Unsupervised Learning, Clustering, and Dimensionality Reduction

Nikhil Solanki

March 2026

1 Introduction

This study explores the effects of clustering and dimensionality reduction techniques on data structure and model classification performance. By using the Adult Income dataset (binary classification) and the Wine dataset (multiclass classification), I investigate how unsupervised learning methods can reveal underlying patterns in the data. By implementing these techniques, I hope to see how these transformations influence overall model performance.

The experiment consisted of various smaller experiments, where clustering and dimensionality reduction algorithms were applied in varying capacities. I began by applying clustering algorithms, such as K-Means and Expectation Maximization (with Gaussian Mixture Models). These algorithms were applied to raw datasets to analyze their inherent structure. I then applied dimensionality reduction techniques (like Principal Component Analysis – PCA, Independent Component Analysis – ICA, and Random Projection – RP) to each dataset and evaluated how these methods compressed and transformed the data.

Building on these representations of our data, I then reapplied the clustering algorithms in these reduced-dimensional spaces to assess whether dimensionality reduction improved cluster quality and separability. Lastly, I examined the downstream effects of these previous experiments by training our previously defined neural networks on both reduced representations and cluster-derived feature augmentations. This enabled me to evaluate the impact of unsupervised learning methods and whether they enhance predictive performance when integrated into a supervised learning pipeline.

2 Exploratory Data Analysis and Hypotheses

This report uses the same preprocessing pipeline, dataset splits, and backbone architecture established in the previous supervised and optimization learning reports to ensure fair comparisons. All model selection and hyperparameter tuning was performed using the training and validation sets only, with the held-out test set was used a single time at the end of each experiment to report final performance. The same 70-30 train-validation split was used for the PyTorch MLP, due to the additional complexity of implementing cross-validation in PyTorch. The same fixed hidden layer sizes of $(64, 32, 16, 8)$ were used for the neural network. A fixed random seed of 66 was used across all experiments to ensure reproducibility and consistent data splits. EDA is shortened for brevity.

Throughout the experiments, two datasets are used: the Adult Income (Census Income) dataset and Wine Quality dataset (red + white). The Adult dataset is a binary classification problem, consisting of income levels for people $\leq 50k$ or $> 50k$. The dataset contains over 45,000 samples with a mix of categorical and numerical features, exhibiting moderate class imbalance ($\sim$ 3:1). The target is *income*, a binary and imbalanced variable. This motivated the use of preprocessing techniques like one-hot encoding and feature scaling (standardization) to ensure compatibility for all models.

The Wine dataset represents a multiclass classification problem, with the target of classifying wines based on their *quality* ratings. It consists of roughly 6500 entries, of solely numerical data describing aspects of the wine. The class distribution is favored towards mid-quality ratings, where models must distinguish subtle differences between neighboring quality levels.

EDA from prior work indicated that both datasets contained meaningful structure but differ in complexity. This motivated the use of clustering and dimensionality reduction methods to better understand the underlying structure of the data and improve downstream classification performance. The following hypotheses aim to evaluate how different unsupervised learning techniques reveal and utilize structure in the data, while potentially improving downstream supervised learning performance:

Hypothesis 1: Dimensionality reduction techniques will improve clustering performance by removing noise and potentially redundant features. I presume that PCA can produce the most stable and effective clustering results, whereas ICA and RP may introduce more variability due to their assumptions and randomness. **Hypothesis 2**: Training my neural networks on dimensionality-reduced data will reduce the input complexity and improve generalization to an extent. Aggressive reductions will likely degrade classification performance by removing important information. **Hypothesis 3**: Incorporating cluster-derived features into neural networks will not improve classification performance, since clustering significantly lessens data complexity which my deep neural network thrives on.

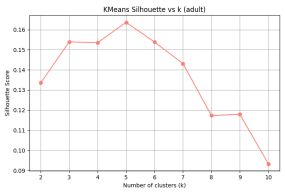
3 Clustering on Raw Data

3.1 Adult Dataset

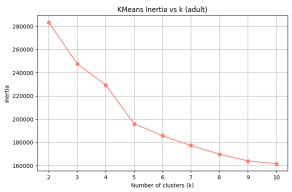
3.1.1 K-Means

K-Means clustering was applied to each dataset with the number of clusters k, selected based on the silhouette score. The silhouette score is a metric for how separated and cohesive clusters can be, ranging from $[-1, +1]$. A high score indicates points are close to their cluster and far from others, with a low score implying overlap between clusters. Values of k were swept over a predefined range, and configuration yielding the highest silhouette score was chosen. All features were preprocessed with the preexisting technique of one-hot encoding for categorical variables and standardization for numerical features. The algorithm was initialized using the *k-means++* strategy with a random seed of 66.

