ANALYSIS OF UNSUPERVISED CLUSTERING ALGORITHMS AND IMPACT OF DIMENSIONALITY REDUCTION: A DATA DRIVEN APPROACH

Palak Narula

Adobe Inc., India

ABSTRACT

Clustering is a widely used unsupervised learning technique for discovering hidden patterns in data.however, high-dimensional datasets often pose challenges in terms of computational efficiency and clustering effectiveness. This study investigates the impact of dimensionality reduction on clustering performance by applying principal component analysis (pca),independent component analysis (ica),randomized projection, and feature agglomeration before clustering. The research utilizes k-means and expectation-maximization (em) clustering algorithms on two real-world datasets: bankruptcy prediction and breast cancer diagnosis. The study examines how different dimensionality reduction techniques influence cluster formation, computational efficiency, and interpretability. The results indicate that dimensionality reduction improves processing time and, in some cases, enhances clustering performance by removing noise and redundant features. However, certain techniques may lead to information loss, reducing cluster separability. This research provides insights into selecting appropriate dimensionality reduction methods to optimize clustering in unsupervised learning applications.

KEYWORDS

Machine Learning, Unsupervised Learning, Clustering, Dimensionality Reduction, K-means clustering, EM clustering, Principal Component Analysis, Independent Component Analysis, Randomised Projection, Feature Agglomeration

1. INTRODUCTION

Clustering is a fundamental machine learning technique used to group similar data points into clusters, enabling pattern discovery, data simplification, and various real-world applications like customer segmentation, anomaly detection, and image analysis, all without needing labeled data. However, high-dimensional data often leads to computational inefficiencies and suboptimal clustering performance due to the curse of dimensionality.

Dimensionality reduction techniques are commonly employed to address these challenges by transforming highdimensional data into a more compact representation while preserving essential information. Popular techniques include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projection, and Feature Agglomeration, each with unique properties.

The research is conducted on two real-world datasets: a bankruptcy prediction dataset and a breast cancer diagnosis dataset, both of which contain high-dimensional feature spaces that can benefit from dimensionality reduction. However, it is important to note that these datasets are used for reference purposes only—in real-world applications such as bankruptcy risk assessment and breast cancer diagnosis, accuracy is of utmost importance, and any loss of critical information

due to dimensionality reduction may lead to incorrect conclusions. While dimensionality reduction can improve computational efficiency and clustering speed, it must be applied with caution in high-stakes domains where data integrity is crucial. The primary objectives of this study are:

- To evaluate how different dimensionality reduction techniques affect clustering performance in terms of cluster formation, separability, and computational efficiency.

- To compare K-Means and EM clustering algorithms in reduced feature spaces to assess their robustness across different datasets.
- To determine the trade-offs between accuracy and processing time, offering insights into selecting the most suitable dimensionality reduction method for clustering tasks

2. RELATED WORK

Clustering has been widely studied in machine learning and data mining, particularly for applications in financial risk assessment and medical diagnostics. Traditional clustering algorithms such as K-Means[1] and Expectation- Maximization (EM) [2] have been extensively used for identifying patterns in large datasets. In bankruptcy prediction, clustering techniques have been applied to financial ratios to classify firms based on insolvency risk [3]. Similarly, in medical research, clustering methods have been used to analyze patient data and classify cancer subtypes [4].

Dimensionality reduction plays a crucial role in improving clustering performance by reducing computational complexity and mitigating the curse of dimensionality [5]. Principal Component Analysis (PCA) [6] is one of the most commonly used techniques for feature reduction by transforming data into orthogonal components. Independent Component Analysis (ICA) [7] has been used in biomedical signal processing and financial modeling to separate independent sources from mixed data. More recent approaches, such as Randomized Projection [8], provide computationally efficient alternatives by approximating feature space transformations, while Feature Agglomeration groups correlated features to improve interpretability.

Several studies have investigated the impact of dimensionality reduction on clustering. Van der Maaten & Hinton [9] demonstrated how t-SNE improves cluster visualization in high-dimensional datasets. Ding & He [10] analyzed the relationship between PCA and K-Means clustering, showing that PCA can improve cluster separation by removing noise. However, other studies have highlighted the trade-offs, noting that excessive dimensionality reduction can remove key distinguishing features, leading to poor cluster formation.

Unlike previous research that applies clustering and dimensionality reduction separately, this study systematically evaluates the impact of multiple dimensionality reduction techniques (PCA, ICA, Randomized Projection, and Feature Agglomeration) on clustering performance using K-Means and EM algorithms. By analyzing two distinct datasets (bankruptcy prediction and breast cancer diagnosis), this research provides insights into the trade-offs between computational efficiency and clustering effectiveness, contributing to the broader understanding of unsupervised learning workflows.

