLUSTERING ON DIMENSIONALITY REDUCED FEATURES

5.1 Employee Attrition

Combinations of dimensionality reduction and clustering techniques were run on the employee attrition dataset. Homogeneity evaluations of the resulting clusters are shown on Figure 18. The homogeneity of clusters did not improve after dimensionality reduction.

PCA resulted in much lower homogeneity scores and no useful separation of ground truth labels. ICA performs better than PCA but still scores lower than the raw features. This can be explained by the lost of information when reducing the number of components. PCA used three components, and ICA used four components, so it makes sense that ICA performs slightly better. However, considering the fact

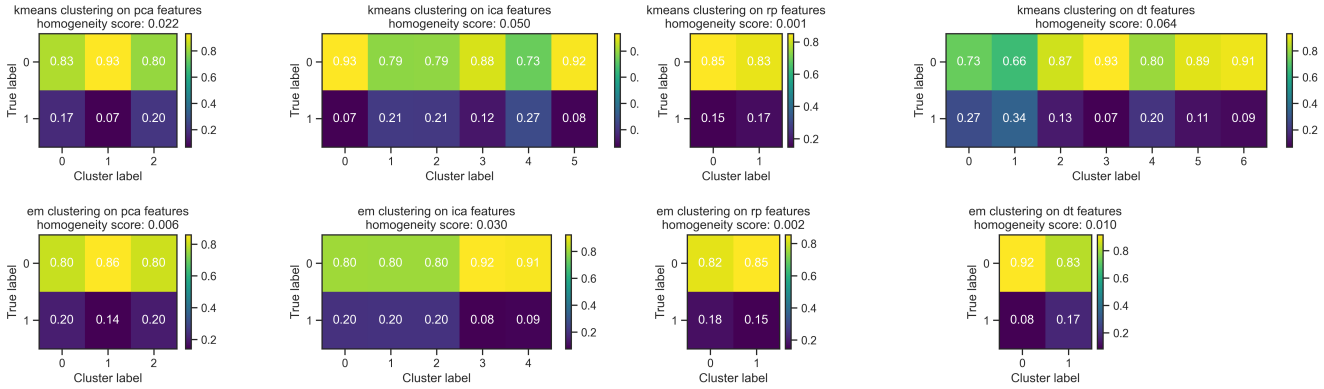

Figure 18 – For clustering on employee attrition data after applying dimensionality reduction techniques, the percentage of each cluster that belongs to a ground-truth label is shown.

the the samples started out with 49 features, the results from ICA indicate that a lot of information was still retained in those four features. Since it was noted earlier that the ICA features were influenced mostly by whitening, it's possible that the PCA component selection process was too aggressive, and using four components would have improved results.

Random projection formed the least homogeneous clusters. In fact, the distribution between truth labels within each cluster are very close to the label distributions over the training set (Figure 1). Since random projection creates random component vectors, it makes sense that clustering done on these vectors finds random samples of the dataset.

Decision tree feature importance performed only slightly worse than the raw features using k-Means clustering. It's surprising that the decision tree did not improve homogeneity as the features used should have been valuable in determining ground truth labels. This can be explained by the low training scores observed when performing reduction and using the dummy decision tree classifier on the employee attrition dataset. It's likely that the decision boundaries formed by decision trees are not a natural fit to the data. Additionally, the ideal number of EM clusters was much lower than when using raw features. This indicates that some of the features unused by the decision tree had high enough variances when doing EM to be significant.

Clusters for the PCA/k-Means case are shown in Figure 19. This clustering particularly highlights the ordering of PCA components. k-Means recognizes features with greater variance, so the 0^{th} and 1^{st} PCA components appear to have clearly definied clusters while the 2^{nd} component does not.

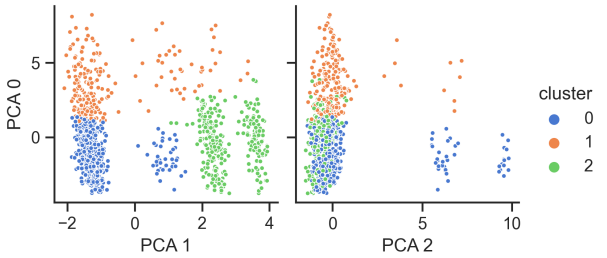

Figure 19 – Running k-Means clustering on PCA components of the employee attrition dataset results in the clusters shown.