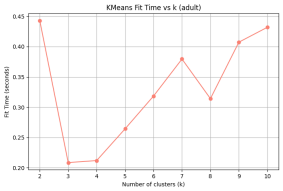
To determine the optimal number of clusters, the value of k ranging from $[2, 10]$ were evaluated on the Adult dataset. As seen in Figure 1, the silhouette score peaked at $k = 5$,



(a) Silhouette Score vs. k



(b) Inertia vs. k



(c) Fit-Time vs. k

Figure 1: K-Means Metrics vs. k – Adult

achieving a value of 0.163624. This represents the best balance between cluster cohesion and separation among the tested configurations. This is indicative of the clusters having not been separated from each other, likely having overlapping boundaries; low cohesion with the points within each cluster not being held tightly around the centroid.

As k increased, the inertia (which measures the internal coherence of clusters) decreased monotonically. This indicates that the data points on average were closer to their centroids, with lower inertia indicating that the clusters are more compact and tight. This suggests that even though cluster quality (silhouette score) improves until $k = 5$ before declining, additional clusters beyond this point likely introduce fragmentation rather than any meaningful structure.

The peak silhouette score (0.1636 at $k = 5$), is significantly lower than the typical threshold for well-separated clusters (> 0.5). This is indicative of the Adult dataset lacking strong cluster separability. This is likely due to the dataset's high dimensionality, mixed feature types, and overlapping class distributions. A reasonable explanation may be the presence of many one-hot encoded categorical variables, which may further reduce the effectiveness of distance-based clustering methods like K-Means.

3.1.2 Expectation Maximization (Gaussian Mixture Models

Expectation Maximization (EM) is an iterative optimization algorithm that finds the maximum likelihood of parameters in statistical models that depend on unobserved latent variables or missing data. It is used to train Gaussian Mixture Models (GMM), which are probabilistic models that assume all data points are generated from a mix of a finite number of Gaussian distributions with unknown parameters. GMM models were applied to the Adult dataset to capture more flexible cluster structures compared to K-Means. By modeling the data as a mixture of Gaussian distributions, there is room for overlapping and non-spherical clusters.

The number of components k was selected using both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), evaluated over a range $k = [2, 14]$. AIC and BIC are statistical tools used to identify the best-fitting model while penalizing complexity to prevent overfitting. AIC estimates the relative quality of statistical models, aiming to minimize information loss. BIC selects the best model based on Bayesian probability principles.

As shown in Figure 2, both AIC and BIC decreased steadily as k increased. The optimal value occurs at $k = 14$, where both metrics reached their minimum values. The monotonic decrease in AIC and BIC suggests that the model continues to improve its fit as more components are added, indicating a highly complex underlying data distribution. The model's likelihood is improving faster than the complexity penalty is growing. In comparison to K-Means, which identifies a relatively small number of clusters, GMM favors larger numbers of components to model data effectively. The larger k suggests a potential trade-off between model fit and interpretability. While GMM captures more nuanced structure, the resulting clusters may be harder to interpret and may reflect local variations rather than meaningful global groupings.

The Adult dataset appears to have a complex and overlapping structure, making it challenging for clustering algorithms to identify clearly separable groups. These findings suggest that clustering alone may not be sufficient to uncover strong structure in this dataset and motivate the use of dimensionality reduction techniques in subsequent experiments.

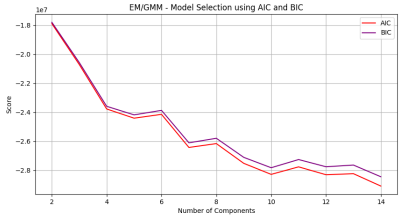


Figure 2: AIC and BIC score vs. Number of Components k – Adult

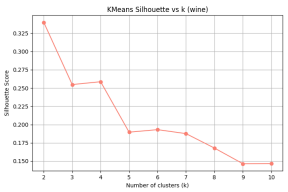
3.2 Wine Dataset

3.2.1 K-Means

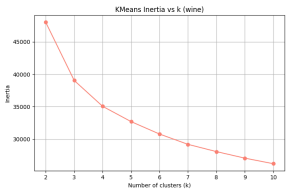
K-Means clustering was applied to the Wine dataset with the number of clusters k ranging from $[2, 10]$. Model selection was performed using the silhouette score again. As shown in Figure 3, the silhouette score peaks at $k = 2$, achieving a value of 0.339907. This score is significantly higher than that observed in the Adult dataset, indicating the strength of cluster separation in the Wine dataset. As k increases beyond the maximum, the silhouette score decays, suggesting that additional clusters can lead to over-segmentation and perhaps reduced cluster quality.

The inertia also decreases monotonically as expected, but does not exhibit a clear "elbow" point beyond $k = [2, 4]$. This supports the conclusion that a small number of clusters best captures the structure of the Wine dataset, which is expected for it being a smaller dataset.

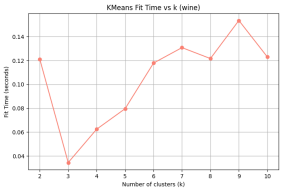
Overall, these results portray the Wine dataset as having a more well-defined and separable cluster structure compared to the Adult dataset. The optimal choice of $k = 2$ suggests that the Wine dataset naturally partitions into a small number of dominant groups, regardless of the multiclass structure of the data. This implies that some classes may be closely related or overlapping in the feature space.