3. ABOUT THE DATA

3.1. Company Bankruptcy Prediction

Bankruptcy data is a dataset to be used for the experiments. This dataset is derived from Kaggle [11] and includes features defining the financial status of the company. There are multiple features in the dataset like operating gross margin, cash rate, equity to liability ratio etc. The historical data also contains labels specifying whether the company was declared bankrupt or not. This data corresponds to a binary classification problem where the financial features of the company are available, and the problem is to predict whether the company is heading towards bankruptcy or not. To use this dataset for experiments with unsupervised learning and dimensionality reduction algorithms, only the training data without labels would be used.

3.2. Breast Cancer Classification

This is a classification problem derived from Kaggle [12] which includes historical data representing various features of the breast like area mean, texture mean, radius mean etc. The data includes label malignant (cancerous) or benign (non-cancerous) based on the diagnosis of the breast tissue. The problem involves predicting the diagnosis of the breast tissue given the measurable features. Again, this dataset will be used without the final labels for running experiments for unsupervised learning and dimensionality reduction algorithms

4. UNSUPERVISED CLUSTERING OF DATA

In this paper, we mainly focus on two common clustering techniques – K-means clustering and EM clustering

4.1. K- Means Clustering

It is a type of clustering technique [13] where the data point are classified based on their distance from the cluster centres. It is a hard clustering technique where each data point belongs to only one cluster and there is no sharing of data points. Applying the k means clustering to each of the datasets and analysing the silhouette score for different number of clusters would help deduce the value of k (most efficient value for number of clusters)

A silhouette score determines the quality of the cluster. It varies between [-1,1] where the best value 1 indicates that the datapoint are compact within the cluster where they belong and are far away from the other clusters so analysing the silhouette score helps create a clear picture of the dataset.

4.1.1. Bankruptcy Data

Applying K-means clustering to bankruptcy data and plotting a graph between the number of clusters and the silhouette score to get the optimal number of clusters-

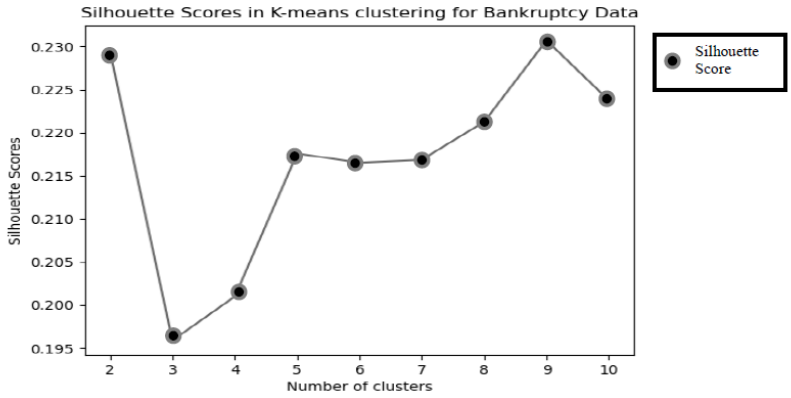

Figure 1. Silhouette score vs clusters (Bankruptcy data)

The silhouette scores are maximum for 2 and 9 number of clusters. After achieving the maximum score, there seems to be a steep decrease because the data can't be divided efficiently in higher number of clusters.

Using number of clusters = 2 and applying the K means algorithm to bankruptcy data, a score of 0.2294 is achieved which indicates that though most of the datapoint are in their cluster but the clusters are overlapping.

To visualize the clusters generated, plotting a biplot using two important features based on domain knowledge (cash flow rate and operating expense rate) with each datapoint marked based on its cluster.

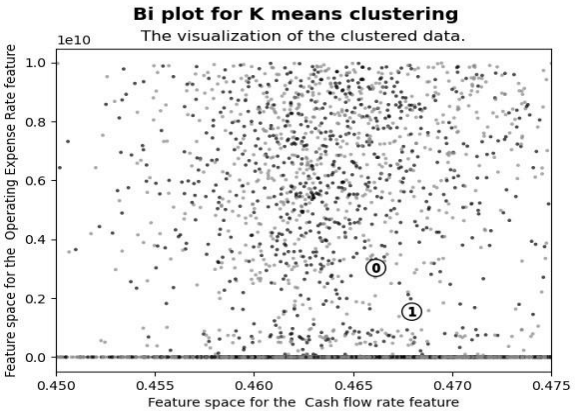

Figure 2. Bi plot for K means clustering (Bankruptcy data)

The biplot graph clearly shows overlapped clusters as indicated by the silhouette score.

