UPC UB URV

MASTER IN ARTIFICIAL INTELIGENCE

Project 2: Dimensionality Reduction and Visualization

Introduction to Machine Learning

Professor:
Dr. Maria Salamó
Course: 2023-2024.
Laboratory group 12L

Students:
Wafaa Guendouz
Alam Lopez
Laura Roldán
Mario Rosas

November 13, 2023

Project 2

Dimensionality Reduction and Visualization

Table of Contents

1 Introduction

In the realm of designing machine learning algorithms, a common challenge we encounter is the issue of data dimensionality. While it might seem like having more features would automatically make our data more descriptive and our models perform better, the reality usually differs from this. Not all features contribute equally to a given task. On the contrary, having too many features complicates both the processing and interpretation of data, leading to what is known as the "curse of dimensionality.", which makes it challenging to comprehend and visualize a multi-dimensional array.

Hence, it's crucial to address the challenge posed by high-dimensional data, aiming for more efficient processing and meaningful interpretation. Specific techniques, such as dimensionality reduction and visualization, are tailored to deal with these challenges. They help to extract essential information from the data while avoiding the computational burden of excessive dimensions. At the same time, we need to be mindful of the potential impact of applying these techniques on the performance of our models.

The focus of this assignment is to understand how to practically implement these techniques. This involves gaining proficiency in applying them across various datasets, understanding when to use them, and assessing how they might affect the performance of our models. The aim is to equip ourselves with the knowledge and skills needed to navigate the complexities of high-dimensional data, facilitating both efficient processing and accurate interpretation—an essential balance in creating effective machine-learning models.

2 PCA, TruncatedSVD, Isomap

- **Principal Component Analysis (PCA)**: Is a linear dimensionality reduction technique that identifies the principal components, or axes, in the data that capture the maximum variance. By projecting the data onto these principal components, PCA simplifies high-dimensional data while retaining its essential variability.
- **Truncated Singular Value Decomposition (TruncatedSVD)**: Is another linear dimensionality reduction method, particularly useful for sparse data. It operates on the singular value decomposition of the data matrix, retaining a specified number of singular values and their corresponding vectors. TruncatedSVD is efficient for large, sparse datasets and often used in text mining and natural language processing.
- **Isomap (Isometric Mapping)** : Is a non-linear dimensionality reduction technique that focuses on preserving the geodesic distances between all data points. It constructs a neighborhood graph based on pairwise distances, capturing the intrinsic geometry of the data. Isomap is especially effective for datasets with non-linear structures, providing a low-dimensional representation that respects the underlying manifold's topology.

3 Experiments description

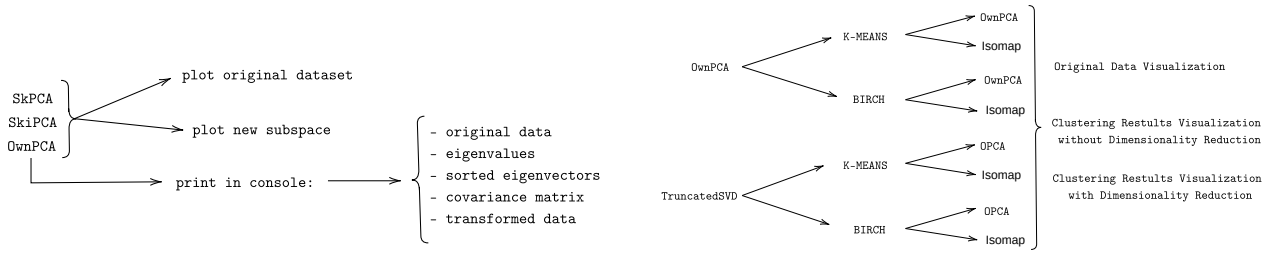
In our experimentation, we implemented a custom Principal Component Analysis (PCA) algorithm from scratch and compared its performance with Sklearn's implementations of PCA, Incremental PCA, and Truncated Singular Value Decomposition (TruncatedSVD). We assessed mainly the quality of dimensionality reduction achieved by each method.

Subsequently, we integrated these dimensionality reduction techniques into our clustering experiments using our own implementation of Kmeans and Sklearn's Birch. To gain insights into the impact of feature visualization, we employed PCA and Isomap to visualize the data before and after clustering with and without dimensionality reduction, exploring both linear and non-linear representations. A graphical representation of the process is shown in the figure 1.

The PCA implementation in the code follows the assignment guidelines, displaying key information on the console. This includes details about the original matrix, covariance matrix, eigenvectors, eigenvalues, and the transformed dataset post-PCA.

An essential consideration in PCA is determining the number of components. This is explored through the cumulative explained variance ratio, assessing whether it retains enough information from the original data. The goal is to select a number of components where the cumulative explained variance surpasses a specified threshold, set at 85% in this case.

Once the number of components is established, it is used to execute sklearn's PCA and IncrementalPCA methods.



(a) Dimensionality Reduction

(b) Feature Visualization

Figure 1: Experiments

4 Dataset 1: Waveform

4.1 Dimensionality reduction

The first dataset is numeric, featuring 5000 samples, 40 features, and 3 classes. Following the correlation matrix analysis from the previous report, it is spotlighted pairs of features with the highest positive and negative correlations: x15 and x16 with a correlation of 0.714, and x7 and x14 with a correlation of -0.710. Observe figure 2 for a visual representation of these features.