(a) Silhouette Score vs. k



(b) Inertia vs. k



(c) Fit-Time vs. k

Figure 3: K-Means Metrics vs. k – Wine

3.2.2 Expectation Maximization (Gaussian Mixture Models

Expectation Maximization (EM) with Gaussian Mixture Models (GMM) were also applied to the Wine dataset, with the number of components k ranging from $[2, 14]$. Model selection was performed again using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), where lower values indicate better model fit while penalizing model complexity. As shown in Figure 4, both AIC and BIC decrease substantially as k increases. This indicates an improved model fit with the addition of more components. There is a tradeoff between model fit and an increase in model complexity; the rate of improvement decays beyond $k = 10$, suggesting diminishing returns from an increasingly complex model.

The optimal number of components differed slightly depending on the criterion. AIC selected $k = 14$, and BIC selected $k = 12$. AIC which is often used for prediction, favored a more complex model. However, since the Wine dataset only has 12 original features, the choice of k was effectively limited to at most 12 independent components. $k = 8$ was chosen to as the best alternative to the full feature space, since it portrayed the lowest score prior to the expected dramatic drop in score because of full feature space.

These results point to the Wine dataset having a rich underlying structure, requiring multiple Gaussian components to adequately model its distribution. Compared to K-Means which only identified two dominant clusters, EM/GMM reveal a finer complex substructure of the data, likely capturing overlapping or more nuanced relationships between features.

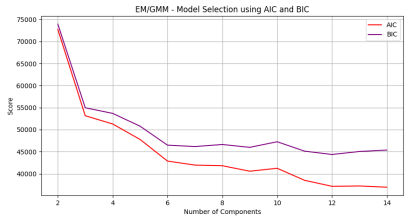


Figure 4: AIC and BIC score vs. k – Wine

Overall, the results from this experiment suggest that the Wine dataset is more amenable to clustering than the Adult dataset. Higher silhouette scores and clearer model selection (via AIC/BIC) indicate stronger inherent structure within the Wine data. In contrast, the Adult dataset's complexity and less separable tendency, requires more components in EM without achieving a stronger cluster separation.

4 Dimensionality Reduction on Raw Data

Three dimensionality-reduction techniques (Principal Component Analysis – PCA, Independent Component Analysis – ICA, and Random Projection – RP) were applied to both datasets. Prior to transformation, the same preprocessing/standardization techniques were applied. For each method, the target dimensionality was selected based on criteria such as explained variance ratio (PCA), component stability & non-Gaussianity (ICA) and reconstruction or distance presevation (RP).

4.1 Adult Dataset

4.1.1 Principal Component Analysis

PCA was applied to the Adult dataset after preprocessing (one-hot encoding and standardization as before), ensuring that all features contributed equally to this variance-based decomposition.

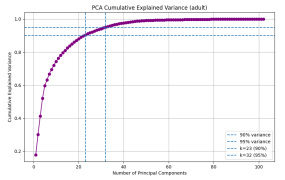
The primary metrics used to analyze this variance analysis was the explained variance ratio (EVR), which measures how much variance each component captures, and cumulative explained variance (CEV). The number of components k was selected based on CEV, which measures the proportion of the total dataset variance captured by the principal components. This metric is important for PCA since the method explicitly maximizes variance along orthogonal directions, prompting variance retention as a criterion for dimensionality selection.

As shown in Figure 5, the first few principal components capture a large portion of the dataset's variance, with diminishing contributions from later components. Specifically, 23 components were required to retain about 90% of the total variance, and 32 components to retain about 95%.

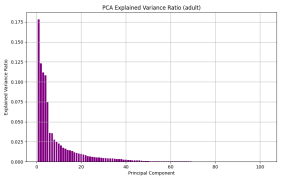
Based on the trade-off between dimensionality reduction and information retention, $k = 23$ was selected as the final dimensionality. This choice still is a significant compression from the initial 102 dimensions, while preserving dataset structure. The scree plot also supports this decision, showing a clear "elbow" in the early components with it eventually plateauing afterwards, indicative of diminishing returns. Overall, PCA reveals the Adult dataset contains a significant amount of redundancy and potentially irrelevant features, since a small subset of components capture most of the variance. 23 components retaining about 90% of the variance (from the original 102 features), indicates a substantial redundancy in the feature space. This motivates the use of dimensionality reduction to remove noise without major information loss.

4.1.2 Independent Component Analysis

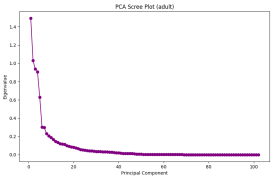
ICA was applied to the Adult dataset using a whitening transformation (*unit-variance*) using sklearn's FastICA, to ensure that components are uncorrelated and scaled.