4.1.2. Breast Cancer Data

Applying K-means clustering on Breast Cancer dataset and plotting a graph for analysis-

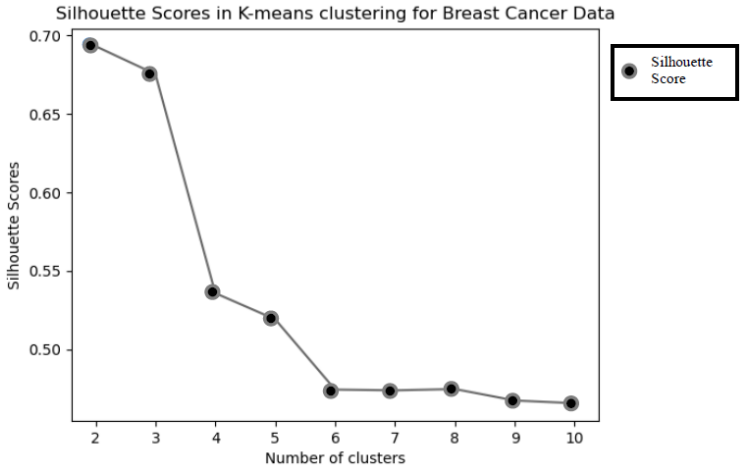

Figure 3. Silhouette score vs clusters (Breast cancer data)

The graph clearly shows that the algorithm is able to best classify the data in two clusters. As the number of clusters increases, the data-points start overlapping between the clusters with a drastic decrease in the silhouette score.

Choosing number of clusters = 2 and applying the algorithm on the train data of the dataset, the silhouette score achieved = 0.693 which indicates a fairly decent performance of the algorithm with the data.

The following biplot between area mean and smoothness mean features (features are selected based on domain knowledge) help visualise the K means clustering with breast cancer data-

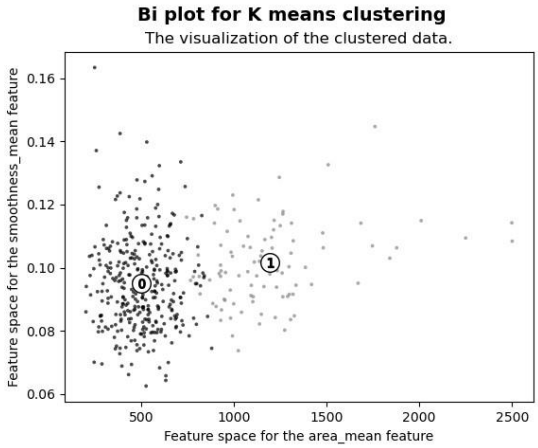

Figure 4. Bi plot for K means clustering (Breast cancer data)

As indicated by the silhouette score, the clusters are compact and there is not much overlapping between the clusters.

4.2. Expectation Maximisation

It is a type of technique [14] that performs maximum likelihood expectation on each data point and based on the probability of a datapoint to belong to a cluster, it is assigned the label. This

technique is a type of soft clustering where a data point may belong to more than one cluster with some probability

4.2.1. Bankruptcy Data

Applying the expectation maximisation technique to bankruptcy data and plotting a graph between the silhouette score and number of clusters-

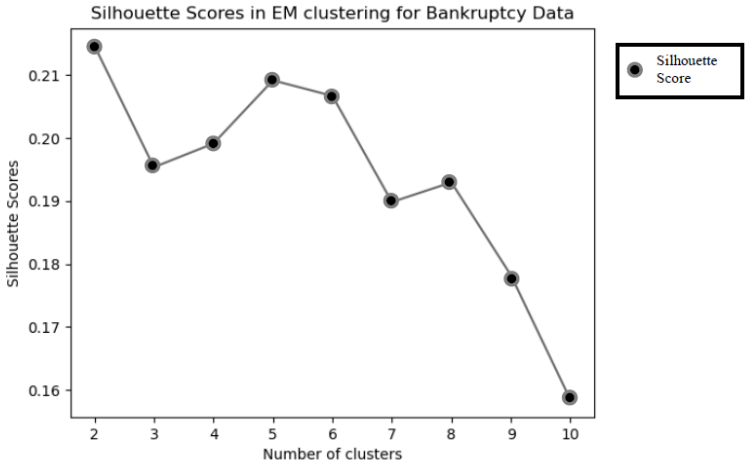

Figure 5. Silhouette score vs clusters (Bankruptcy data)

Similar to K means clustering, the EM algorithm achieves the highest silhouette score for number of clusters = 2. The silhouette score while applying EM with 2 clusters on the data is 0.2146 which again indicates that though most of the data points belong to its own cluster, there is a high overlapping between the clusters. Visualising the clusters by creating a biplot for the same features chosen for K means clustering-