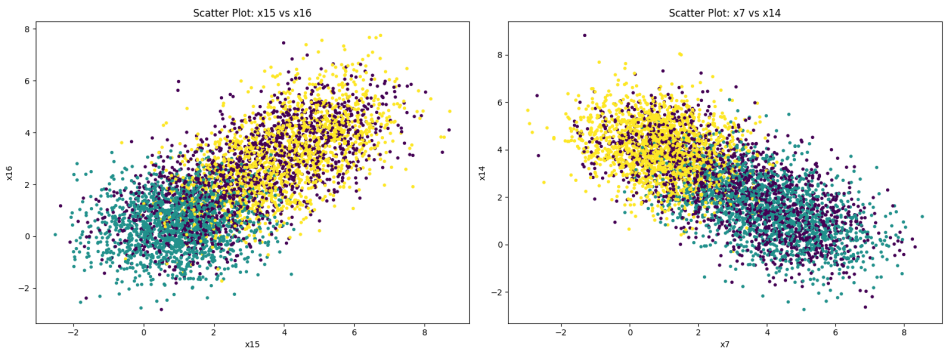


Figure 2: Pair of features from original dataset

Post-PCA, the dataset's dimensionality shrinks to 26 components, with figure 3 illustrating the components selected to maintain at least 85% of the information.

Applying the same number of components to both custom PCA and Sklearn methods, it is observed the following information retention percentages:

Retention Percentages		
Custom PCA	Sklearn PCA	Sklearn In. PCA
86.7%	86.72%	86.44%

For visualization purposes, Figure 4 illustrates the representation of the first two PCA components for each PCA method.

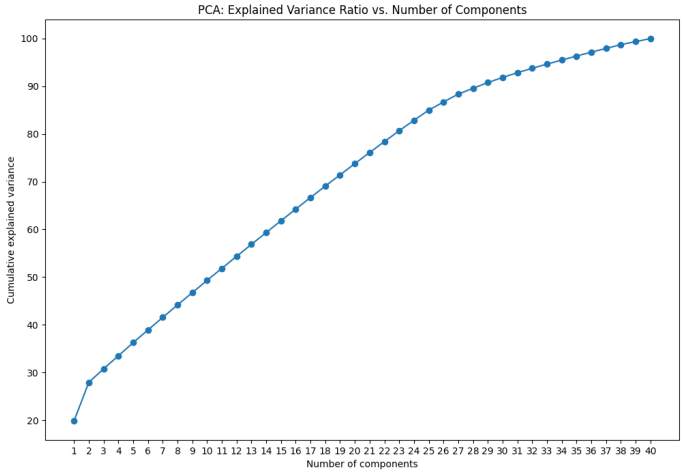


Figure 3: Cumulative explained variance

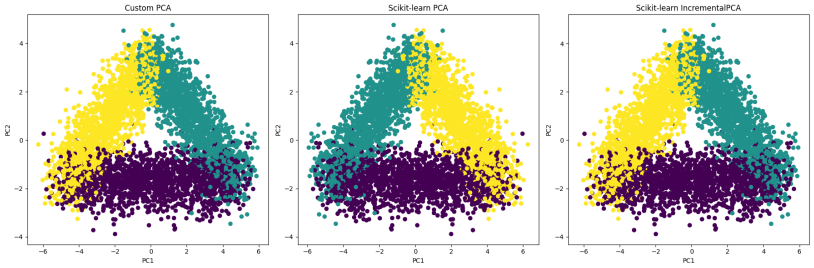


Figure 4: Comparison PCA methods

Upon scrutinizing the results, minimal distinctions are observed among the methods for this dataset. Prior to engaging in clustering, TruncatedSVD is performed with the same number of components.

4.2 Clustering

4.2.1 K-means

Table 1: Kmeans clustering results

Method	**Accuracy**	**Homogenei.**	**Completen.**	**V_measure**	**ARI**	**AMI**	**Silhouette**	**Davies**
No Reduc.	0.6448	0.2928	0.3434	0.3161	0.2294	0.3158	0.0463	3.5467
PCA	0.655	0.2833	0.3452	0.3112	0.2138	0.3109	0.0646	1.5613
TSVD	0.6454	0.2964	0.3462	0.3194	0.2344	0.3191	0.2047	1.4239

Dimensionality reduction improved the clustering performance in terms of cluster separation and cohesion, as evidenced by the silhouette scores and Davies Index. but the impact on other metrics like homogeneity, completeness, and adjusted mutual information is less pronounced. The waveform dataset seemed to respond positively to dimensionality reduction, suggesting that the feature space might have contained redundant or less informative features. The choice between TSVD and PCA might depend on specific requirements of the analysis, such as the desired balance between accuracy and interpretability