(a) CEVR vs. k



(b) EV vs. k



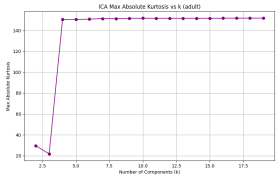
(c) Scree Plot

Figure 5: Principal Component Analysis – Adult

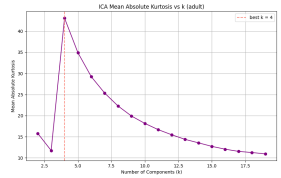
Unlike maximizing variance with PCA, ICA aims to recover statistically independent components by maximizing non-Gaussianity. The number of components k was selected using absolute kurtosis, a standard measurement of non-Gaussianity. A higher kurtosis is indicative of more non-Gaussian, more informative, components.

Two metrics were evaluated, mean absolute kurtosis and maximum absolute kurtosis. The mean absolute kurtosis was the primary selection criterion, since it depicts the overall quality of extracted components rather than any single component dominating. As shown in Figure 6, the mean absolute kurtosis peaks at $k = 4$, and declines afterwards. This means that increasing the number of components beyond 4 introduces less independent and more Gaussian components.

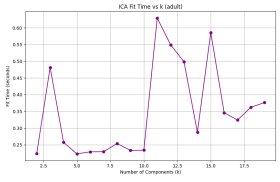
$k = 4$ was selected as the optimal dimensionality for ICA (mean absolute kurtosis of 43.111107, and max absolute kurtosis of 150.415299). Overall, ICA depicts a much more aggressive dimensionality reduction compared to PCA (4 vs. 23 components). This suggests that there are a small number of strongly independent features in the dataset. The rapid decline in kurtosis does suggest however, that ICA is sensitive in over-decomposition where adding more components may introduce noise rather than structure.



(a) Max Abs Kurtosis vs. k



(b) Max Abs Kurtosis vs. k



(c) Fit Time vs. k

Figure 6: Independent Component Analysis – Adult

4.1.3 Random Projection

RP was applied using Gaussian random matrices to reduce dimensionality while approximately preserving pairwise distances (*Johnson-Lindenstrauss lemma*).

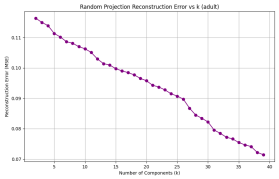
Unlike PCA and ICA, RP does not optimize a specific statistical property. It instead relies on random linear transformations, therefore evaluated using reconstruction error (mean squared error). This identifies how well the original data can be recovered after projection and inverse transformation.

As shown in Figure 7, reconstruction error decreases steadily as the number of components k increases. This monotonic behavior is expected, since higher-dimensional projections likely retain more information. Beyond $k = 25$, diminishing returns are shown. This indicates that increases in dimensionality yield only marginal improvements in reconstruction error.

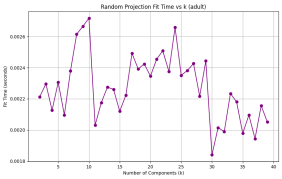
Based on this tradeoff between dimensionality reduction and information preservation, $k = 25$ was selected. Although reconstruction error decreases as the number of components increases, this is expected behavior for RP and does not provide a clear optimal point.

RP also exhibits extremely fast and stable fit times across all k values, outperforming PCA and ICA with its fastest fit time at $k = 10$. Since RP does not produce interpretable components and does not explicitly optimize for structure in the data, it will likely produce weaker representations for downstream use.

The differences observed in the application of three different dimensionality reduction methods highlight the tradeoff between each method. PCA preserves global variance and structure (opting for a larger number of components), ICA emphasizes independent components (with a smaller set of informative components), and RP provides a computationally efficient but less structured representation (higher dimensionality to maintain low reconstruction error). PCA likely will provide the most balanced representation for downstream tasks, with the latter two methods potentially suffering from aggressive reduction or lackluster feature representations.



(a) MSE vs. k



(b) Fit Time vs. k

Figure 7: Random Projection Analysis – Adult

4.2 Wine Dataset

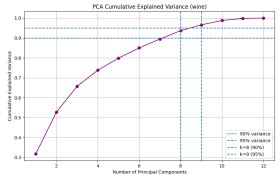
4.2.1 Principal Component Analysis

Following the same procedure and metrics for the Adult dataset, PCA was applied to the Wine dataset. As shown in Figure 8, the first principal component alone accounts for about 31.72% of the (explained) variance and the first three components accounting for about 66% of the total variance. This indicates that the Wine dataset has a compact structure, where important information is likely concentrated in a small number of dimensions.

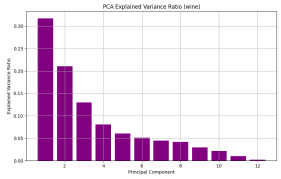
The cumulative explained variance plot depicts approximately 90% of the variance is captured with $k = 8$ components, while approximately 95% is captured with $k = 9$ components. $k = 8$ was chosen as the dimensionality, providing strong variance (cumulative variance of 0.936659) while achiev-

ing a strong reduction.