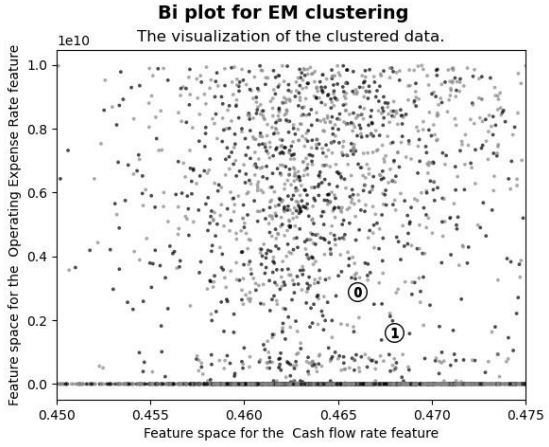

Figure 6. Bi plot for EM clustering (Bankruptcy data)

The plot looks very similar to K means clustering with high overlapping of clusters as expected from the silhouette score.

4.2.2. Breast Cancer Data

Applying EM algorithm to the breast cancer data and plotting a graph between silhouette score and number of clusters-

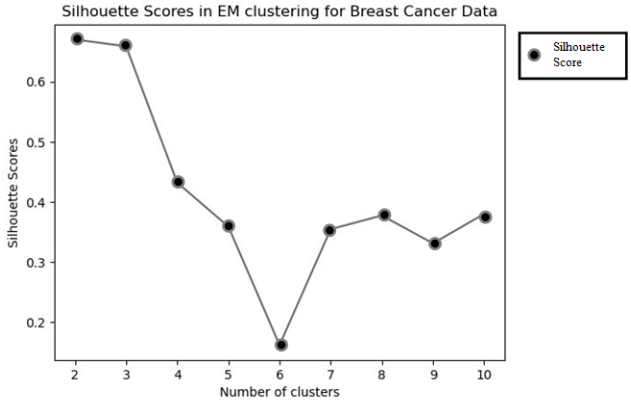

Figure 7. Silhouette score vs clusters (Breast cancer data)

The graph indicates that 2/3 clusters is the most suitable configuration for breast cancer data with EM clustering. Choosing number of clusters = 2, a silhouette score of 0.6704 is achieved which indicates that the clusters have low overlapping and the datapoint are well bound to their clusters. Visualising the clustering with a biplot between area mean and smoothness mean features (features are selected based on domain knowledge) of the data-

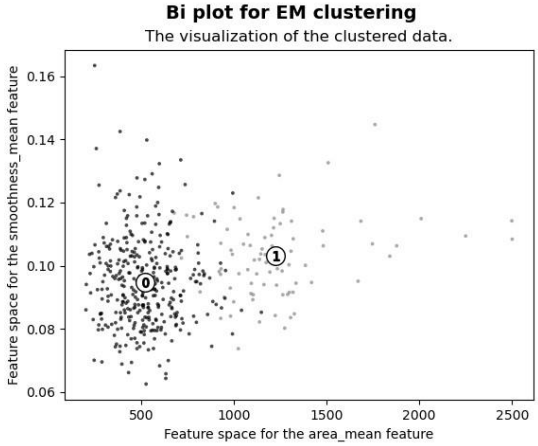

Figure 8. Bi plot for EM clustering (Breast cancer data)

As indicated by the silhouette score, the clusters created by EM clustering for Breast cancer data are compact and have low overlapping.

5. DIMENSIONALITY REDUCTION OF DATA

Dimensionality Reduction algorithms are a type of unsupervised learning algorithms that transforms the dataset from a high dimension to a low dimension in order to reduce the processing time and computation power of algorithms being applied on the dataset. In this paper

we analyse the following dimensionality reduction techniques by applying these techniques on the Bankruptcy and Breast Cancer Dataset-

- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)
- Randomised Projection
- Feature Agglomeration

5.1. Principal Component Analysis

PCA [15] is a dimensionality reduction technique that works by finding the correlation between the features of the dataset. It works by finding the axis with the highest variance in the data and then projects the dataset onto this axis.