4.2.2 BIRCH

Table 2: BIRCH clustering

		Homogeneity	Completeness	V_measure	ARI	AMI	Silhouette
No dimensionality reduction	Threshold	0.1	0.1	0.1	0.1	0.1	0.1
	Branching Factor	10	10	10	10	10	10
	#Clusters	3	3	3	3	3	3
	SCORE	**0.349**	**0.35**	**0.349**	**0.286**	**0.349**	**0.085**
PCA	Threshold	0.1	0.1	0.1	0.1	0.1	0.1
	Branching Factor	10	10	10	10	10	10
	#Clusters	3	3	3	3	3	3
	SCORE	**0.337**	**0.349**	**0.343**	**0.264**	**0.342**	**0.101**
TruncatedSVD	Threshold	0.1	0.1	0.1	0.1	0.1	0.1
	Branching Factor	10	10	10	10	10	10
	#Clusters	3	3	3	3	3	3
	SCORE	**0.34**	**0.351**	**0.346**	**0.263**	**0.345**	**0.102**

The algorithm successfully identified three distinct clusters in the data for the three cases. Despite adjustments to hyperparameters such as threshold, branching factor, and fixed cluster count at 3, the clustering performance remained moderate. Silhouette scores across all scenarios indicated challenges in achieving well-separated clusters. Interestingly, the choice between PCA and TruncatedSVD did not significantly influence the outcomes.

4.3 Visualization

In the feature visualization experiments conducted for the Waveform dataset, the initial observation of the original data revealed a complex and dense arrangement, with classes intricately intertwined.

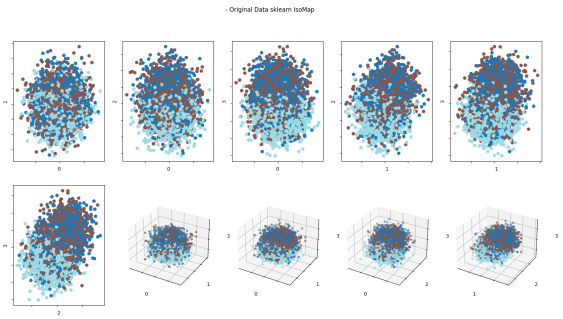
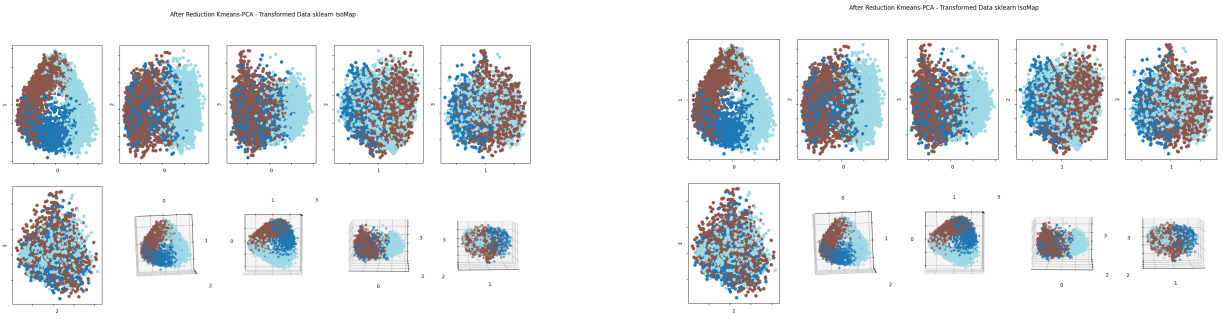


Figure 5: Original data visualization

However, after applying k-means clustering with PCA for dimensionality reduction and Isomap for feature visualization in the first experiment, a notable transformation occurred. The data became well-separated within the first three

principal components, suggesting that the combination of k-means and PCA successfully unraveled the underlying patterns.

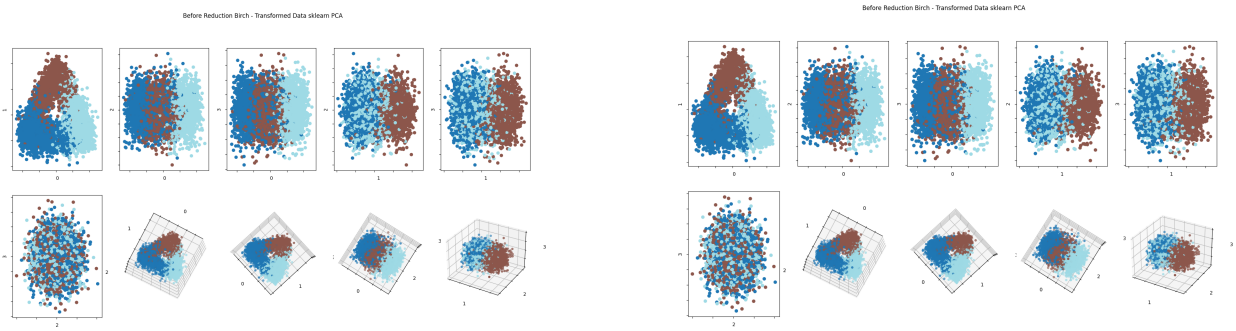


(a) Clustering before reduction

(b) Clustering after reduction

Figure 6: Waveform feature visualization with Kmeans-PCA-Isomap

The previous enhanced separation was further improved in the second experiment, where BIRCH clustering with PCA for both dimensionality reduction and feature visualization was employed. The results indicated a slight but discernible refinement in the data's separation, emphasizing the efficacy of BIRCH in achieving a more refined clustering and dimensionality reduction outcome.