5.2 Red Wine Quality

In the red wine quality dataset, many dimensionality reduction techniques improved cluster homogeneity (Figure 20). PCA nearly tripled the homogeneity scores in both k-Means and EM. As mentioned above, PCA was able to combine features that were saying the same thing (i.e. citric acid is a fixed acid, so they do not need to be two separate features). PCA was also better at splitting the truth labels to create clusters that could be useful in classification.

ICA performed better than the raw features using k-Means but not as well as PCA. This could indicate that the independent components found, while more useful than the raw features, were not more useful than the whitened components.

Random projection, as expected, performed poorly compared to the raw features and any other dimensionality reduction technique. Although it did not suffer as much as random projection on the employee attrition dataset, it's clear that information is lost.

Decision tree features performed the best out of all techniques. Not only are the homogeneity scores higher but the clusters separate the truth labels fairly

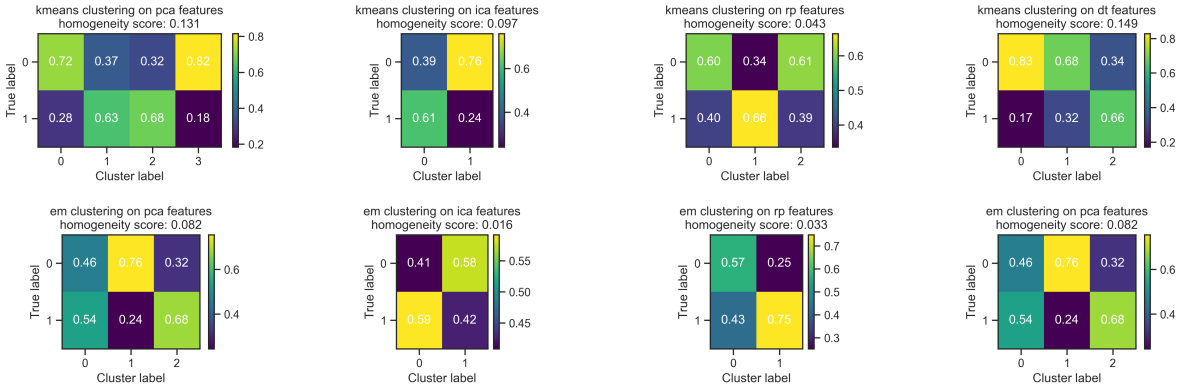

Figure 20 – For clustering on red wine quality data after applying dimensionality reduction techniques, the percentage of each cluster that belongs to a ground-truth label is shown.

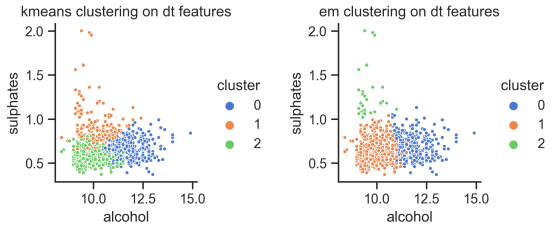

Figure 21 – The clusters formed by applying k-Means and EM on features found in a decision tree are shown.

Input Type	Fit Time	Train Score	Test Score
Raw	140.2	0.791	0.765
PCA	174.4	0.626	0.545
ICA	146.0	0.597	0.546
RP	21.0	0.456	0.456
DT	79.7	0.723	0.635

Table 3 – Comparison of neural network performance with different dimensionality reduction techniques

well. This method has the advantage of using supervised learning and looking at the truth labels. In this problem, decision trees performed better than in the employee attrition problem. Therefore, the expected result occurs and the features presented to the cluster-er are ones that have already been determined to play a role in determining the truth label.