5.1.1. Bankruptcy Data

Applying PCA dimensionality reduction to Bankruptcy Dataset and plotting a graph between the cumulative explained variance ratio (ratio of Eigen value and the total sum of Eigen values) and number of features to deduce the optimal number of features post reduction-

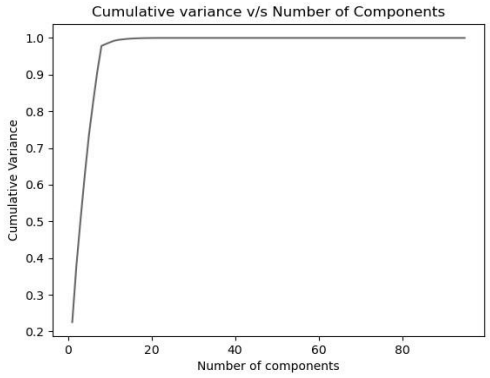

Figure 9. Cumulative variance v/s number of components (Bankruptcy data)

The above graph shows an increase in the cumulative variance with the increase in the number of features for lower values but once the highest variance is achieved there is no improvement with increase in number of features indicating that the features can be reduced to 25 (where highest variance is achieved) with PCA algorithm.

5.1.2. Breast Cancer Data

Applying PCA dimensionality reduction technique to breast cancer data and plotting a graph between cumulative variance and number of components –

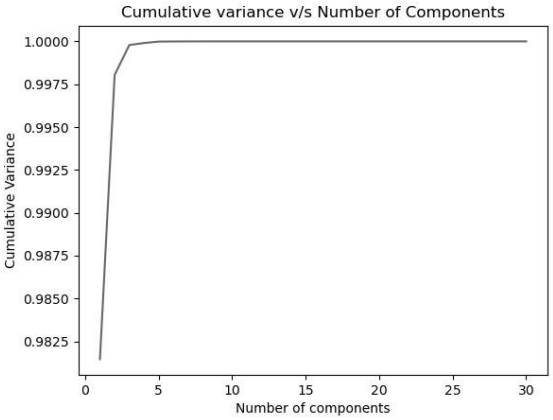

Figure 10. Cumulative variance v/s number of components (Breast cancer data)

The nature of the plot is similar to the plot for Bankruptcy data with a monotonic increase in variance upto a particular value and then the value remains constant. Choosing number of components = 3 as the optimal value to represent the dataset.

5.2. Independent Component Analysis

Independent component analysis [16] is a dimensionality reduction technique that works by identifying how independent the features are from each other and then reduces the features by projecting the features such that the independence of the features is maximised. Kurtosis is a measure of combined weight of the distribution of the tail with respect to the centre. The optimal number of components using ICA can be found by maximising the non-gaussianity (kurtosis) of the given dataset.

5.2.1. Bankruptcy Data

Applying ICA dimensionality reduction to Bankruptcy dataset and plotting a graph between kurtosis and the number of components-

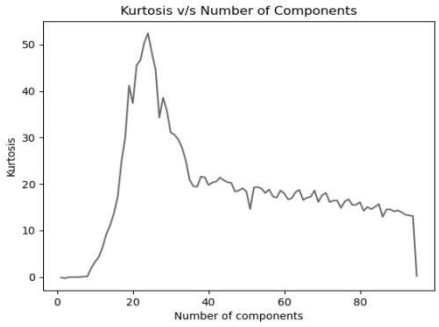

Figure 11. Kurtosis variance v/s number of components (Bankruptcy data)

The graph shows that the maximum Kurtosis is achieved for ‘number of components’ = 24 so ICA should reduce the dimensions to 24 features to achieve the most independent features that contain the entire information.

5.2.2. Breast Cancer Data

Applying ICA to breast cancer data and plotting a graph between Kurtosis and number of components-

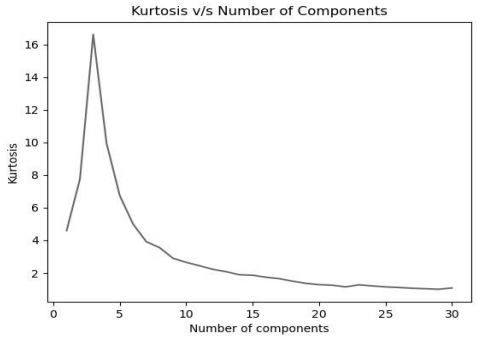

Figure 12. Kurtosis variance v/s number of components (Breast cancer data)

The nature of the graph for Breast cancer data is similar to the previous dataset with increase in Kurtosis with increasing number of components and once the peak is achieved, the value of Kurtosis starts decreasing. This is because once the peak is achieved increasing the number of features would increase the inter-dependency. Choosing number of components = 3 for this dataset with ICA Dimensionality Reduction.

5.3. Randomised Projection

Randomised Projection [17] is a dimensionality reduction technique that works by projecting data to a lower dimensional subspace by using a random matrix whose columns have unit length.

One way to deduce the optimal number of components with randomised Projection algorithm is to inspect reconstruction error. The point where the reconstruction error is minimum should reflect to the minimum error when the original data is reconstructed.