(a) Clustering before reduction

(b) Clustering after reduction

Figure 7: Waveform feature visualization with Birch-TSVD-PCA

5 Dataset 2: Kr-vs-Kp

5.1 Dimensionality reduction

The second dataset is Kr-vs-Kp. It is categorical, has 3196 samples, 36 features and 2 binary classes. Each sample means a position of the board.

For better selection of the features, we analysed the meaning of each one. We came up with this theoretical relations:

- `bxqsq` (Black controls queening square): It may influence the ability of Black to renew the skewer threat (`reskr`). The control of the queening square could contribute to the simplicity of the position (`simpl`).

- `reskr` (Black alone can renew the skewer threat): This feature might increase the risk of the Black rook being captured safely (`rimmx`). It could be influenced by Black's control of the queening square (`bxqsq`).

- `rimmx` (Black rook can be captured safely): The safety of the Black rook might be influenced by the ability to renew the skewer threat (`reskr`). A captured rook could indicate a straightforward position (`simpl`).

- `simpl` (A very simple pattern applies): Indicates the presence of a very simple pattern, which may correlate with the safety of the Black rook (`rimmx`) and Black's control of the queening square (`bxqsq`).

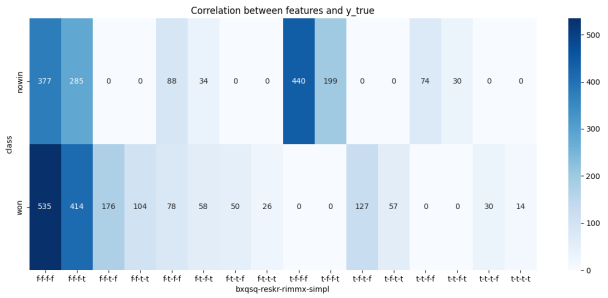


Figure 8: Correlation between features `bxqsq`, `reskr`, `rimmx`, `simpl`

Each of these features can take on values of true or false. Figure 8 visually represents the correlation between these features and the class distinctions (win or no win). Notably, the first two configurations suggest a potential for confusion during clustering. However, the ninth configuration appears more distinct and clear in its correlation.

Post-PCA, the dataset's dimensionality shrinks to 17 components, with figure 9 illustrating the components selected to maintain at least 85% of the information.

Applying the same number of components to both custom PCA and Sklearn methods, it is observed the following information retention percentages:

Retention Percentages		
Custom PCA	Sklearn PCA	Sklearn In. PCA
86.09%	86.10%	84.55%

For visualization purposes, Figure 10 illustrates the representation of the first two PCA components for each PCA method.

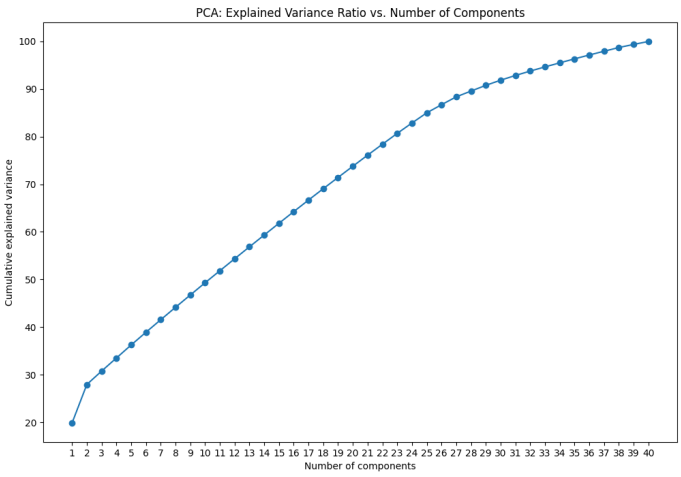


Figure 9: Cumulative explained variance

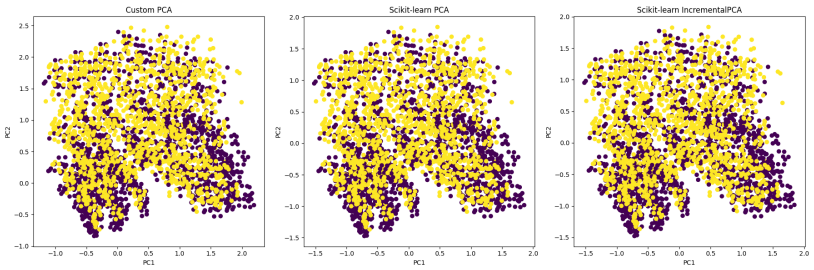


Figure 10: Comparison PCA methods

Upon scrutinizing the results, minimal distinctions are observed among the methods for this dataset. Prior to engaging in clustering, TruncatedSVD is performed with the same number of components.