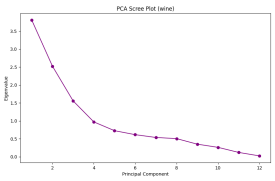
The scree plot also shows a distinct "elbow" around 3-4 components, followed by decaying performance. Compared to PCA on the Adult dataset, the Wine dataset requires significantly less components to capture most of the variance. This highlights the lower dimensionality and more structured nature of the Wine dataset, where features are more correlated and less redundant.



(a) CEVR vs. k



(b) EV vs. k



(c) Scree Plot

Figure 8: Principal Component Analysis – Wine

4.2.2 Independent Component Analysis

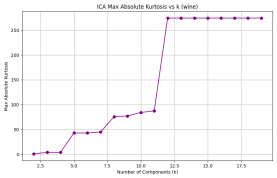
Following the same procedure for the Adult dataset, ICA was applied to the Wine dataset. As shown in Figure 9, the mean absolute kurtosis increases as the number of components k increases, with a sharp jump observed at $k = 12$. Beyond this, the kurtosis plateaus, indicating that additional components do not provide further gains in non-Gaussianity. This is also shown in the maximum absolute kurtosis, which also saturates at high values of k.

Since the Wine dataset only has 12 original features, ICA is effectively limited to at most 12 independent components. $k = 8$ was chosen since it still maintained the highest kurtosis after $k = 12$, with a mean absolute kurtosis of 13.161181. This effectively reduced the dimensionality from 12 to 8 features, avoiding overfitting to the full-dimensional representation while preserving the independent component structure.

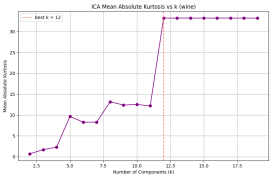
Unlike the Adult dataset, ICA on the Wine dataset fails to identify a small subset of dominant independent components. Since the highest kurtosis is achieved by retaining the full dimensionality (12 features), it suggests that the underlying structure of the Wine dataset is more evenly distributed across its features rather than concentrated in a few latent features. Therefore, ICA offers limited compression on the Wine dataset, which may lead to limited effectiveness for downstream usage.

4.2.3 Random Projection

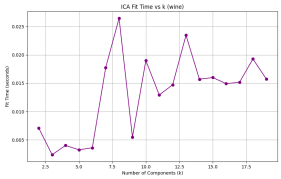
Following the same procedure for the Adult dataset, RP was applied to the Wine dataset. As shown in Figure 10, reconstruction error decreases monotonically as the number of components k increases. At $k = 12$, the reconstruction error approaches zero, which is expected since this corresponds to the original dimensionality of the dataset.



(a) Max Abs Kurtosis vs. k



(b) Max Abs Kurtosis vs. k

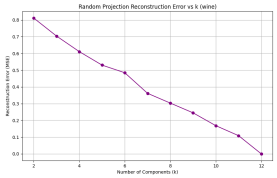


(c) Fit Time vs. k

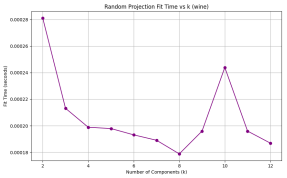
Figure 9: Independent Component Analysis – Wine

A trade-off between dimensionality reduction and information preservation was considered. Although the largest reductions in error occur for smaller k values, the reconstruction error still remained high. The error continues to decrease substantially through $k = 10$. Based on this balance, $k = 10$ was selected as the working dimensionality, preserving most of the original structure while still reducing the the dataset from 12 features to 10 features.

Compared to the Adult dataset, the Wine dataset exhibits a steeper reduction in reconstruction error. This indicates the Wine dataset's structure is easier to preserve in lower-dimensional random projections. This suggests that RP is more effective on the Wine dataset than on the Adult dataset, yet still is less interpretable than PCA and less structured than ICA. RP provides a computationally efficient representation that preserves much of the geometry of the Wine dataset, but the relatively mild dimensionality reduction with the best reconstruction error suggests limited compression benefits than PCA.



(a) MSE vs. k



(b) Fit Time vs. k

Figure 10: Random Projection Analysis – Wine

Overall, PCA provided the most efficient dimensionality reduction by preserving majority of the variance with relatively fewer components. ICA captured non-Gaussian structure in the datasets, but was less stable and less interpretable. RP preserved distance structure with low computational cost, but required higher dimensionality to achieve comparable reconstruction error, especially on the Adult dataset.

5 Clustering in Reduced Dimensional Spaces

Having established reduced-dimensional representations using PCA, ICA, and RP, the next step is to evaluate how these transformations affect clustering performance. K-Means & EM/GMM were reapplied to each transformed space to deter-

mine whether dimensionality reduction improves cluster quality and separability relative to the original feature spaces. The number of components k used defined in previous steps were: **Adult** — $PCA = 23, ICA = 4, RP = 25$; **Wine** — $PCA = 8, ICA = 8, RP = 10$. The number of components k for the clustering algorithms were set: **Adult** — $kmeans = 5, EM = 14$; **Wine** — $kmeans = 2, EM = 8$.

