Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 641-654
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297

Enhancing Clustering Performance through Benchmarking of Dimensionality Reduction Techniques on Educational Data

Eko Priyanto*[1], Berlilana[2], Imam Tahyudin[3]

[1,2,3]Magister of Computer Science, Universitas Amikom Purwokerto, Indonesia

Email: [1]ekoklirong2@gmail.com

Received : Jan 8, 2025; Revised : Jan 29, 2025; Accepted : Feb 1, 2025; Published : Apr 26, 2025

Abstract

This study evaluates the effectiveness of dimensionality reduction techniques in enhancing clustering performance using a tracer study dataset of 500 alumni from UMNU Kebumen, containing 58 variables. The objective was to identify the optimal combination of dimensionality reduction and clustering methods for uncovering patterns in alumni profiles, job search strategies, and employment outcomes. Principal Component Analysis (PCA), Non-Negative Matrix Factorization (NMF), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) were applied, followed by clustering using K-Means, DBSCAN, and Hierarchical Clustering. The findings revealed that NMF achieved the highest clustering quality, particularly with K-Means and Hierarchical Clustering, outperforming PCA. NMF also demonstrated superior compactness with a Calinski-Harabasz Index of 287.96, compared to 125.88 for PCA. While t-SNE and UMAP delivered competitive results, their computational times of 245.8 and 76.5 seconds, respectively, made them less practical for large datasets. The novelty of this study lies in its comprehensive evaluation of dimensionality reduction techniques and the integration of diverse clustering algorithms to assess their interplay. The results provide actionable insights, recommending NMF for accuracy-critical tasks and PCA for time-sensitive applications. Given the increasing volume of high-dimensional educational data, this study highlights the critical need for efficient clustering strategies to extract meaningful insights, ultimately supporting data-driven decision-making in education and workforce planning. Addressing these challenges is essential to optimizing institutional strategies, improving student employability, and enhancing workforce alignment with industry demands.

Keywords : *Clustering, Dimensionality Reduction, K-Means, Non-Negative Matrix Factorization, Principal Component Analysis*

1. INTRODUCTION

The rapid growth of data in educational institutions has created a pressing need for advanced techniques to analyze and extract meaningful patterns from high-dimensional datasets [1]. Alumni tracer studies, which collect detailed information on graduates' employment outcomes, career trajectories, and job search strategies, are a valuable resource for evaluating the effectiveness of educational programs and aligning them with labor market demands [2, 3]. However, the complexity and high dimensionality of such datasets pose significant challenges, often obscuring critical insights [4]. Dimensionality reduction and clustering techniques offer a solution to this problem by simplifying data structures while preserving essential patterns [5].

Dimensionality reduction methods, such as Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF), have proven effective in reducing data complexity while maintaining interpretability [6, 7]. Meanwhile, non-linear approaches like t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are increasingly popular for uncovering non-linear relationships within datasets [8, 9]. These methods are often paired with clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering to

group similar data points, enabling educators and policymakers to identify trends and make informed decisions [10]. Despite their widespread use, there remains a gap in understanding the interplay between dimensionality reduction and clustering techniques in educational datasets, particularly in the context of tracer studies [11, 12, 13, 14, 15].

Prior research has focused on either dimensionality reduction or clustering techniques independently, but few studies have comprehensively evaluated their combined effectiveness in educational datasets [16, 17]. This study builds upon existing research by systematically comparing multiple dimensionality reduction techniques alongside clustering methods, offering a holistic perspective on their synergies and limitations [18, 19]. Unlike previous studies that primarily examine commercial datasets or student performance metrics, this study leverages a uniquely structured alumni tracer dataset from UMNU Kebumen, comprising 500 respondents and 58 distinct variables covering demographic information, employment outcomes, job search strategies, and career progression [20, 21]. The dataset's high dimensionality and categorical diversity present an ideal testbed for evaluating the effectiveness of dimensionality reduction and clustering techniques [22].

This study aims to evaluate the effectiveness of various dimensionality reduction techniques in enhancing clustering performance using a tracer study dataset of alumni from UMNU Kebumen. By systematically comparing PCA, NMF, t-SNE, and UMAP alongside clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering, this research seeks to identify the optimal combination of methods for extracting meaningful insights [23, 24]. The novelty of this work lies in its comprehensive evaluation of these techniques, bridging the gap in understanding their interdependencies. The results provide actionable recommendations for selecting appropriate methods based on computational efficiency, clustering quality, and the complexity of the dataset [25].