5.2 Clustering

5.2.1 K-means

Table 3: Kmeans clustering results

Method	Accuracy	Homogeneity	Completeness	V_measure	ARI	AMI	Silhouette	Davies
No Reduction	0.6008	0.0342	0.0461	0.0393	0.0398	0.0390	0.0986	2.6369
TSVD	0.5876	0.0236	0.0237	0.0237	0.0304	0.0235	0.2066	1.9351
PCA	0.5854	0.0218	0.0275	0.0243	0.0284	0.0241	0.2156	1.6337

TSVD and PCA improved the clustering quality in terms of cluster separation and cohesion. However, the overall effectiveness of K-means on this dataset appears limited, as seen in the low homogeneity, completeness, and other related metrics.This might suggest the need for alternative clustering algorithms or a re-evaluation of the dataset's suitability for clustering.

5.2.2 BIRCH

Table 4: BIRCH clustering

		Homogeneity	Completeness	V_measure	ARI	AMI	Silhouette
No dimensionality reduction	Threshold	0.5	0.5	0.5	0.5	0.5	0.1
	Branching Factor	10	10	10	10	10	20
	SCORE	**0.056**	**0.038**	**0.045**	**0.035**	**0.044**	**0.101**
PCA	Threshold	0.9	0.9	0.9	0.9	0.9	0.1
	Branching Factor	40	40	40	40	40	20
	SCORE	**0.059**	**0.039**	**0.047**	**0.049**	**0.047**	**0.116**
TruncatedSVD	Threshold	0.9	0.9	0.9	0.9	0.9	0.2
	Branching Factor	20	20	20	20	20	20
	SCORE	**0.061**	**0.039**	**0.048**	**0.046**	**0.048**	**0.117**

The algorithm successfully identified three distinct clusters in the data for the three cases. Silhouette scores consistently indicate challenges in achieving well-defined and separated clusters. Homogeneity, Completeness, and V_measure scores are balanced but suggest suboptimal clustering. Notably, the choice between PCA and TruncatedSVD does not significantly impact outcomes.

5.3 Visualization

In the initial examination of the Kr-vs-Kp dataset, characterized by its fully categorical nature, the visualization of the original data exhibited a distinctive pattern of focal points. This concentration indicated that the classes were strongly defined but in a limited region of the space, reflecting the categorical nature of the features.

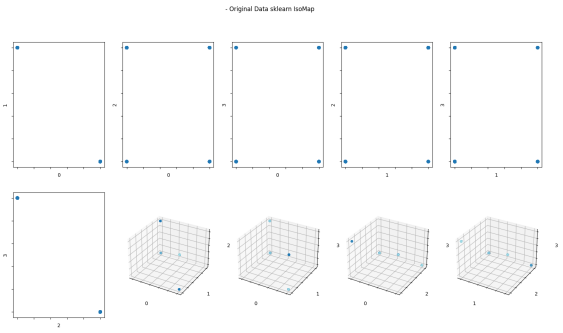
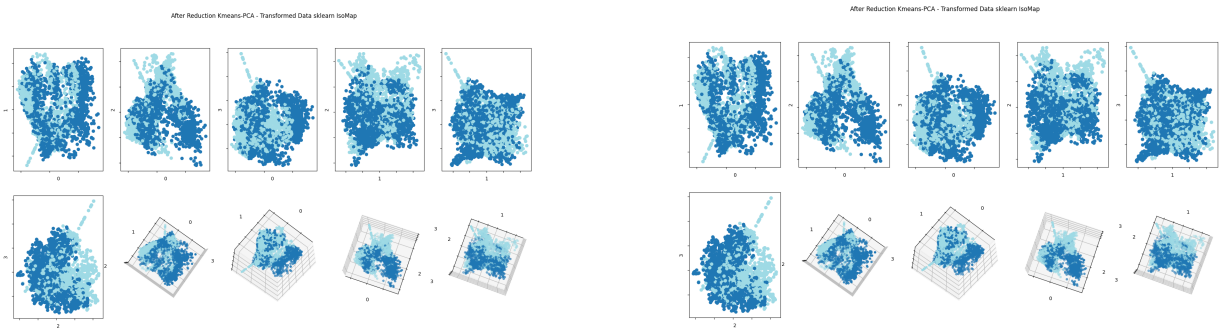


Figure 11: Original data visualization

Following the first experiment, which involved applying k-means clustering with PCA for dimensionality reduction and Isomap for feature visualization, a notable transformation occurred. The data became visibly dispersed across a broader region of the space, revealing a more intricate and spread-out distribution. In some perspectives, clusters appeared somewhat blurry, suggesting a more nuanced and complex relationship among the categorical features. This

dispersion indicated that the combination of k-means and PCA with Isomap successfully brought out subtle patterns, contributing to a richer representation of the categorical data.