5.1 Adult Dataset

Dimensionality reduction had a significant but algorithm-dependent impact on clustering performance and training time. For K-Means, ICA achieved the highest silhouette score compared to the RAW (original) dataset: 0.357 vs. 0.164. ICA improves K-Means performance because it transforms the data into statistically independent components that reduce redundancy and correlation between features. In the original Adult dataset, many features (including one-hot encoded categorical variables) introduce sparsity and a correlated structure that can distort Euclidean distances. K-Means utilizes Euclidean distances to assign cluster membership. Highly correlated features lead to less meaningful distance calculations. ICA "untangles" overlapping feature interactions, producing a distinction of latent factors. This explains why ICA also had the significantly lower inertia (41388.04), where the points are close to their centroids with much better cluster separation (in line with K-Means' assumption of isotropic cluster structure). In contrast, PCA and RP provided modest improvements over the RAW feature space (silhouette score: 0.185, 0.192).

Table 1: Clustering performance by representation – Adult

Algorithm	Representation	Silhouette	Inertia	Fit Time	BIC	AIC
KMeans	RAW	0.1636	196008.75	0.5757	–	–
KMeans	PCA	0.1848	165308.04	0.1914	–	–
KMeans	ICA	**0.3571**	**41388.04**	**0.1156**	–	–
KMeans	RP	0.1922	175910.16	0.1699	–	–
EM	RAW	–	–	7.5019	**-2.8454e+07**	**-2.9095e+07**
EM	PCA	–	–	2.0556	-1.5998e+06	-1.6357e+06
EM	ICA	–	–	**1.0313**	-1.2657e+05	-1.2836e+05
EM	RP	–	–	2.5355	-1.4374e+06	-1.4794e+06

For EM/GMM, clustering on the RAW feature space yielded the best results, as indicated by the lowest (most negative) BIC and AIC values. The degradation in EM performance after dimensionality reduction is likely due to EM's assumption that the data is generated from a mixture of Gaussian distributions (covariance structure is important for component likelihoods). Among the reduced representations, PCA and RP performed moderately, likely discarding important features need for the estimating components. ICA performed the worst since it maximizes non-Gaussianity, contradicting Gaussian assumptions of EM.

These interesting findings highlight that the effectiveness of dimensionality reduction is highly algorithm-dependent. ICA is particularly beneficial for distance-based methods like K-Means, while EM benefits from retaining the original full feature space. These results partially support Hypothesis 1, as ICA significantly improved clustering performance for K-Means, while PCA and RP provided only modest gains. However, dimensionality reduction degraded EM performance, which suggests the hypothesis holds only for certain clustering algorithms.

(a) Silhouette Score vs. k (b) Inertia vs. k

(c) BIC vs. k

Figure 11: Clustering Reduced Analysis – Adult

5.2 Wine Dataset

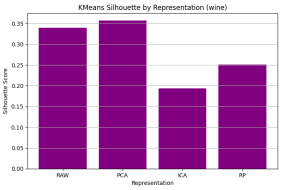
Like for the Adult dataset, dimensionality reduction had a significant impact on clustering performance and training time for the Wine dataset. For K-Means, PCA provided the best clustering performance with the highest silhouette score (0.357 vs. 0.340 - RAW). This suggests that variance-preserving transformations are well-suited for this dataset. ICA portrayed the lowest inertia (39232.89), indicating an improved cluster separation. However, ICA produced the lowest silhouette score (0.1939), with RP providing moderate performance between RAW and ICA.

For EM/GMM, clustering on the raw feature space yielded the best results (lowest BIC/AIC values) yet again. All reduced representations resulted in worse performance, especially PCA and ICA. This suggests that reducing dimensionality may have removed important distributional information required for probabilistic clustering. ICA attempts to decompose the data into independent non-Gaussian components, which may not exist strongly in this dataset (which explains the weak cluster performance).

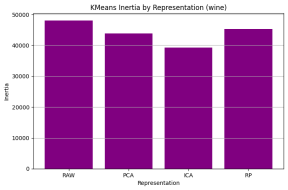
Overall, PCA improved clustering performance for K-Means on the Wine dataset, partially supporting Hypothesis 1. Conversely, dimensionality reduction consistently degraded performance for EM on this dataset. Unlike the Adult dataset where ICA was the most beneficial, the Wine dataset benefits from PCA because its features are continuous, low-dimensional, and more strongly correlated. Preserving variance through PCA reinforces cluster separability without distorting the underlying structure. This interesting find further supports the claim that the effectiveness of dimensionality reduction methods depends on the underlying structure of the data.