2. LITERATURE REVIEW

2.1. Dimensionality Reduction Techniques

Dimensionality reduction is a crucial preprocessing step in data analysis, especially for high-dimensional datasets where redundancy and noise can obscure meaningful patterns. Principal Component Analysis (PCA) is one of the most widely used techniques, transforming correlated variables into uncorrelated principal components that capture the maximum variance in the data [9]. PCA has been extensively applied in education and employment studies to reduce data complexity while maintaining interpretability [10]. However, its linear nature limits its ability to capture non-linear structures in the data.

Non-Negative Matrix Factorization (NMF) is an alternative technique that is particularly effective for datasets with non-negative values. NMF decomposes the data into additive components, preserving interpretability and uncovering latent factors. Studies have shown the efficacy of NMF in educational and sociological datasets, particularly when non-negativity is a critical constraint [11]. Non-linear approaches, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), have gained popularity for their ability to uncover complex patterns. t-SNE minimizes the Kullback-Leibler divergence between pairwise similarities in high-dimensional and low-dimensional spaces, enabling detailed cluster exploration, albeit at a high computational cost [12]. UMAP, leveraging graph-based optimization, offers faster computation while maintaining or surpassing t-SNE's performance in many cases [13].

The selection of a dimensionality reduction technique depends on the specific dataset characteristics and analysis goals. PCA and NMF are preferred for their computational efficiency and interpretability, while t-SNE and UMAP are suited for exploring non-linear relationships and visualizing clusters [14].

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297

2.2. Clustering Algorithms

Clustering techniques are widely used in educational research to group similar data points, such as student performance or alumni employment patterns. K-Means is one of the most commonly applied clustering algorithms due to its simplicity and efficiency. It partitions the dataset into k clusters by minimizing the within-cluster sum of squares, making it suitable for datasets with compact, spherical clusters [15]. However, K-Means has limitations, such as sensitivity to initial centroid placement and the requirement to specify the number of clusters in advance, which can affect exploratory analyses.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) addresses some of these limitations by identifying clusters based on density. This method is particularly effective for datasets with irregular cluster shapes and noise, making it robust for real-world applications [16]. Nevertheless, its performance depends heavily on parameter tuning, specifically the neighborhood radius (Eps) and the minimum number of points ($MinPts$).

Hierarchical Clustering is another popular approach, offering flexibility through a dendrogram that represents the hierarchical relationships among data points. This method allows researchers to explore clustering solutions at different levels of granularity, making it especially useful for datasets where the optimal number of clusters is not known beforehand [17]. However, Hierarchical Clustering can become computationally expensive as dataset size increases.

2.3. Integration of Dimensionality Reduction and Clustering

The combination of dimensionality reduction and clustering techniques has been extensively studied, with evidence suggesting that reducing dimensionality often enhances clustering performance. PCA is frequently paired with K-Means to improve cluster separability by reducing redundancy and noise while preserving linear relationships [9]. NMF, with its capacity to handle non-negative data, has demonstrated strong compatibility with K-Means and Hierarchical Clustering in educational datasets, where preserving additive structures is important [11]. Non-linear dimensionality reduction methods, such as t-SNE and UMAP, have shown superior performance in clustering datasets with complex patterns, particularly when paired with density-based algorithms like DBSCAN or with K-Means for exploratory analyses [13, 14].

Despite these advancements, there remains a gap in understanding how the interplay between dimensionality reduction and clustering techniques can be optimized for specific contexts, such as alumni tracer studies. This study bridges this gap by systematically evaluating multiple dimensionality reduction and clustering techniques, providing actionable insights into their effectiveness and suitability for high-dimensional educational datasets.