To look closer at these well-performing features, see Figure 21. Dimensionality reduction makes the differences between k-Means and EM more easier to visualize since there are less variables determining distances that are not depicted. In k-Means, all three clusters appear to meet in the middle of the plot where they have similar densities. In EM, the cluster describing high sulphate content is better separated from the others. It takes advantage of a more sparse region around a sulphate content of 1.0 and decides that this region must be on the boundary of different Gaussians.

6 NEURAL NETWORKS

6.1 Dimensionality Reduction

Dimensionality reduction did not improve the accuracy of neural networks on the employee attrition data. PCA and ICA not only saw a reduction in training and testing score but the wall clock time did not improve either. This is most likely because parameter tuning showed that larger networks performed better on these features. While the number of dimensions was reduced, the number of weights to learn increased, and ultimately more iterations were required.

Random projection performed the quickest and the least accurate. Tuning of the random projection network showed that adding perceptrons did not change performance. Therefore, the neural network consists of only one perceptron. Combined with the small number of features, this explains the dramatic decrease in wall clock time observed. However, the scores are significantly lower than when using any other technique.

The decision tree features seem like the most appealing dimensionality reduction option. Not only is the wall clock time cut almost in half but also the scores suffer the least. This technique was designed to choose features that were important in determining the class label, so naturally it performs better on neural networks. While ICA and PCA finds interesting features, there is no guarantee that these will be significant for the classification problem.

The loss curves for each technique are plotted in Figure 22. It's clear that random projection starts out with the most error and is unable to minimize it

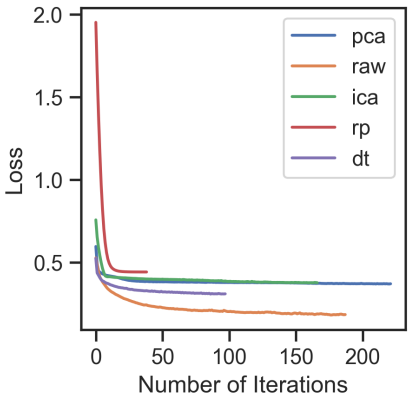

Figure 22 – The loss curves for training neural networks using dimensionality reduction as a preprocessing step are shown.

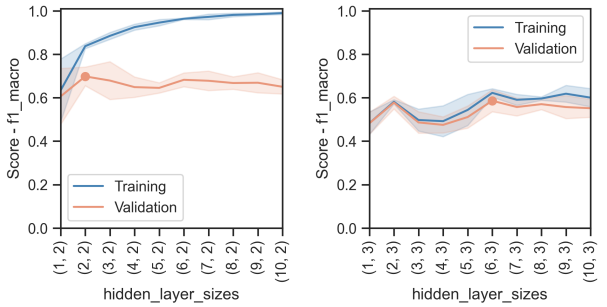

Figure 23 – When tuning the hidden layer size, validation curves for the raw features (top) show a clear trade-off between bias and variance. The PCA features neural network (bottom) does not show signs of overfitting.

Input Type	Fit Time	Train Score	Test Score
None	133.3	0.791	0.765
k-Means	2476.4	0.768	0.784
EM	2309.4	0.767	0.792

Table 4 – Comparison of neural network performance with different dimensionality reduction techniques

enough to create meaningful results. PCA and ICA seem to behave similarly with PCA continuing for more iterations, despite having less features. As observed previously, PCA and ICA had very similar features for this dataset. This can explain the similar behavior up to 150 iterations. It's possible that PCA continues further because the model has trouble converging when missing the additional feature included in ICA.

The decision tree features create a smoother and steeper loss curve. The converged loss is above that of the law features but lower than the other techniques, confirming what was observed in training scores. This approach seems to directly address the curse of dimensionality. By discarding less relevant features, the neural network was able to learn the same number of training examples in fewer iterations.

When hyper-tuning the hidden layer size with reduced features, the model was able to use more perceptrons without risk for overfitting. Figure 23 shows a validation curve used with raw features and one used with PCA features. The raw features show a clear bias-variance tradeoff. At smaller neural networks, validation scores found in cross-validation have a higher variance as the model has more degrees of freedom. As the number of perceptrons increase, the model's variance decreases but the results are heavily biased towards the training data. Using PCA features, the model seems to have high variance and low bias no matter how many perceptrons you add. A moderate amount is selected since there seems to be no improvement after a certain number of perceptrons.