5.3.1. Bankruptcy Data

Applying randomised projection algorithm to bankruptcy data and plotting a graph between reconstruction error and number of components-

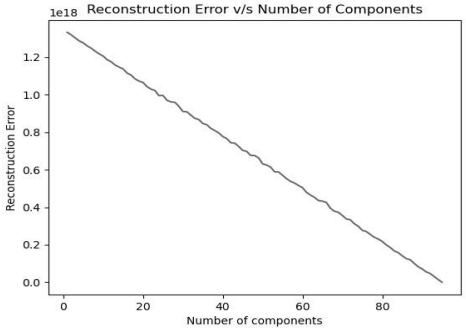

Figure 12. Reconstruction error v/s number of components (Bankruptcy data)

The graph shows a continuous decrease in the error with the increasing number of components. This is an expected behaviour as the original data can be best reconstructed when all the available features are present. Choosing number of components = 1 to understand the effect when all the points are randomly projected to a single dimension.

5.3.2. Breast Cancer Data

Applying Randomised Projection technique to Breast cancer dataset and again plotting a graph between reconstruction error and number of components.

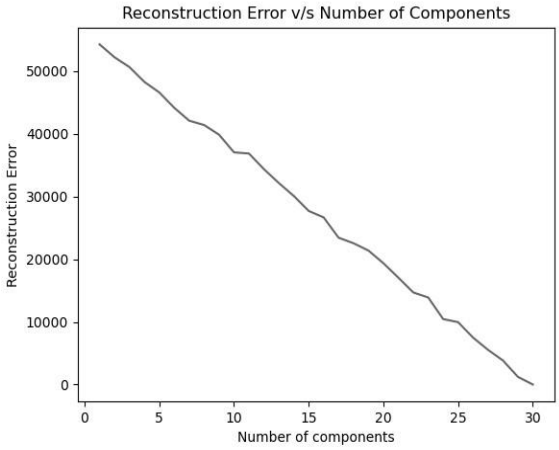

Figure 13. Reconstruction error v/s number of components (Breast Cancer data)

The nature of the curve remains same as that generated for the previous dataset, that is, there is a continuous decrease in the reconstruction error until all the available features are used and the dataset can be reconstructed with minimum error.

5.4. Feature Agglomeration

Feature agglomeration [18] is a dimensionality reduction technique where the features are grouped into clusters based on their similarity and then the features in a cluster are recursively merged in order to reduce the number of features.

Similar to ICA, the optimal number of features post dimensionality reduction can be derived by analysing the Kurtosis value. The higher the Kurtosis value, the more independent are the merged features from each other.

5.4.1. Bankruptcy Data

Applying Feature Agglomeration technique on the bankruptcy data and plotting a graph between Kurtosis and the number of components-

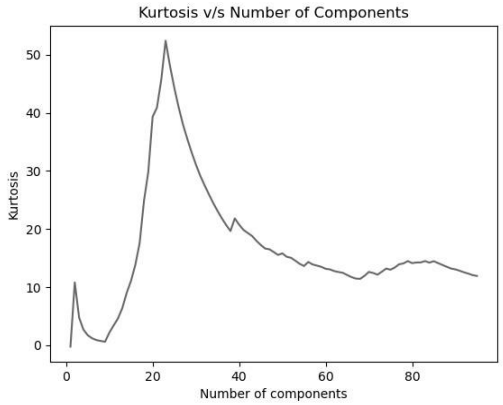

Figure 14. Kurtosis v/s number of components (Bankruptcy data)

The graph achieves the peak at number of components = 23 which is the optimal value when reducing the dimensions of bankruptcy data using feature agglomeration. The nature of the graph indicates that independence of data increases with increase in number of components until the peak is achieved and then the value decreases continuously upto a certain value.

5.4.2. Breast Cancer Data

Applying feature agglomeration reduction algorithm to breast cancer data-

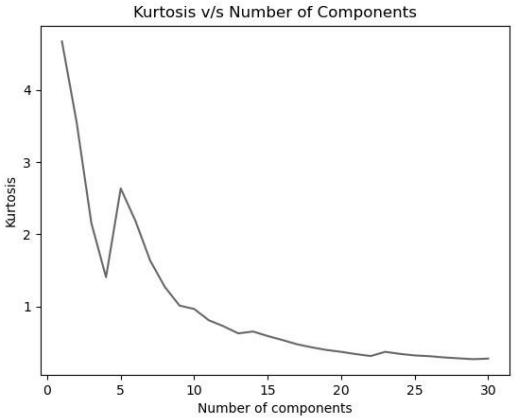

Figure 15. Kurtosis v/s number of components (Breast Cancer data)

The above graph shows that the highest Kurtosis value is achieved for number of components = 1 and then there is an overall decrease in the value with a small increase in value at 6 but the overall nature remains decreasing.

