A3- Unsupervised Learning & Dimensionality Reduction

Yumna Zehra Rizvi
Georgia Institute of Technology
yrizvi3@gatech.edu

Abstract- This analysis applies and optimizes two clustering methods, Expectation Maximization and K-Means to the Wine Class Dataset and the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset. It then employs four dimensionality reduction techniques: PCA, ICA, Randomized Projections, and Manifold Learning, to improve the datasets before and after clustering. The study investigates how these methods affect the data structure. For the wine dataset, the research further explores the use of a Neural Network classifier with both the clustered and dimension-reduced data. The classifier's performance is evaluated and compared to understand the benefits of combining dimensionality reduction and clustering with Neural Network classification providing insight into the differences as well as potential synergies between the techniques.

1. Introduction

1.1 Datasets and Expectations

The Wine Class Dataset, with 178 entries and 13 features across three classes, is subjected to EM and K-Means clustering to uncover its complex structure. EM is expected to leverage the wine data's intricacies, aligning clusters with underlying Gaussian distributions. K-Means seeks to create distinct, feature-based clusters, but might struggle with the data's multivariate nature and uneven variances. EM's probabilistic approach may better capture the unique distributions of the classes, especially when they overlap or vary in size.

The WDBC Dataset, for binary classification, contains 569 entries with 30 features. Clustering and unsupervised learning are used to discern the most influential features and observe trends in tumor characteristics. EM clustering and K-Means are expected to provide different insights: K-Means is projected to group data based on feature similarity which may reflect tumor classifications, but it may falter with complex cluster shapes or sizes. EM clustering, using Gaussian mixtures, should offer a nuanced, probabilistic clustering that recognizes subtle overlaps and distinctions in data. PCA is anticipated to condense the datasets while preserving variance, aiding in the visualization of critical features and possibly boosting the efficacy of learning algorithms. RP is expected to maintain instance distances and be computationally efficient, albeit less interpretable than PCA.

ICA is anticipated to uncover independent underlying factors in the datasets. UMAP aims to display geometric clustering in a reduced dimension, capturing complex dataset structures and potentially offering insightful visualizations of wine varieties or breast cancer subtypes. It is expected to outperform PCA and ICA in demonstrating relationships and structural flexibility, and to be more noise-resistant, but it may be challenging to interpret and prone to tuning sensitivity and overfitting. The wine dataset is likely easier to analyze visually than the more feature-rich and class-imbalanced WDBC dataset.

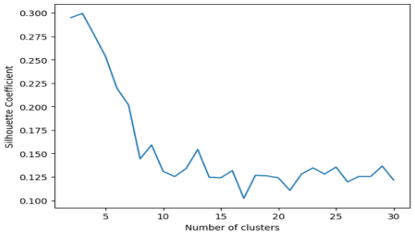
1.2 Approach

The preprocessing routine standardizes and scales both datasets to ensure uniformity, while also addressing any missing values. We employ stratified sampling for dividing the data into training and test sets, reserving the test set exclusively as a holdout set untouched by the unsupervised learning algorithms. Dimensionality reduction techniques are fine-tuned using GridSearchCV to optimize their performance. A baseline implementation of each algorithm is considered in the comparisons.

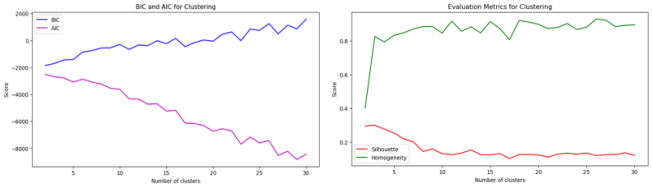
2. EM-Clustering and K-Means Clustering

2.1.1 EM Clustering on Wine Data:

The evaluation of various cluster numbers using GMM involved analyzing silhouette and homogeneity scores, along with the average log-likelihood, to gauge clustering effectiveness. The Euclidean distance was selected for its straightforwardness and suitability for the ratio-scaled features of the datasets.

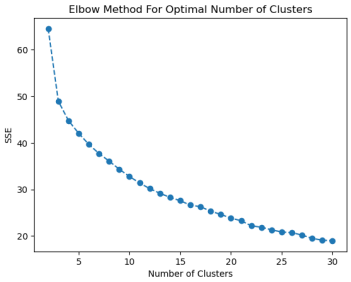


As the number of clusters rises, silhouette coefficients decline, indicating that cluster separation does not significantly enhance and may result in over-segmentation with less distinct clusters. Notably, the plot lacks values for 2 or 3 clusters, which are typically relevant for distinguishing wine types. The trend suggests limited benefits to cluster quality with more clusters. While the BIC does not favor increasing clusters due to minimal gains in model fit, the AIC suggests more clusters improve the model, possibly due to its lesser penalty for extra parameters, raising concerns about overfitting.

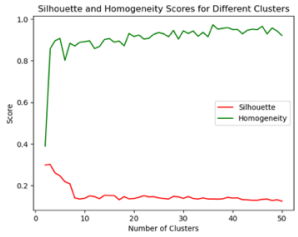


An ideal cluster arrangement for the wine dataset would balance the BIC and AIC metrics, prevent overfitting as indicated by stable silhouette coefficients, and maintain high homogeneity. Accuracy jumps from 1.7% with 3 clusters to 40.4% with 5, highlighting that the data contains discernible subgroups that the EM algorithm can capture with a larger cluster count. The GMM's distinct centroids across 13 dimensions suggest varied wine characteristics. Thus, a 5-cluster model is likely the best fit to represent the wine data's inherent variation without overfitting, according to EM clustering results.