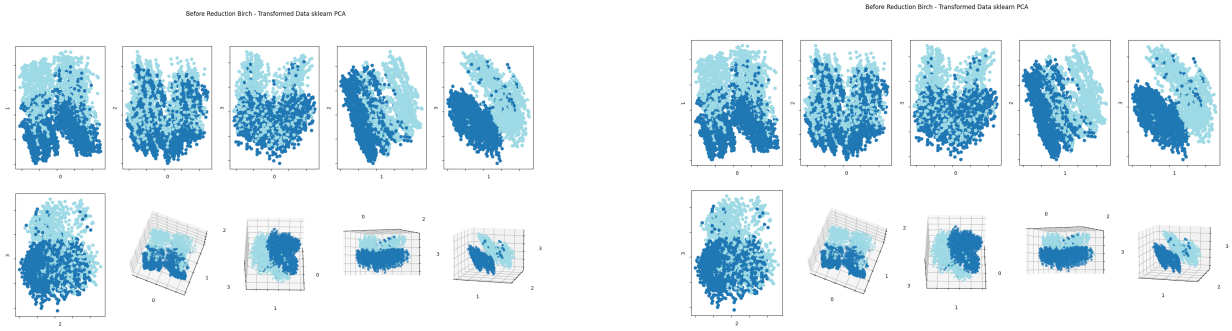


(a) Clustering before reduction

(b) Clustering after reduction

Figure 12: Kr-vs-Kp feature visualization with Kmeans-PCA-Isomap

In the second experiment introduced BIRCH clustering with PCA for both dimensionality reduction and feature visualization. While not achieving perfection, this approach contributed to a more refined separation with increased distances between the groups of classes. The resulting visualization demonstrated a clearer distinction among the classes, highlighting the potential of BIRCH in handling fully categorical datasets and improving separation.



(a) Clustering before reduction

(b) Clustering after reduction

Figure 13: Kr-vs-Kp feature visualization with Birch-TSVD-PCA

6 Dataset 3: Vowel

6.1 Dimensionality reduction

The third dataset comprises a mixture of features, totaling 990 samples. It encompasses 10 numerical features, 3 categorical features, and is categorized into 11 classes, vowel sounds.

For analysis, the focus is on the numerical features representing speech signals, along with a categorical feature denoting gender. In Figure 14, the visualization illustrates the relationship between gender and three specific features. Notably, the initial feature exhibits comparable values for both genders. Feature 7 distinctly highlights vowel variations

in females, while feature 8 is more indicative for males.

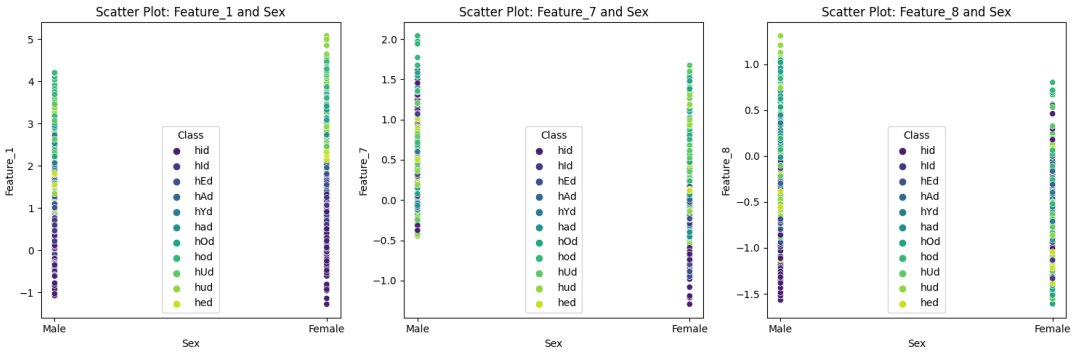


Figure 14: Correlation between features 1, 7 and 8 with sex

Post-PCA, the dataset's dimensionality shrinks to 4 components, with figure 15 illustrating the components selected to maintain at least 85% of the information.

Applying the same number of components to both custom PCA and Sklearn methods, it is observed the following information retention percentages:

Retention Percentages		
Custom PCA	Sklearn PCA	Sklearn In. PCA
86.73%	86.73%	85.75%

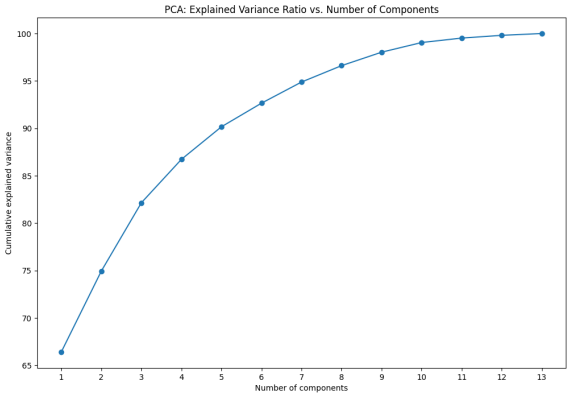


Figure 15: Cumulative explained variance

For visualization purposes, Figure 16 illustrates the representation of the first two PCA components for each PCA method.

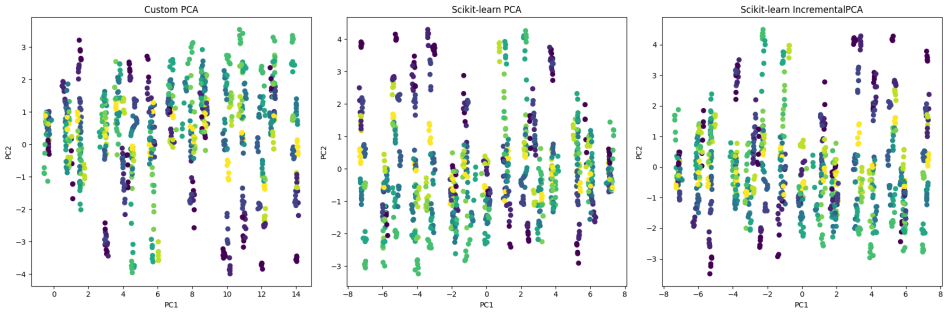


Figure 16: Comparison PCA methods

Upon scrutinizing the results, minimal distinctions are observed among the methods for this dataset.