6. CLUSTERING ALGORITHMS ON REDUCED DATASET

For the experiments in this section, using the reduced dataset with the optimal number of components derived in the previous section and applying clustering algorithms with the optimal number of clusters for each dataset derived earlier and analysing the results.

The silhouette score that measures the quality of the clusters is the best measure to compare how the clustering algorithms work with the original dataset and the dataset derived after applying the dimensionality reduction algorithms.

6.1. Bankruptcy Data

Applying K means and EM clustering algorithms with 2 clusters (number of clusters based on the clustering experiment) on the reduced bankruptcy dataset. Using the following number of reduced features for this experiment –

- PCA - 25
- ICA - 24
- Randomised Projection - 1
- Feature Agglomeration – 23

Table 1. Comparison of dimensionality reduction algorithms with K-means clustering (Bankruptcy data)

	Silhouette Score	**Davies-Bouldin Index**	**Calinski-Harabasz Score**	**Computation Time**
No reduction	0.2294	1.9231	1143.6301	0.48s
PCA	0.2294	1.9231	1143.6301	0.45s
ICA	0.7568	0.5704	205.1123	0.43s
Randomised Projection	0.5945	0.4631	12960.5986	0.41s
Feature Agglomeration	0.2294	1.9231	1143.6371	0.42s

Table 2. Comparison of dimensionality reduction algorithms with EM clustering (Bankruptcy data)

	Silhouette Score	**Davies-Bouldin Index**	**Calinski-Harabasz Score**	**Computation Time**
No reduction	0.2146	1.9653	1087.8286	0.52s
PCA	0.2277	1.9071	1107.5265	0.43s
ICA	0.1193	3.1879	189.2792	0.40s
Randomised Projection	0.5974	0.4321	12128.9256	0.39s
Feature Agglomeration	0.2278	1.9093	1106.6271	0.42s

6.2. Breast Cancer Data

Applying K means and EM clustering algorithms with 2 clusters (number of clusters based on the clustering experiment) on the reduced breast cancer dataset. Using the following number of reduced features for this experiment –

- PCA - 3
- ICA - 3
- Randomised Projection - 1
- Feature Agglomeration – 1

Table 3. Comparison of dimensionality reduction algorithms with K-means clustering (Breast cancer data)

	Silhouette Score	**Davies-Bouldin Index**	**Calinski-Harabasz Score**	**Computation Time**
No reduction	0.6930	0.5175	905.6937	0.024s
PCA	0.6934	0.5171	906.3006	0.015s
ICA	0.8752	0.5341	135.6421	0.018s
Randomised Projection	0.7086	0.4634	967.0403	0.017s
Feature Agglomeration	0.7042	0.4821	955.6881	0.017s

Table 4. Comparison of dimensionality reduction algorithms with EM clustering (Breast cancer data)

	Silhouette Score	**Davies-Bouldin Index**	**Calinski-Harabasz Score**	**Computation Time**
No reduction	0.6705	0.5325	794.2776	0.015s
PCA	0.6911	0.4942	798.7168	0.011s
ICA	0.5533	1.3636	121.1993	0.009s
Randomised Projection	0.7086	0.4374	873.3010	0.007s
Feature Agglomeration	0.6974	0.4616	821.5227	0.008s

7. SUMMARY

This study examines the impact of dimensionality reduction on clustering performance using PCA, ICA, Randomized Projection, and Feature Agglomeration in combination with K-Means and Expectation-Maximization (EM) clustering algorithms. The research is conducted on two real-world datasets: bankruptcy prediction and breast cancer diagnosis, both of which involve high-dimensional data where feature reduction techniques can enhance computational efficiency and clustering effectiveness.

The findings demonstrate that dimensionality reduction significantly improves processing time, making clustering more efficient, particularly for large datasets. PCA and Feature Agglomeration help remove noise and redundant features, leading to more defined clusters, whereas Randomized Projection and ICA show mixed results depending on dataset characteristics. While dimensionality reduction often enhances cluster formation, some techniques may introduce information loss, affecting cluster separability and overall clustering performance.

It is important to emphasize that the chosen datasets are used for reference purposes only—in real-world applications such as bankruptcy prediction and breast cancer diagnosis, where accurate decision-making is critical, dimensionality reduction must be applied with caution. While it enhances computation speed and efficiency, any loss of essential data could lead to inaccurate risk assessments or misdiagnoses. Therefore, the trade-off between processing efficiency and data integrity must be carefully evaluated before implementing dimensionality reduction in high-stakes applications.