2.1.2 K-Means Clustering on Wine Data:



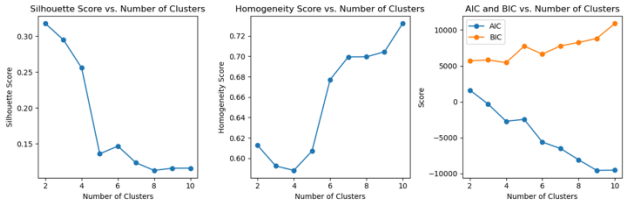
The Elbow Method doesn't show a distinct elbow, but larger decreases in SSE at lower cluster counts suggest that fewer clusters may better capture major data structures. K-Means clustering with 3 clusters shows an accuracy of about 0.624 and an AMI Score of 0.851, indicating strong correspondence with true labels and that three clusters likely reflect significant data patterns. Increasing to five clusters lowers accuracy to 0.079 and AMI to 0.751, implying potential overfitting or less meaningful data segmentation. Although AMI stays high with more clusters, the better performance with three clusters suggests it's a more natural division of the data. Despite expectations, the Elbow Method's ambiguity requires considering SSE and external validation scores, such as AMI, for optimal clustering. Comparatively, the EM algorithm doesn't clearly favor 2-3 clusters, perhaps due to its flexibility in capturing data complexity.



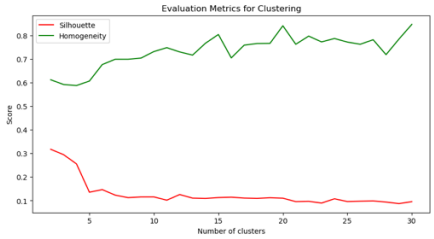
On the other hand, it indicates that accuracy is not a good metric to use for these techniques.

2.2.1 EM Clustering on WDBC Data

The silhouette score, measuring the quality of the clusters, started to decline noticeably after two clusters, suggesting that the separation between clusters is maximized at two clusters. Despite this, the silhouette scores overall are relatively low, indicating there might be significant overlap between clusters or that the clusters are not well-defined.



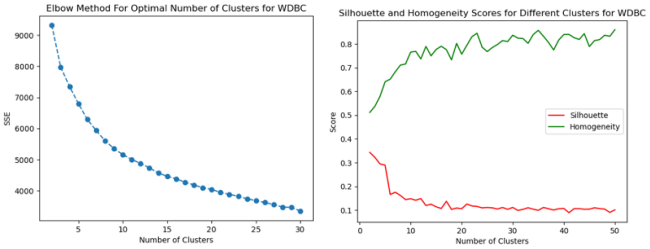
Considering these results, it can be observed that the silhouette coefficient curve peaks for 2 clusters and then decreases to stabilize, remaining low overall. Homogeneity improves as we increase the number of clusters, reaching its peak around 30 clusters. A simpler model with fewer clusters would be ideal for 2 per the silhouette score. However, for an aim of better homogeneity a tradeoff can be considered for more clusters.



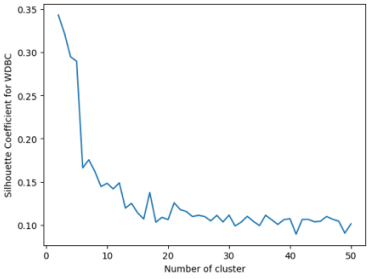
Considering the binary nature of the WDBC dataset (benign and malignant), one would anticipate a two-cluster solution to align closely with the true labels; however, low accuracy suggests the two clusters derived from the GMM may be capturing variance unrelated to the benign/malignant classification. Notably, only a seven-cluster solution yielded non-zero accuracy, hinting at some correlation with the known labels, but this alignment isn't consistent across other cluster numbers, necessitating a deeper examination of the data structure. This discrepancy emphasizes that unsupervised clustering, such as GMM, might not always reflect expected structures, particularly regarding classification accuracy.

2.2.2 K-Means Clustering on WDBC Data

In the plot there is not clear distinct elbow point (although around 10 clusters would be a reasonable guess), suggesting the data isn't portioned distinctly.



The silhouette and homogeneity scores, on the other hand indicate better defined clusters. The homogeneity seems to gradually increase, while the silhouette score tends to decrease. It's interesting to note that even as the silhouette score decreases, the homogeneity score keeps increasing. This means that clusters become more homogenous but might be less distinct from each other- more clusters capture the distribution of the data more accurately in terms of each cluster containing only members of a single class. When taking a closer look at the silhouette coefficient the trend is brought more into focus- the silhouette score is highest for lower number of clusters (around 2 clusters) it decreases as the number of clusters increase.



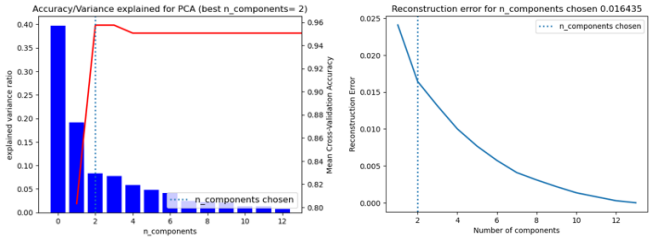
The optimal cluster count for the WDBC dataset is two, reflecting its binary nature, but a low silhouette score suggests that more clusters might be fitting noise or variations within classes rather than revealing distinct subgroups. With more clusters, there's improved homogeneity, where clusters contain data points predominantly from a single class. K-Means on the wine dataset attains better accuracy with fewer clusters, suggesting clear groupings, while on the WDBC dataset, accuracy declines as the number of clusters grows, implying a more complex and potentially overlapping data structure.