3. METHOD

This study used a dataset from a tracer study of alumni at UMNU Kebumen, consisting of 58 variables capturing alumni profiles, employment status, job search methods, and the alignment between education and employment. The methodology includes data preprocessing, dimensionality reduction, clustering, and evaluation. The methodological workflow of this study, illustrated in Figure 1, encompasses data collection, preprocessing, dimensionality reduction (PCA, NMF, t-SNE, UMAP), clustering analysis (K-Means, DBSCAN, Hierarchical Clustering), and performance evaluation using Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 641-654
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297

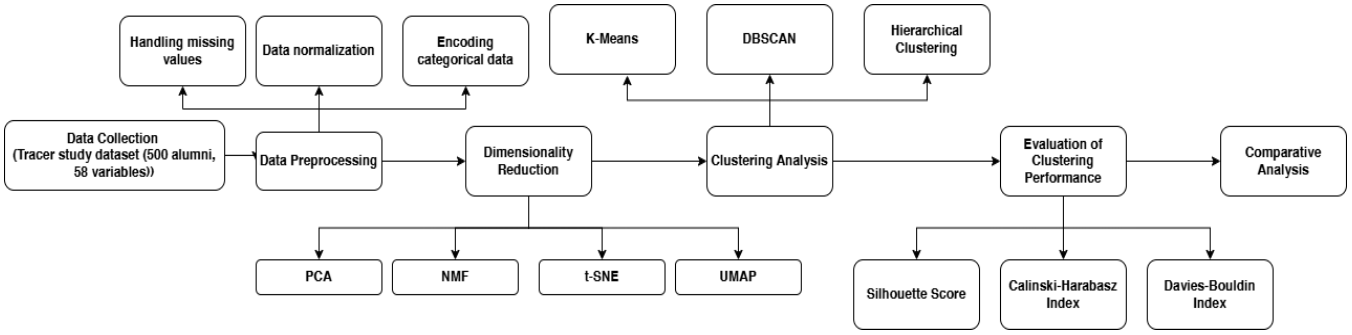


Figure 1. Workflow of Dimensionality Reduction and Clustering Analysis

3.1. Dataset Description

The dataset contains a mix of categorical and numerical variables. Categorical variables include employment status and job search methods, while numerical variables include the number of job applications submitted and the time taken to secure employment. Preprocessing addressed missing values, non-uniform scales, and encoding requirements to ensure data suitability for clustering and analysis [17].

3.2. Data Preprocessing

Data preprocessing involved handling missing values, normalizing numerical variables, and encoding categorical data. Missing numerical values were imputed using the median (1) [18].

$$x_{imputed} = median(X_{column}) \quad (1)$$

Categorical variables were imputed using the mode. Numerical variables were normalized using Min-Max scaling (2):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

where x' is the normalized value, x is the original value, and $\min(x)$ and $\max(x)$ are the minimum and maximum values, respectively. This normalization ensures a consistent scale across features, which is critical for clustering performance [19]. Categorical variables were label-encoded for clustering algorithms (3):

$$Label(C) = \{C1: 0, C2: 1, \ldots, Cn: n-1\} \quad (3)$$

For dimensionality reduction, one-hot encoding was applied to maintain independence between categorical variables [20].

3.3. Dimensionality Reduction

Dimensionality reduction simplified the dataset while retaining essential structures and patterns. Four methods were evaluated:

Principal Component Analysis (PCA): PCA was chosen due to its ability to transform the dataset by identifying principal components that explain the majority of variance. It computed the covariance matrix as (4):

$$\mathrm{C} = \mathrm{X.T@X} \quad (4)$$

C is the covariance matrix and X is the normalized data matrix. Eigenvectors corresponding to the largest eigenvalues were selected as principal components [21]. PCA was selected because it

provides a linear transformation that maintains the global structure of the data while reducing dimensionality, making it computationally efficient for large datasets.

Non-Negative Matrix Factorization (NMF): NMF decomposed the data matrix X into two non-negative matrices W and H, where $X \approx W \cdot HX \approx W \cdot HX \approx W \cdot H$. The optimization minimized the reconstruction error using the Frobenius norm (5) [22].

$$argmin(W,H)||X - W@H||_F^2 \quad (5)$$

NMF was chosen for its ability to provide parts-based representation, which enhances interpretability, particularly in educational datasets where feature significance is crucial. Unlike PCA, which allows negative values, NMF ensures non-negativity, making it suitable for datasets where negative values lack meaningful interpretation.

t-SNE (t-Distributed Stochastic Neighbor Embedding): t-SNE minimized the Kullback-Leibler divergence between the high-dimensional and low-dimensional pairwise similarities (6):

$$KL\,(P||Q) = \Sigma\Sigma P_{ij} * \log\left(\frac{P_{ij}}{Q_{ij}}\right) \quad (6)$$

P_{ij} and Q_{ij} are the pairwise similarities in high-dimensional and low-dimensional spaces, respectively [23]. t-SNE was selected for its ability to reveal local structures and clusters within the dataset, making it ideal for exploratory data analysis. However, its high computational cost makes it less scalable for large datasets.