Prior to engaging in clustering, TruncatedSVD is performed with the same number of components.

6.2 Clustering

6.2.1 K-means

Table 5: Kmeans clustering results on vowel.arff dataset

Method	Accuracy	Homogeneity	Completeness	V_measure	ARI	AMI	Silhouette	Davies
No Reduction	0.1959	0.1741	0.3049	0.2217	0.0912	0.2153	0.1796	1.7439
TSVD	0.2030	0.1834	0.3249	0.2345	0.0958	0.2282	0.2887	1.1881
PCA	0.2030	0.1834	0.3249	0.2345	0.0958	0.2282	0.2887	1.1881

Both TSVD and PCA improved the Silhouette Score significantly, suggesting better cluster separation, but the overall effectiveness of K-means on this dataset appears limited, as seen in the low accuracy and moderate scores in other metrics. The low scores across metrics indicate that K-means, even with dimensionality reduction, may not be the best fit for this dataset, possibly due to the inherent structure of the data not aligning well with K-means' assumptions.

6.2.2 BIRCH

Table 6: BIRCH clustering

		Homogeneity	Completeness	V_measure	ARI	AMI	Silhouette
No dimensionality reduction	Threshold	0.4	0.4	0.4	0.4	0.4	0.5
	Branching Factor	30	30	30	30	30	10
	SCORE	**0.013**	**0.03**	**0.018**	**0.005**	**0.012**	**0.281**
PCA	Threshold	0.8	0.8	0.8	0.8	0.8	0.5
	Branching Factor	10	10	10	50	10	10
	SCORE	**0.023**	**0.055**	**0.032**	**0.007**	**0.027**	**0.356**
TruncatedSVD	Threshold	0.8	0.8	0.8	0.8	0.8	0.2
	Branching Factor	20	20	20	20	20	10
	SCORE	**0.038**	**0.09**	**0.053**	**0.011**	**0.047**	**0.357**

The algorithm successfully identified three distinct clusters in the data for the three cases. Without dimensionality reduction, clustering shows modest performance with a high Silhouette score (0.281) suggesting relatively well-separated clusters. PCA brings slight improvements, with a higher Silhouette score (0.356), but limited agreement with the true clusters (low ARI and AMI). Notably, TruncatedSVD exhibits the best performance, achieving a Silhouette score of 0.357 and showing improved cluster separation.

6.3 Visualization

In our exploration of the Vowel dataset, characterized by its mixed nature encompassing both numerical and categorical features, the initial visualization of the original data revealed a sparse distribution without clearly defined groups over the space. This indicated a degree of complexity in the dataset where distinguishing between distinct groups was challenging.

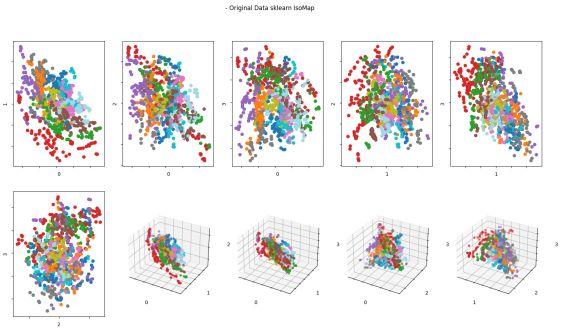
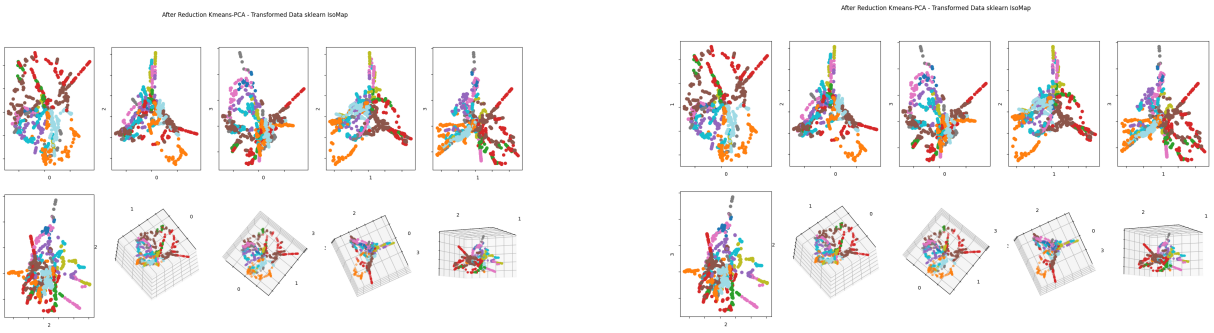


Figure 17: Original data visualization

However, after the first experiment, where k-means clustering with PCA for dimensionality reduction and Isomap for feature visualization were applied, a noticeable improvement emerged. Despite the limited distances between groups, points with similar characteristics appeared to be closer, suggesting that the combination of k-means and PCA with Isomap facilitated a more cohesive representation of the dataset. This outcome indicated that the clustering algorithm, when coupled with PCA and Isomap, contributed to revealing subtle relationships among the mixed features, enhancing the understanding of the dataset's structure.