3. Dimensionality Reduction

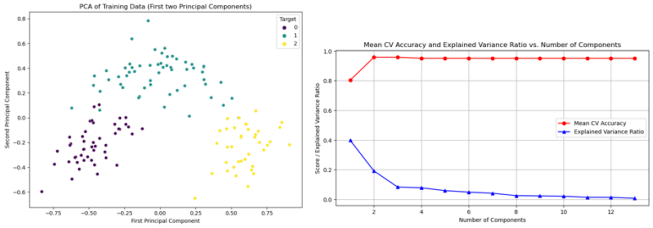
3.1.1 Wine Data PCA

The first two principal components account for 58.88% of the variance, with about 80-85% explained by five or six components, and around 95% by nine or ten components. Cross-validation accuracy peaks around 93-94% and indicates high accuracy with even two components, despite diminishing returns after 90% variance explained, showing that two components largely maintain the necessary data structure for classification.

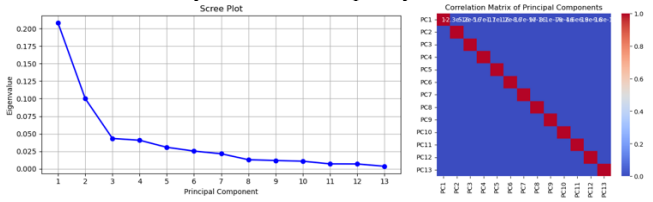
As seen by the Reconstruction Error vs. Number of Components graphs- as expected the error decreases as more components are included but the reconstruction error is very low at the point where two components are chosen, which again supports the idea that two dimensions capture the essential information.



The scatter plot shows well-separated classes when projected down to two dimensions- this indicates good class discriminability. There is a clear elbow point in the scree plot at the second component- further justifying the choice of two components for reduction.



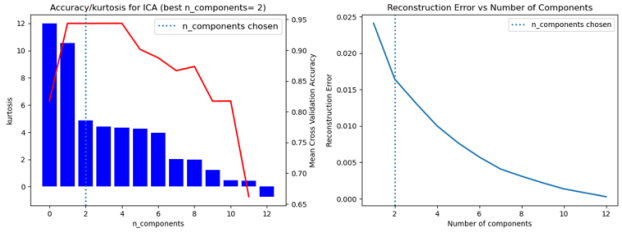
The cumulative explained variance described 10 components are needed to explain at least 95% of the variance but given the high classification accuracy with just two components, there is a trade-off between accuracy and model simplicity.



A decision tree classifier is used to evaluate the results and the cross-validation scores are consistently high, with a mean accuracy of about 91.60%, meaning the features are well-represented even after dimensionality reduction. The mean CV accuracy highlight this as well. Furthermore, the correlation matrix produced displays that the principal components are effectively uncorrelated – affirming the success of PCA deriving orthogonal features.

3.1.2 Wine Data ICA

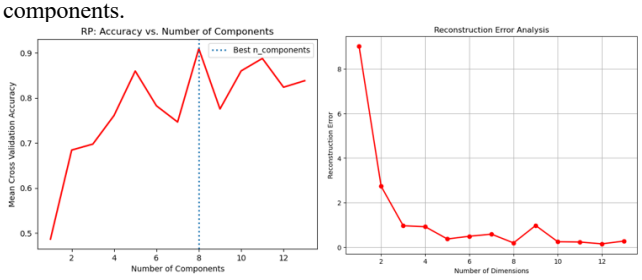
ICA analysis of the wine dataset with two components results in a low reconstruction error of 0.016435, indicating a successful dimensionality reduction while preserving important features. The classification accuracy is very high at 97.22%, confirming that the two components maintain crucial discriminative information. The kurtosis notably drops after the first two components, aligning with the optimal component choice.



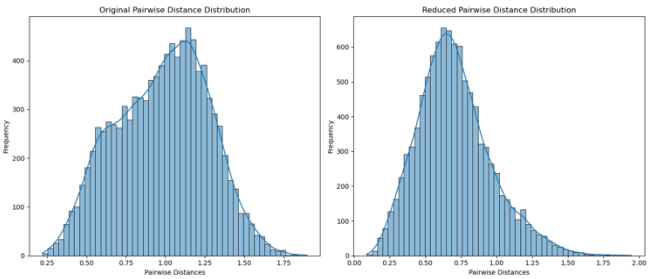
Although the accuracy is consistent across different component numbers, the first two show significant kurtosis. Additionally, reconstruction error falls sharply when increasing components from two, but stabilizes with additional components, reflecting typical diminishing benefits with more components in dimensionality reduction.

3.1.3 Wine Data RP

Mean cross validation accuracy for RP peaks at 8 components achieving the best classification accuracy. This is further confirmed with visualizing the reconstruction error which is minimized at 8 components.

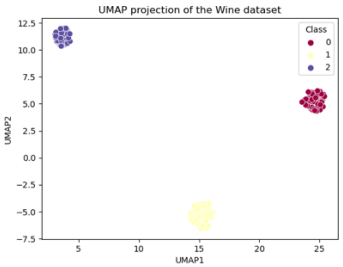


A further increase in dimensions does not significantly improve reconstruction past 8 components. Moreover, the histograms are visualized to compare the distribution of pairwise distances between samples in the original dataset and after dimension reduction.