Dimensionality reduction introduces noise that was not there before. By either removing features completely or compressing them, more samples may be considered identical (or similar) even if their ground truth labels are different. While a decision tree helped make a more educated guess on which features are safer to discard, it's clear that without all (or at least most) of the original features, the model does not fit well to the training data, and some pieces of information are missing.

6.2 Clusters as Features

Neural networks for inputs including clusters required slower learning rates and more perceptrons. As a result, more iterations were required and the fit time increased significantly with the addition of eight new features.

The training score seems to be slightly worse while the testing score is slightly better in both cases. The learning curves show that training and validation scores are very close during fitting (Figure 24. This indicates that the chosen parameters result in a model with low bias, and the slightly higher testing scores can be attributed to high variance.

The loss curves clearly show the impact of lower learning rates used in neural networks with clusters as features (Figure 25). The model requires more iterations to converge and the curves are smoother than when using raw features. The final loss in both cases is higher than the original, which makes sense with the slightly lower training scores observed. This indicates that the clusters added more dimensions and complexity without adding much information useful to the classification problem. This can be expected

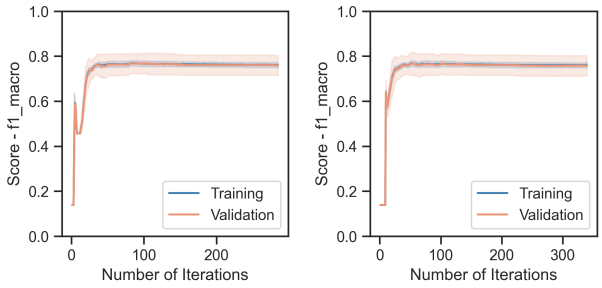

Figure 24 – Learning curves were generated for neural networks using both the raw features and calculated clusters as inputs. k-Means is shown on the right, and EM is shown on the left.

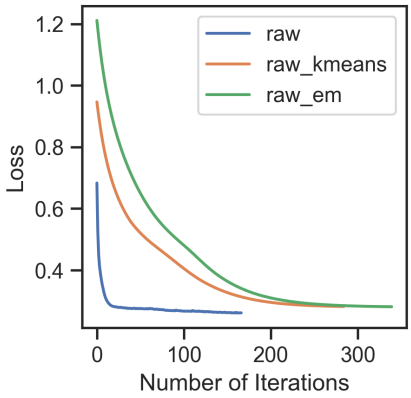

Figure 25 – Loss curves were generated for neural networks using both the raw features and calculated clusters as inputs.

due to the low homogeneity scores of the clusters themselves and the fact that there was no clear relationship to the ground truth labels. The implications of the curse of dimensionality are exemplified in the dramatic increase in fit time and iterations without any gain of information.

6.3 Reflections

A common theme in EM clustering was a preference for using tied variances when using a silhouette analysis. Upon reflection, this method may not have been suitable to use with EM since EM clusters should naturally be softer and of varying sizes. Therefore, in many cases k-Means and EM gave similar results. In order to highlight attributes of soft clustering another evaluation method, such as Bayes Information Criteria (BIC) could've been considered.

Dimensionality reduction with PCA and ICA succeeded in providing interesting insight into the domain where they may have failed in optimizing neural networks. PCA was able to find correlated features that could provide or confirm domain knowledge. ICA found interesting relationships. In cases where ICA finds independent components with a high kurtosis, it can be interesting to the domain to determine what these independent features may be representing in the real world. From a clustering perspective, these methods improve wall clock time and form more comprehensible clusters.

The selection of the number of components used in PCA and ICA could be more rigorous and more applicable to the goal of the problem. The elbow analysis is imprecise and does not shed any light on the classification of the features. Supervised dimensionality reduction, such as using decision tree features, was the most advantageous since it was able to retain features with high information gain. For PCA and ICA to be more effective, they could be performed in a wrapping configuration rather than a filtering configuration. Rather than using kurtosis and explained variance to select the number of components, training and testing scores could be used in order to directly affect neural network performance.