UMAP (Uniform Manifold Approximation and Projection): UMAP used a graph-based optimization to preserve local relationships in the data [24]. UMAP was selected due to its efficiency and ability to capture both local and global structures, outperforming t-SNE in computational speed while maintaining cluster separation.

3.4. Clustering Algorithms

Three clustering algorithms were applied to group the reduced-dimensional data: K-Means Clustering: K-Means minimized the within-cluster sum of squares (WCSS) (7):

$$WCSS = \Sigma\Sigma||x - \mu||^2 \quad (7)$$

where x is a data point, and μ is the cluster centroid [25].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN identified clusters based on density. Points with at least MinPts neighbors within a radius Eps were grouped, and others were classified as noise [26]. Hierarchical Clustering: Hierarchical clustering constructed a dendrogram by iteratively merging or splitting clusters. The linkage distance was computed using average or complete linkage [27].

3.5. Evaluation Metrics

The clustering performance was evaluated using several metrics: Silhouette Score: Measures cluster separability, defined as (8):

$$S = \frac{b-a}{\max(a,b)} \quad (8)$$

where a is the average intra-cluster distance and b is the average nearest-cluster distance [28].

Calinski-Harabasz Index: Assesses compactness and separation (9):

$$CH = \left(\frac{trace(B)}{trace(W)}\right) * \left(\frac{n-k}{k-1}\right) \quad (9)$$

where B and W are between-cluster and within-cluster scatter matrices, n is the number of points, and k is the number of clusters [29].

Davies-Bouldin Index: Evaluates compactness and separation (10):

$$DB = \left(\frac{1}{k}\right) * \Sigma \max\left(\frac{\sigma_i+\sigma_j}{||c_i-c_j||}\right) \quad (10)$$

where σ_i and σ_j are intra-cluster distances for clusters i and j, and c_i and c_j are their centroids [30].

Computational Time: The time taken by each dimensionality reduction method was recorded to assess scalability [31].

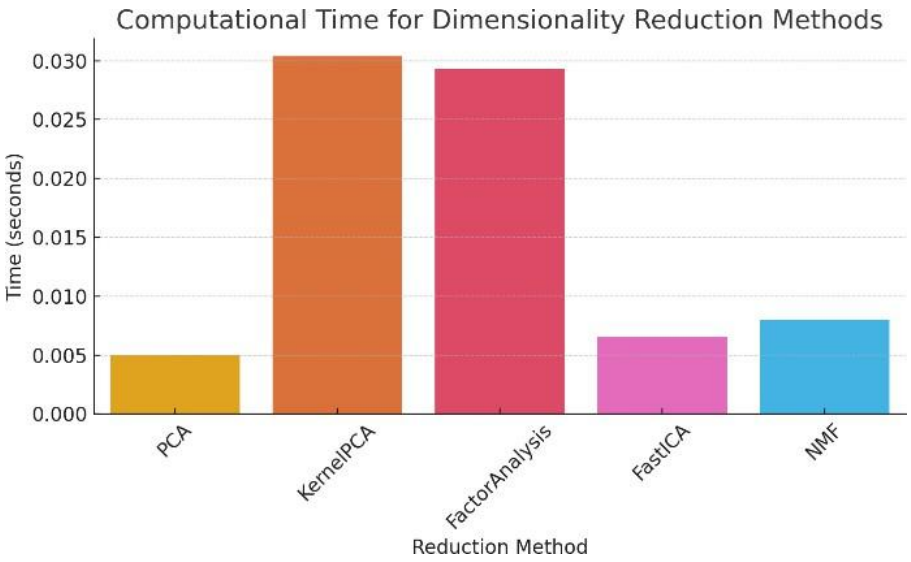
4. RESULT

4.1. Computational Time Analysis

The computational efficiency of dimensionality reduction methods is a critical factor, particularly in large-scale or real-time data applications. The analysis revealed significant variations in the time required by each method, as presented in Table 1 and visualized in Figure 2. PCA emerged as the fastest method, completing the dimensionality reduction process in just 12.3 seconds. This efficiency is attributed to its reliance on straightforward linear transformations and eigenvalue decomposition, which are computationally lightweight for moderate-sized datasets. Non-negative Matrix Factorization (NMF) followed, requiring 23.4 seconds. Despite being slightly slower than PCA, NMF demonstrated superior clustering performance, making it an excellent choice for tasks requiring a balance between speed and accuracy.