The distribution looks approximately normal- it appears that the reduced dimensionality has preserved the original distribution as desirable in RP. The result of a reconstruction error at approximately 0.198 with 8 components is good considering RP is a dimensionality reduction technique that aims to reduce the number of random variables under consideration. This is a good balance between preserving information (low reconstruction error) and simplifying the dataset (fewer dimensions).

3.1.4 Wine Data Manifold Learning

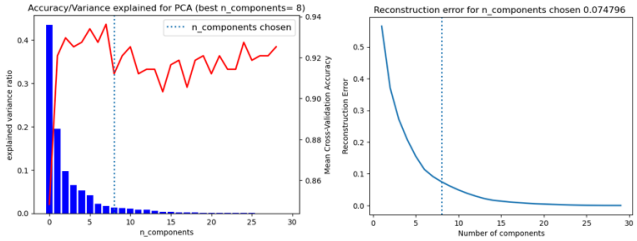


Using UMAP for manifold learning, the wine dataset's dimensionality was effectively reduced to two dimensions, which facilitated the visualization and interpretation of data clusters. Optimal UMAP parameters, determined via GridSearchCV, were 30 neighbors and a minimum distance of 0.1. Post-reduction, distinct clusters emerged, indicating a clear distinction of wine classes in the reduced feature space. The Decision Tree classifier applied thereafter yielded a high cross-validation score of 0.96, showcasing strong performance. UMAP proved to be the most straightforward and effective algorithm for classifying the wine

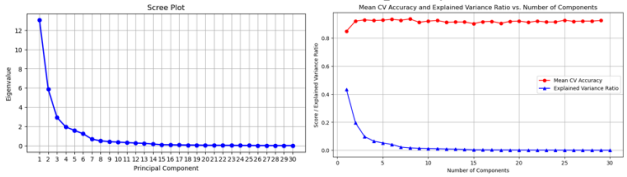
data. For additional insights, it will be interesting to explore feature engineering to see if creating or selecting specific feature subsets could improve classification further. In conclusion for dimension reduction on wine class- UMAP performed the best for wine data – separating and visualizing the results in exactly 3 classes. The ICA analysis was the most different – indicating 8 components as optimal, whereas the other techniques indicated 2 components as optimal.

3.2.1 WDBC PCA

The PCA explained variance ratio is a typical PCA where initial components capture the bulk of the variance – this is a trend that was seen in wine data as well. The accuracy is consistent after 8 components- suggesting additional components do not contribute to improvement in prediction.

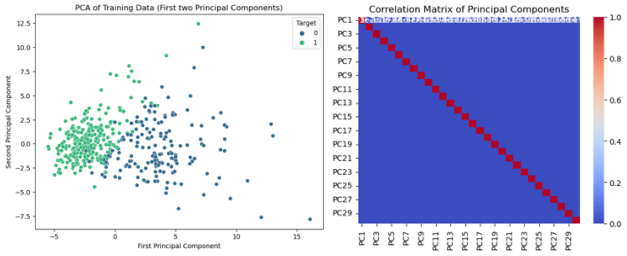


The reconstruction error decreases sharply after the first few components reinforcing the choice of 8 components as a tradeoff between error reduction and model complexity.



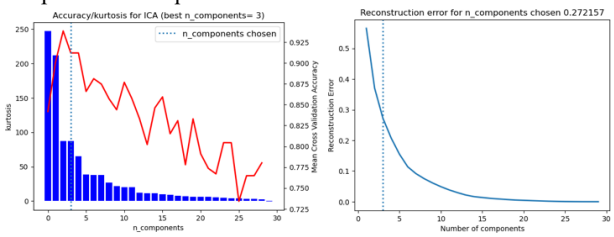
The Scree plot for PCA on the WDBC dataset shows the explained variance for each component in descending order. The first few components capture significant amount of variance which tapers off as it moved to the higher numbered components. When looking at cumulative explained variance, the first two principal components alone explain over 63% of the variance in the dataset. By the 10th component, there is 95% of the variance in the dataset and can be reduced to this without losing much information.

The scatter plot shows that the first 2 principal components effectively separate the training data into groups with some overlap. A decision tree classifier achieved 90.99% accuracy, and a heatmap indicated low feature correlation. It is concluded that using 10 components is suitable for capturing 95% variance in the data, while reducing to 8 components may offer a simpler model with a slight decrease in accuracy.

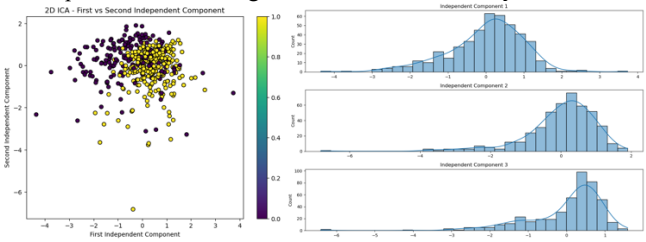


3.2.2 WDBC ICA

ICA on WDBC data highlights 3 as the optimum number of components with a low reconstruction error of 0.272. When visualizing the plots, the first two components account for most of the variance- this has been typical trend so far. The 3rd and 4th components are equal.



The 2D plot visualizes the data points from the WDBC data in terms of the first two ICs – the two classes (malignant and benign) exhibit a degree of separation from the first two ICs -implying that they capture some of the underlying structure or differences between the tumors in the dataset. However, there's also a significant overlap between the two classes in the center of the plot – the ICs capture some discriminatory information and don't perfectly segregate – more features are needed or different techniques to achieve a higher classification accuracy.