(a) Clustering before reduction

(b) Clustering after reduction

Figure 18: vowel feature visualization with Kmeans-PCA-Isomap

Expanding on these observations, the second experiment incorporated BIRCH clustering with PCA for both dimensionality reduction and feature visualization. While the overall number of distinct groups reduced, the resulting visualization showcased clearer delineations, particularly in specific components. This refinement in the separation of groups, even with a reduced color palette, emphasized the effectiveness of BIRCH in revealing distinct patterns within the mixed dataset.

(a) Clustering before reduction

(b) Clustering after reduction

Figure 19: vowel feature visualization with Birch-TSVD-PCA

7 Conclusions

- **Is PCA giving more advice for knowing the underlying information in the dataset?**

  PCA seems to be effective in simplifying the high-dimensional data while retaining its essential variability. For each dataset, the PCA implementation, whether custom or from Sklearn, retained a significant portion of the original information (over 85% in most cases). This suggests that PCA helps in understanding the underlying structure and variability in the datasets.

- **Is BIRCH giving you more advice for knowing the underlying information in the data set?** BIRCH was used for clustering, and it showed moderate performance. The Silhouette scores suggest some challenges in achieving well-separated clusters. The effectiveness of BIRCH in revealing the underlying information might be limited, as indicated by the modest scores in homogeneity, completeness, and other metrics.

- **Can you explain the setup that you have used for PCA algorithm?** In summary, the code defines our CustomPCA class, performs PCA on a given dataset, and provides visualizations to aid in understanding the principal components and their impact on the data. The CustomPCA setup involves calculating the covariance matrix, eigenvalues, and eigenvectors of the data, selecting principal components based on the cumulative explained variance ratio (threshold set at 85%), and transforming the data accordingly.

  It works based on three methods:

  - **Init**: Initializes various parameters, including the dataset (X), dataset name (data_name), the number of principal components (k), and a threshold for cumulative explained variance (threshold).
  - **Fit**: Performs the actual PCA analysis on the dataset. It computes the mean-centered vector and the covariance matrix of the dataset. Calculates eigenvalues and eigenvectors of the covariance matrix. Adjusts the signs of the eigenvectors to ensure consistency. Sorts eigenvectors and eigenvalues by their magnitude. Determines the number of principal components (k) based on the cumulative explained variance, or uses the provided k. Projects the original data onto the selected principal components. Computes the transformed dataset and the reconstructed dataset.

  - **Visualize**: This method creates the plots of the original, reconstructed, or transformed data: Plots scatter plots or 3D scatter plots based on the specified axes. If visualizing the transformed data, it includes a scree plot showing the explained variance.

- **Can you reduce the dimensionality of the data set?** In case of an affirmative answer, **detail how do you do and how many features have been reduced from the original data set.**

  The dimensionality was successfully reduced in all datasets. For instance, in the Waveform dataset, as said before the number of components was reduced to 26 to maintain over 85% of the information. The specific number of features reduced varied across datasets as follow:

  - Waveform Dataset: Original Features is 40 reduced to 26 components, this was achieved by maintaining at least 85% of the information from the original dataset. The PCA process involved calculating the covariance matrix, eigenvectors, and eigenvalues of the dataset, and then transforming the dataset based on these components.
  - Kr-vs-Kp Dataset: Original Features is 36 reduced to 17 components, similar to the Waveform dataset, the number of components was chosen to preserve at least 85% of the original data's variance. The process involved analyzing and retaining the principal components that captured the most variance in the data.
  - Vowel Dataset: Original Features is 13 (10 numerical and 3 categorical) reduced to 4 components, the PCA here aimed to keep over 85% of the information, focusing on numerical features and one categorical feature (gender). The reduction involved a similar approach of extracting principal components that accounted for the majority of the variance in the dataset.

- **Can you detail how you decided to make the reduction using truncatedSVD have you obtained and how much have been reduced from the original data set.**

  TruncatedSVD was particularly noted for its efficiency in sparse datasets. It retained a specified number of singular values and their corresponding vectors.

- **Do you obtain the same results from your code of PCA to the code in sklearn? Explain the similarities and the differences among the two implementations.**

  The custom PCA and Sklearn's PCA show similar results in terms of information retention, suggesting that both implementations are effective. The custom PCA provides additional insights through detailed visualization.

- **Can you explain if you obtain similar clusters with the data set reduced to those obtained in the original data? In addition, have you observed a reduction in time?**

  Clustering on reduced datasets shows improvements in metrics like silhouette scores, indicating better cluster separation. The effectiveness varies across datasets, but generally, dimensionality reduction aids in revealing clearer patterns and groupings.

- **Which are the similarities and differences in the visualizations obtained using PCA and ISOMAP?**

  Visualizations using PCA and Isomap display different aspects of the data. PCA focuses on variance, while Isomap emphasizes the intrinsic geometric structure. This difference is particularly evident in the Vowel dataset, where Isomap contributes to a more cohesive representation of the dataset.