A key takeaway from this study is that no single dimensionality reduction technique is universally superior, and the choice of method depends on the dataset, clustering algorithm, and the trade-off between computational efficiency and clustering effectiveness. This research highlights the

importance of carefully selecting dimensionality reduction techniques based on dataset properties and clustering objectives.

By systematically comparing multiple dimensionality reduction approaches, this study provides valuable insights for optimizing clustering workflows in machine learning. The results are particularly relevant for applications in finance and healthcare, where high-dimensional data is common, and efficient, interpretable clustering is essential for decisionmaking. Future research could extend this analysis to other clustering methods, hybrid feature reduction approaches, and unstructured data, further refining best practices for dimensionality reduction in unsupervised learning.

8. FUTURE SCOPE

While this study provides insights into the impact of dimensionality reduction on clustering performance, several avenues for future research remain:

Exploring Additional Clustering Algorithms – This study focuses on K-Means and Expectation-Maximization (EM) clustering. Future work could explore the impact of dimensionality reduction on other clustering techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Agglomerative Clustering, and Spectral Clustering to assess their suitability for high-dimensional data.

Hybrid Dimensionality Reduction Approaches – Instead of applying a single method, future studies could evaluate the effectiveness of hybrid approaches such as PCA followed by Feature Agglomeration or ICA combined with Randomized Projection. These combinations might balance the trade-offs between variance preservation, computational efficiency, and clustering performance.

Application to Unstructured and High-Dimensional Data – The datasets used in this research are structured and tabular. Future work could extend the analysis to unstructured data, such as text, images, and time-series data, where dimensionality reduction techniques like word embeddings (e.g., Word2Vec, BERT), autoencoders, and t-SNE may significantly impact clustering performance.

Scalability and Performance Optimization – As datasets grow larger, the efficiency of dimensionality reduction techniques becomes crucial. Future studies could analyze how these methods scale in big data environments using distributed computing frameworks such as Apache Spark or GPU-accelerated implementations to optimize performance in real-world applications.

Impact on Cluster Quality Metrics – This study primarily evaluates clustering qualitatively. Future research could use a broader set of clustering validation metrics, such as Silhouette Score, Davies-Bouldin Index, Adjusted Rand Index, and Mutual Information Score, to quantitatively assess the effects of different dimensionality reduction methods on clustering quality.

Robustness in Noisy and Imbalanced Datasets – Real-world datasets often contain noise, outliers, and class imbalances that can affect clustering results. Future studies could investigate how well different dimensionality reduction methods handle noisy or imbalanced data and whether pre-processing techniques, such as outlier removal or synthetic data augmentation, improve clustering performance.

Domain-Specific Optimization and Interpretability – The choice of dimensionality reduction techniques may vary depending on the application domain. Future research could explore how

these techniques impact clustering in specific fields such as healthcare, finance, cybersecurity, and bioinformatics. Additionally, integrating explainable AI (XAI) methods could improve the interpretability of cluster assignments, making results more actionable for domain experts.

By addressing these areas, future research can enhance the practical application of dimensionality reduction in clustering, optimizing trade-offs between computational efficiency, clustering accuracy, and interpretability across different machine learning domains.

REFERENCES

[1] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

[2] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological).

[3] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance.

[4] Golub, T. R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science.

[5] Bellman, R. (1957). Dynamic Programming. Princeton University Press.

[6] Jolliffe, I. T. (2002). Principal Component Analysis. Springer.

[7] Hyvärinen, A.,& Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Networks.

[8] Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

[9] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research.

[10] Ding, C., & He, X. (2004). K-Means clustering via principal component analysis. Proceedings of the 21st International Conference on Machine Learning (ICML).

[11] Company Bankruptcy Prediction --- kaggle.com.https://www.kaggle.com/fedesoriano/company-bankruptcyprediction,

[12] Breast Cancer Dataset Classification --- kaggle.com.https://www.kaggle.com/dhainjeamita/breast-cancerdataset-classification/data

[13] KMeansscikit-learn.org.https://scikit learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[14] GaussianMixture---scikit-learn.org.https://scikitlearn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

[15] PCA---scikit-learn.org.https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[16] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html

[17] Random Projection --- scikit-learn.org.https://scikit-learn.org/stable/modules/random_projection.html

[18] FeatureAgglomeration---scikit-learn.org.https://scikitlearn.org/stable/modules/generated/sklearn.cluster.FeatureAgglomeration.html

AUTHOR

Palak Narula gained her BTech in Computer Science from H.B.T.I, India and MS in Computer Science (major in Machine Learning) from Georgia Institute of Technology, Atlanta, GA. Currently she is working as a Computer Scientist 2 in Adobe, India.