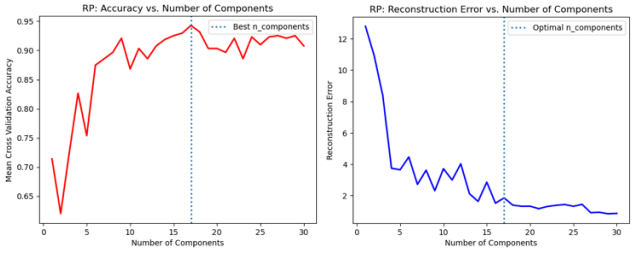


Moreover, both classes spread across the plot, but the benign samples are clustered more towards the right along the first independent component indicating that the first IC captures some variance or feature that defines benign samples. There are also a few outlier datapoints these could be unique cases or noise. Additionally, the distributions of the first three components show an almost normal distribution for the first component. The second and third are skewed left –which could mean that while most of the cells have a characteristic that is relatively high, there are some cells with particularly low values of this characteristic.

ICA on WDBC data reaches a high accuracy at 92.11% and overall the representation provides useful visuals to understand the complex relationships.

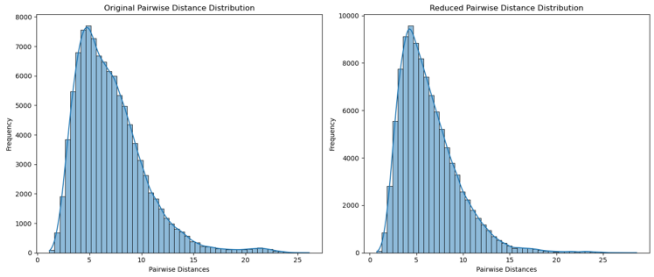
3.2.3 WDBC RP

The Reconstruction error with the chosen number of components which is 17 is about 1.854 which is relatively from other analysis based on the number of components selected- a high reconstruction error is tolerable if the reduced dataset significantly improves efficiency or if the main patterns in the data are preserved.



However, when looking at the pairwise distribution – the similarity in the shapes of the distributions suggests that the dimensionality reduction has somewhat preserved the relative distances between points. However, the shift and spread in the reduced distribution

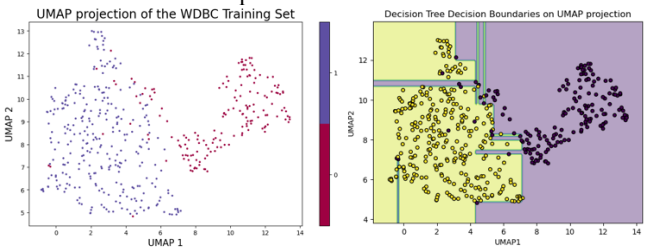
show that some distortion has occurred. This could affect the results in classification since the relationships between points are not the same as in the original space.



In conclusion, RP has not performed as expected on the WDBC data in comparison to other dimensionality reductions.

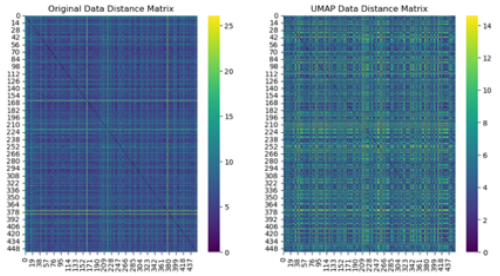
3.2.4 WDBC Manifold Learning

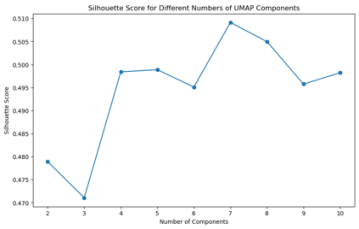
UMAP projection distinctly separates the datapoints in a reduced space as seen by the scatter plots and decision boundaries in the decision tree classifier. The yellow and purple regions indicate a decision area with a complex structure.



The heatmap compares the distance matrices before and after UMAP- the structure after UMAP is more fragmented indicating UMAP has spread out the clusters to expose manifolds structure. The peak silhouette score occurs at 6 components, suggesting that the data may have a natural structure in 6 dimensions that UMAP is able to capture well. Overall, the dataset has a meaningful structure allowing for a separation of classes in reduced space. The distance matrix preserves local structures and spreads out clusters to make the global structure more apparent using UMAP. Finally, the silhouette score indicates that higher dimensions might capture more inherent data.

Overall, the dimensionality reduction techniques, including Manifold Learning, applied to the WDBC dataset, offer insightful visualizations, indicating distinct groupings within the binary classification task. While these methods provide valuable reduction for the training set, the results must be interpreted with caution. The WDBC dataset, though larger than the wine dataset, still presents a limited scope for assessment due to its binary nature, as opposed to the multiclass characteristics of the latter.



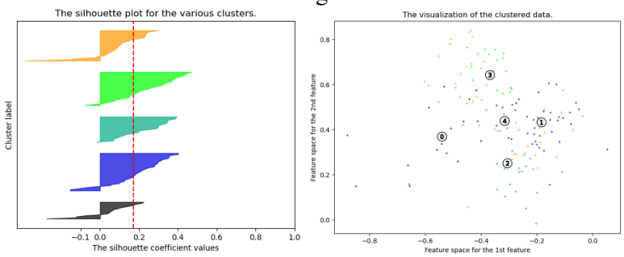


The original high-dimensional nature of these datasets means that some information is lost during the projection process. Despite this, UMAP emerges as the most interpretable technique for both datasets. This underscores the necessity of balancing between the benefits of reduced dimensionality and the potential impacts on model accuracy, as demonstrated by the acceptable accuracy trade-off when using PCA. Future investigations should extend these methods to larger, more complex datasets to fully evaluate their robustness and scalability.