Table 2: Clustering performance by representation – Wine

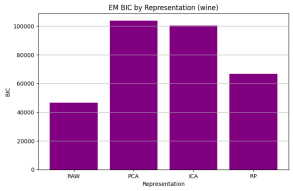
Algorithm	Representation	Silhouette	Inertia	Fit Time	BIC	AIC
KMeans	RAW	0.3399	48002.16	0.1222	–	–
KMeans	PCA	**0.3567**	43901.41	**0.0414**	–	–
KMeans	ICA	0.1939	**39232.89**	0.0612	–	–
KMeans	RP	0.2509	45280.61	0.0520	–	–
EM	RAW	–	–	0.1309	**46633.69**	**41823.49**
EM	PCA	–	–	**0.0874**	103702.45	101327.13
EM	ICA	–	–	0.1252	100481.99	98106.67
EM	RP	–	–	0.1035	66787.57	63300.67



(a) Silhouette Score vs. k



(b) Inertia vs. k



(c) BIC vs. k

Figure 12: Clustering Reduced Analysis – Wine

6 Neural Networks on Dimensionality Reduced Data

To further assess the practical impact of dimensionality reduction, the best-performing neural network model from earlier projects was retrained on each reduced representation (PCA, ICA, and RP), as well as the original feature space (RAW). This allows for a true assessment of whether structural changes introduced by dimensionality reduction can translate into improvements in classification performance, generalization, or training time. The Adult dataset was used here to observe whether dimensionality reduction on a large dataset can improve neural network performance. The number of components k used defined in previous steps were: $PCA = 23, ICA = 4, RP = 25$. The number of components k for the clustering algorithms were set: $kmeans = 5, EM = 14$.

Dimensionality reduction had mixed effects on neural network performance. The same evaluation metrics were used, including test accuracy, test F1-score, and test loss. PCA achieved the best overall performance, yielding the highest test F1-score (0.672), highest test accuracy (0.797), as well as the lowest test loss (0.639). Notably, PCA reduced the feature space from 102 to 23 dimensions, while still maintaining predictive performance compared to the RAW feature space. This reduction demonstrates that similar predictive performance can be maintained with significantly lower input complexity.

The RAW representation performed nearly as well as PCA, with only marginally lower accuracy (0.792) and F1 score (0.672). This is indicative of the neural network's ability to effectively learn from the original feature space (since the neural network architecture is deep), and the dataset does not suffer significantly from high-dimensional noise or redundancy. In contrast, RP resulted in a small but consistent drop in performance (accuracy: 0.791, F1: 0.666), indicating that while it preserves some structure, it does not retain information as effectively as PCA. ICA performed the worst across all metrics, with substantially lower accuracy (0.752), F1-score (0.628), and test loss (0.678). This suggests ICA's objective of maximizing statistically independent components does not suffice for this classification task, likely due to the loss of important predictive information.

In terms of training behavior, reduced representations generally required more updates to reach the best validation performance (especially for ICA, which converged slowly). PCA and RP reduced overall training time due to a lower input dimensionality, leading to improved computational efficiency. Furthermore, the representations differed in convergence behavior. The RAW model converged the fastest, reaching the best validation performance in 900 steps. PCA and RP required more training (3300 and 2100 steps), whereas ICA required substantially longer (6100 steps). This indicates that the RAW representation provides the most direct optimization path for the network.

These results suggest dimensionality reduction is not beneficial universally, but does improve efficiency and performance of the neural network when applied appropriately. PCA removes redundancy while preserving variance (PCA), making it more beneficial for neural network classification performance. This is because in highly complex datasets like the Adult dataset, many features (especially after one-hot encoding) can introduce noise and multicollinearity. This can slow optimization and impact gradient updates (regardless of the optimizer strength). This means a reduction in input complexity can aid in model generalization, which supports Hypothesis 2. Further supporting the hypothesis, aggressive or misaligned dimensionality reduction proved to be detrimental to model performance. ICA reduced the feature space to 4 dimensions, likely discarding important information. This confirms the importance of balancing the right transformation and resulting dimensionality, rather than dimensionality reduction alone.

Table 3: Neural network performance by representation – Adult

Representation	Input Dim	Test Loss	Test Acc	Test F1	Best Val	Best Step	Train Time
RAW	102	0.6395	0.7919	0.6719	**0.6124**	**900**	16.1722
PCA	23	**0.6392**	**0.7966**	**0.6721**	0.6237	3300	**14.3544**
ICA	4	0.6775	0.7518	0.6278	0.6596	6100	14.3669
RP	25	0.6543	0.7910	0.6660	0.6295	2100	14.4943

7 Neural Networks with Cluster Derived Features

To further explore how clustering structure can enhance predictive performance, cluster-derived features were generated from the models in the first experiment, and incorporated into the existing neural network architecture. These features included representations such as cluster assignments and probabilistic outputs. The aim is to capture latent group structure in the data for downstream use. The neural network was retrained using these augmented inputs and compared against the baseline model (using the original feature space – RAW), to evaluate the impact of clustering on classification performance beyond the original feature space.

The same preprocessing technique was used in previous experiments, using the aforementioned number of components k defined in previous steps: $PCA = 23, ICA = 4, RP = 25$. The number of components k for the clustering algorithms were again set as: $kmeans = 5, EM = 14$.

The representations of the cluster-derived features were evaluated independently and combined with the original RAW data. So, five representations were evaluated with the existing neural network architecture: RAW, K-Means, EM/GMM, RAW + K-Means, and RAW + EM/GMM.