4. Clustering on Dimensionality Reduction

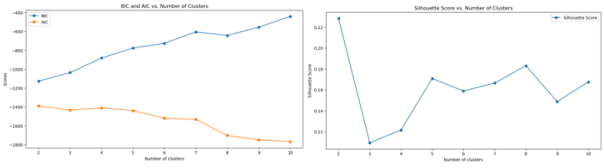
Since the RP dimensionality reduction for both datasets were the most different results from other techniques, RP has been chosen to be highlighted.

4.1.1 Wine RP and EM Clustering

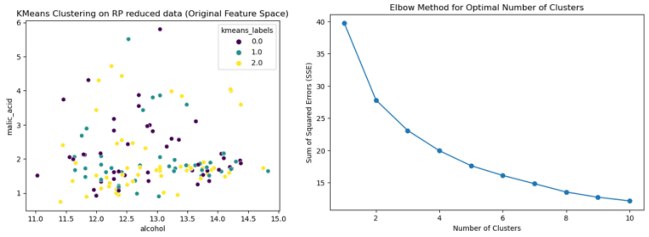


Optimal clusters analyzed are 5.
Silhouette Score on training data: 0.17
Homogeneity Score on training data: 0.60
Adjusted Rand Index on training data: 0.41
Normalized Mutual Information on training data: 0.49
Calinski-Harabasz Index: 32.74
Davies-Bouldin Index: 1.62

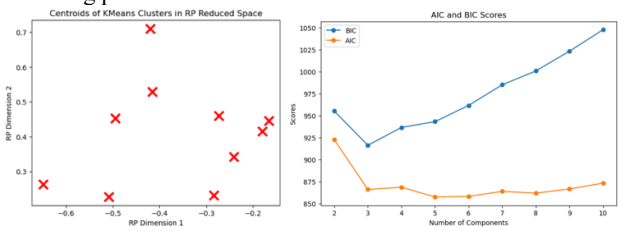
The wine dataset analysis, employing RP and EM with an optimal five clusters, reveals varying performance metrics indicating moderate success. A silhouette score of 0.17 points to some structural overlap but also suggests distinct groupings. The clusters display good purity with a homogeneity score of 0.60. An ARI of 0.41 and NMI of 0.49 both suggest moderate agreement with the true data labels. The clusters appear well-defined and compact as per a Calinski-Harabasz Index of 32.74, with a Davies-Bouldin Index of 1.62 denoting average separation. Two-cluster accuracy is reasonable at 0.82. These measures collectively imply that while the five-cluster model discerns significant patterns in the data, cluster definition could be clearer.



4.1.2 Wine RP and K-Means Clustering



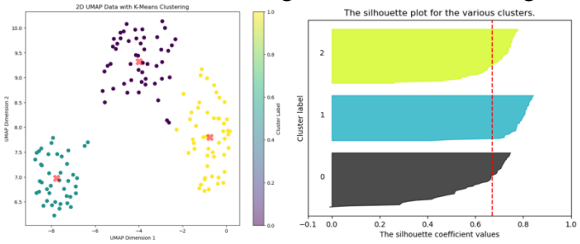
The elbow curve for the wine data suggests an optimal clustering around 3, with a scatter plot showing moderate overlap between clusters, affirmed by a silhouette score of 0.22, indicating that the clusters are not highly distinct. The Calinski-Harabasz score of 50.13 and a Davies-Bouldin index of 1.44 suggest relatively good cluster separation and compactness. Homogeneity, completeness, and V-measure scores of approximately 0.21-0.22, alongside an AMI score of 0.20, point to moderate success in clustering, with clusters mainly containing members of the same class but not exclusively. The SSE of 23.08 shows data points are close to centroids, and balanced cluster sizes with a silhouette-recommended 2 clusters hint at satisfactory but improvable clustering performance.



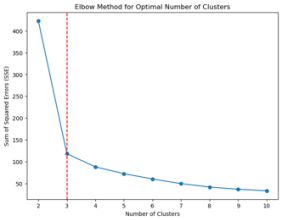
The centroids in the reduced feature space from RP are moderately distanced, indicating some level of cluster separation, yet without stark distinctions. K-Means post-RP identifies structural patterns, although clusters lack crisp definition. The Calinski-Harabasz Index points to decent separation and cluster density, suggesting room for refinement, potentially by varying cluster counts or tweaking methods and preprocessing. A silhouette score of 0.67 for three clusters denotes relatively coherent clusters, with points being more proximal within clusters than between them.

The AIC indicates 5 components as optimal with a score of 857.65, while the BIC suggests 3 components, aligning with the highest silhouette score, indicating a trade-off between model complexity and fit. High Homogeneity (0.84) and Completeness (0.83) Scores denote consistent and class-pure clusters. A low Davies-Bouldin Score (0.45) signifies close-knit within-cluster points and distinct between-cluster separation. The silhouette histogram peaks between 0.6 to 0.8, showing well-clustered data points. Despite AIC and BIC's differing component recommendations, the clustering is robust, with BIC pointing to 3 as optimal, favoring simplicity and fit.

4.1.4 Wine Manifold Learning and K-Means Clustering



Manifold Learning with UMAP yields promising results with with three distinct clusters becoming apparent from the UMAP projection plot.