Cluster-derived features had varying effects on neural network performance, depending on their representation. Using cluster features alone (K-Means or GMM) significantly degraded performance. Both representations individually yielded

lower accuracy and F1 scores compared to the RAW feature space. This suggests that cluster assignments alone likely discard too much fine-grained information necessary for accurate classification.

Augmenting the original feature space with cluster-derived features produced favorable results. The RAW + GMM representation achieved the highest test accuracy (0.808) and test F1-score (0.682), outperforming the baseline model. This indicates that probabilistic cluster features can provide useful complementary structure beyond the original feature space. On the contrary, RAW + K-Means did not improve performance and slightly reduced predictive quality. This suggests that hard cluster assignments are likely less informative than soft probabilistic representations.

In terms of training behavior, models using only cluster-derived features required more training steps (i.e. K-Means requiring 9500 steps to reach the best validation). This is indicative of a more difficult optimization for the model, with augmented representations showing moderate and stable increases in training steps.

These results contradict Hypothesis 3. While cluster-derived features alone degraded performance, augmenting the original feature space with cluster-derived features as complementary inputs proved to be beneficial. GMM-based features provided richer structural information than K-Means, with both improved accuracy and F1-score. This means that probabilistic cluster information has the ability to enhance neural network classification performance.

Table 4: Neural network performance by representation – Cluster Features (Adult)

Representation	Input Dim	Test Loss	Test Acc	Test F1	Best Val	Best Step	Train Time
RAW+GMM	116	0.6396	**0.8081**	**0.6820**	0.6158	1000	17.4031
RAW	102	0.6395	0.7919	0.6719	0.6124	**900**	17.4513
RAW+K-Means	107	**0.6251**	0.7839	0.6664	**0.6032**	2300	17.3780
K-Means	5	0.7531	0.7131	0.5948	0.7363	9500	16.2570
GMM	14	0.8291	0.6872	0.5801	0.8131	900	**16.0817**

8 Conclusion

This study demonstrates that the effectiveness of unsupervised learning methods is highly dependent on both the structure of the dataset and the assumptions of downstream models. Overall, dimensionality reduction had a significant impact on clustering behavior, with PCA and ICA often improving cluster separability compared to the original feature space. In particular, ICA produced the strongest gains for K-Means on the Adult dataset because it transformed the feature space into statistically independent dimensions that better align with K-Means' Euclidean distance reliance. This suggests that non-Gaussian structure can enhance cluster formation when properly captured. (Hypothesis 1 partially supported)

These improvements in clustering quality did not consistently translate to better predictive performance. PCA provided a strong balance between reduction and retention, leading to improved neural network performance. ICA however, often removed too much relevance, resulting in degraded classification performance. RP preserved performance but at the cost of interpretability and not outperforming PCA. (Hypothesis 2 supported)

Incorporating cluster-derived features further highlighted the importance of structural information representation. Cluster features alone were insufficient for prediction, only improving performance when combined with the original feature space. The combination of GMM-based representations improved performance introducing a probabilistic structure to the feature space, whereas hard cluster assignments (K-Means) proved to be less informative. (Hypothesis 3 contradicted)

Overall, these experiments highlight the importance of the application of unsupervised learning methods. When they are applied in a way that benefits both the structure of the data and assumptions of downstream models, they prove to be highly effective. Dimensionality reduction and clustering work best when they preserve or complement the original feature space, rather than replacing it entirely.

9 References

- GeeksforGeeks. "Principal Component Analysis(PCA)." GeeksforGeeks, 7 July 2018, www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/.
- GeeksforGeeks. "Independent Component Analysis ML." GeeksforGeeks, 21 May 2019, www.geeksforgeeks.org/machine-learning/ml-independent-component-analysis/.
- Gupta, Mehul. "Random Projection for Dimension Reduction." Data Science in Your Pocket, 13 June 2023, medium.com/data-science-in-your-pocket/random-projection-for-dimension-reduction-27d2ec7d40cd.
- GeeksforGeeks. "K Means Clustering – Introduction." GeeksforGeeks, 2 May 2017, www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/.
- GeeksforGeeks. "ExpectationMaximization Algorithm ML." GeeksforGeeks, 14 May 2019, www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/.
- GeeksforGeeks. "Gaussian Mixture Model." GeeksforGeeks, 24 Aug. 2018, www.geeksforgeeks.org/machine-learning/gaussian-mixture-model/.
- Diaries, Jani Data. "Choosing the Best Model: A Friendly Guide to AIC and BIC." Medium, 7 Nov. 2024, medium.com/@jshaik2452/choosing-the-best-model-a-friendly-guide-to-aic-and-bic-af220b33255f.
- Macqueen, J. SOME METHODS for CLASSIFICATION and ANALYSIS of MULTIVARIATE OBSERVATIONS. 1967.
- Jolliffe, Ian T, and Jorge Cadima. "Principal component analysis: a review and recent developments." Philosophical transactions. Series A, Mathematical, physical, and engineering sciences vol. 374,2065 (2016): 20150202. doi:10.1098/rsta.2015.0202