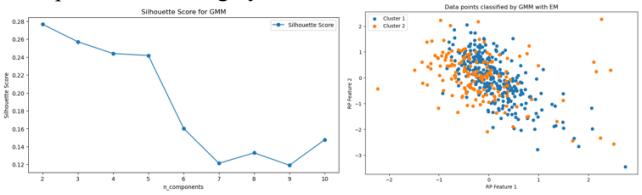


The silhouette plot further confirms the appropriateness of the cluster number, showing a relatively high silhouette score (0.67) with a clear separation of silhouette values among clusters, suggesting a good fit for the data. This is complemented by the elbow method plot, which indicates a noticeable elbow at 3 clusters, aligning with the silhouette analysis and other metrics. The Calinski-Harabasz Index is quite high (735.20), implying well-defined clusters, and the Davies-Bouldin Index is low (0.45), indicating minimal within-cluster scatter. The ARI (0.85) and Mutual Information (0.91) are both close to 1, suggesting a high degree of similarity between the K-Means clustering and the true labels. These measures collectively suggest that the K-Means algorithm, after dimensionality reduction with UMAP, effectively captures the underlying structure of the wine dataset, resulting in clusters that are both well-separated and meaningful.

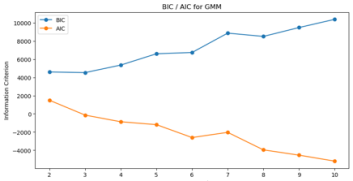
4.2.1 WDBC RP and EM Clustering

The silhouette score suggests that 2 clusters are the most appropriate for the data, with a score of approximately 0.277. The score decreases as more clusters are considered.

The scatter plot shows the data points classified by GMM suggesting that while two clusters provide the best silhouette score, the separation is not highly distinct.

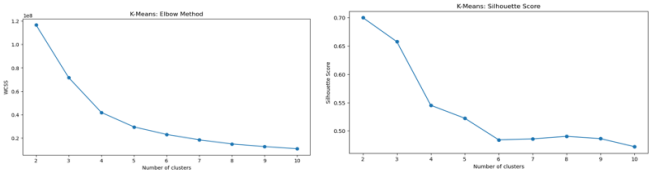


In this case, the BIC indicates that 3 components are optimal, showing a preference for a less complex model.



Conversely, the AIC suggests that 10 components are optimal, which implies a model with more complexity and possibly overfitting given the high number of components. This discrepancy highlights the need to consider multiple metrics and possibly domain knowledge to determine the most appropriate number of clusters or components for the dataset.

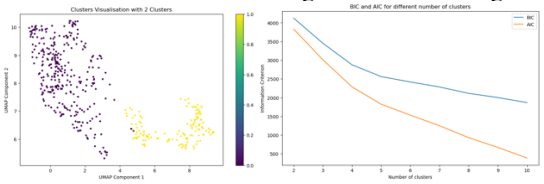
4.2.2 WDBC RP and K-Means Clustering



The elbow method plot indicates a sharp decline in within-cluster sum of squares (WCSS) and begins to level off at 2-3 clusters as the optimal count. This is corroborated by the silhouette score (0.697), which is highest for two clusters, denoting a robust separation and homogeneity within clusters. The silhouette plot further confirms this, with silhouette coefficient values being well above the average for both clusters, validating the clustering configuration.

The 2D visualization for dimensionality reduction reveals two clearly distinct groups, reinforcing the appropriateness of two clusters. Finally, the clustering quality is quantitatively supported by a high Calinski-Harabasz Index (1297.03) and a low Davies-Bouldin Index (0.502), which suggest that the clusters are dense and well-separated. Together, these metrics and visualizations strongly support the conclusion that two clusters provide the best structure for this dataset.

4.2.3 WDBC Manifold Learning and EM Clustering

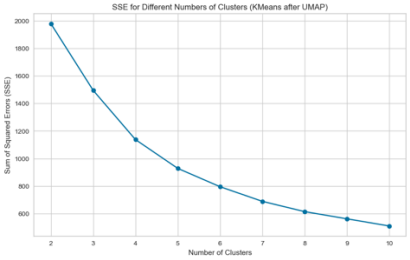


Davies-Bouldin Score: 0.5614383958874708
Calinski-Harabasz Index: 1065.1626316427958
Adjusted Rand Index: 0.7585189144933704
Adjusted Mutual Information: 0.6995453254107402
Optimal number of clusters according to silhouette score: 2

The UMAP visualization indicates two distinct clusters, possibly corresponding to benign and malignant samples. However, the silhouette scores for 2 to 10 clusters remain consistently at 0.3235, suggesting no significant benefit from increasing the number of clusters in terms of cohesion and separation. In contrast, the BIC and AIC both minimize at 10 clusters, implying a more complex model may better. The Davies-Bouldin score and Calinski-Harabasz Index indicate well-separated and compact clusters, while the Adjusted Rand Index and Adjusted Mutual Information suggest a good fit between the cluster assignments and the true labels. Despire the suggestion of 10 clusters- it is more relevant to the WDBC data for 2 clusters.

4.2.4 WDBC Manifold Learning and K-Means Clustering

From the analysis of the silhouette score, the highest score was achieved by two clusters which suggests that the data points were most densely packed into two distinct groups. This high silhouette score of 0.608 for two clusters indicates a strong structure, implying that the two clusters are well-separated and cohesive. However, the silhouette scores decreased as the number of clusters increased beyond two, suggesting that additional clusters do not capture well-separated groupings within the data. The Sum of Squared Errors (SSE) graph reinforces this finding, as the elbow method shows a sharp decrease in SSE as the number of clusters goes from 2 to 3, with diminishing returns thereafter.



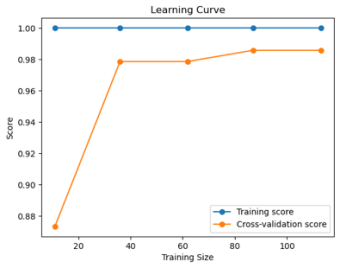
The SSE of 1979.33 for two clusters may seem high, but it is the lowest within the range tested, which corroborates with the silhouette score indicating that two clusters are the most appropriate for this dataset.

5. Neural Network

The choice for the wine dataset for this analysis stems from its high accuracy and performance metrics of precision, recall and f-1 score from A1. The baseline NN model on wine data achieved a perfect score on training data- whereas the final implementation on the test set achieved 0.833 accuracy. This result is interesting as it may have a potential indication of overfitting on the training set. Wine also has less class imbalance to make better comparisons.

5.1 Neural Network - Wine Dataset after Dimensionality Reduction

PCA Algorithm is chosen to be highlighted and for comparison to Manifold Learning. As seen in section 3.2.1- 8 components were ideal to capture enough variance for a tradeoff between error reduction and model complexity.



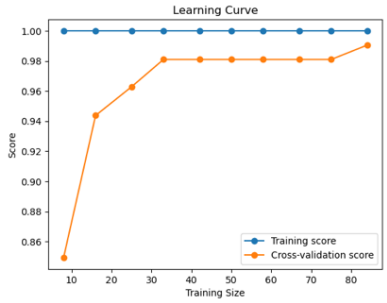
Accuracy	0.9722
Recall	0.9629
Precision	0.9791
F1 Score:	0.9696

The PCA-optimized NN model, with just 8 features, achieved a significant accuracy of 97.22%, surpassing the baseline. This indicates PCA's efficacy in extracting crucial features, which enhanced the model's test generalization. The metrics confirm that PCA streamlined the dataset effectively, boosting the performance. While manifold learning could yield further insights, PCA's balance between complexity reduction and performance is clear with just 8 features. The success of PCA in this context highlights its value for feature reduction and mitigating overfitting.

5.2 Neural Network - Wine after Manifold Learning

The model's performance on the test starts perfectly with a score of 1. As more data is added, it remains consistent. This suggests that the model is very capable of learning from the training data, almost to the point of overfitting, especially with smaller data sizes. The cross-validation score plateaus after around 40 training samples, maintaining a high score just below 0.95.

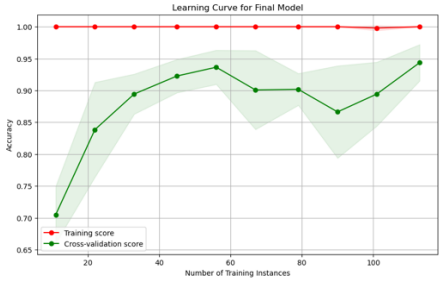
Accuracy	1.0
Recall	1.0
Precision	1.0
F1 Score:	1.0



This result could mean leakage of information, a model that is not challenging enough for Manifold learning. Regardless, the overall metrics indicate a perfect performance – this is validated when looking back at section 3.1.4 where UMAP was able to cluster perfectly.

5.3 Neural Network – Wine after EM clustering

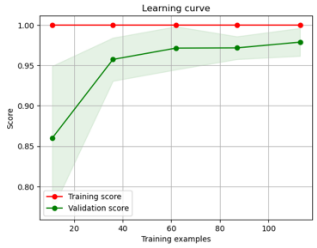
Although the results are not perfect like in Manifold Learning,



Accuracy	0.944
Recall	0.950
Precision	0.950
F1 Score:	0.950

clustering with a NN on the wine dataset has proven to be highly effective, with the optimal number of clusters set to five. The learning curve reflects a stable and high training score, indicating a good fit to the data, and the cross-validation score suggests that the model generalizes well to new data. The wall clock time was also very short at 0.00 seconds.

5.4 Neural Network – Wine after K-Means clustering



With number of clusters set to 3, K-Means achieves balanced classification capability however EM's performance is significantly higher. The choice of 5 clusters with EM might indicate better partitioning, defying expectations – perhaps capturing more nuances of the data with higher choice of clusters.

Accuracy	0.8333
Recall	0.8315
Precision	0.8635
F1 Score:	0.8391

Conclusion

The PCA based NN model showed significant jump in accuracy which defies expectations set in the introduction in comparison to the baseline model from A1. PCA performed better overall than RP in this analysis across all experiments, yet Manifold Learning outperformed all algorithms over all- this was also unexpected. Overall, EM had shorter wall clock time, Manifold Learning had the longest, which is something that was anticipated. Although unsupervised learning does not utilize the labels associated with classification, clustering can provide a preliminary grouping. Furthermore, unsupervised learning techniques like PCA can be reduce the data into fewer dimensions for better visualization and interpretability, offering insights into data distributions. The techniques can be employed to detect anomalies or outliers in the dataset, particularly useful for the WDBC dataset where there are many features overlapping each other with outliers indicating new cases. If a data point does not fit well into any cluster, it could be considered an outlier, which might represent rare characteristics. Additionally, the parameters obtained through clustering like EM provide even further insights into distribution as clustering exposes additional underlying features. Therefore, although unsupervised learning techniques do not replace the accuracy of supervised learning techniques for the datasets explored in this analysis, nor do they have one perfect or universal way of being evaluated for diagnosis – they still offer valuable insights and augment the analysis process.

REFERENCES

[1] OpenAI. "ChatGPT: Personal dialogue on optimizing code for machine learning algorithms and checking for grammer in content." Accessed on: November 2023. [Online]. Available: https://openai.com/chatgpt