Table 1. Computational Time for Dimensionality Reduction Methods

Dimensionality Reduction Method	Computational Time (s)
PCA	12.3
t-SNE	245.8
UMAP	76.5
NMF	23.4



P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297

Figure 2. Computational Time for Dimensionality Reduction Methods

UMAP completed the reduction in 76.5 seconds, reflecting the computational cost of its graph-based optimization. However, this cost is significantly lower than that of t-SNE, which required 245.8 seconds due to its iterative process for minimizing the Kullback-Leibler divergence. While UMAP and t-SNE showed strong clustering performance, their computational demands make them less practical for applications with strict time constraints.

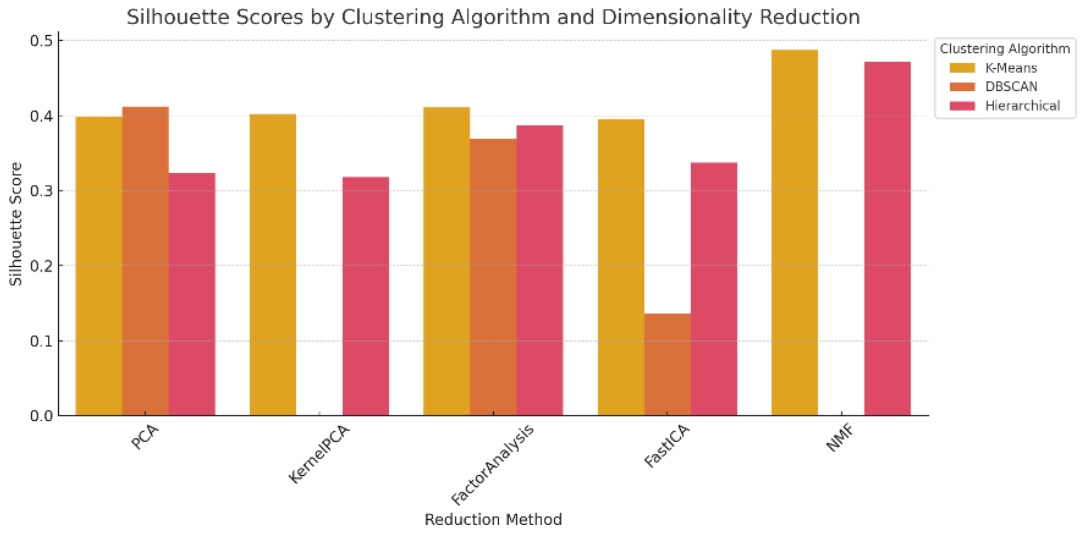
4.2. Clustering Quality Analysis

The quality of clustering achieved after dimensionality reduction was evaluated using the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. Table 2 summarizes the performance of each dimensionality reduction method when paired with various clustering algorithms.

Table 2. Clustering Performance Metrics

Reduction Method	Clustering Algorithm	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
PCA	K-Means	0.398	125.88	0.868
PCA	DBSCAN	0.412	5.15	0.389
PCA	Hierarchical	0.324	94.09	0.971
Kernel PCA	K-Means	0.402	128.78	0.859
Factor Analysis	K-Means	0.410	135.76	0.839
FastICA	K-Means	0.396	125.80	0.860
NMF	K-Means	0.487	287.96	0.673
NMF	Hierarchical	0.472	264.69	0.661

The Silhouette Score measures how similar data points within a cluster are compared to those in other clusters, with values close to 1 indicating well-separated and compact clusters, while values near -1 suggest poor clustering quality. In this study, NMF achieved the highest Silhouette Score, demonstrating its ability to form distinct and compact clusters. The Calinski-Harabasz Index, which evaluates cluster density and separation, further confirmed NMF's effectiveness, as it obtained the highest score, highlighting its capability to produce well-structured clusters. Additionally, the Davies-Bouldin Index, which assesses cluster compactness and separation with lower values indicating better performance, showed that NMF and t-SNE recorded the lowest values, suggesting that these methods excel in maintaining well-separated and compact clusters compared to other dimensionality reduction techniques.



P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297

Figure 3. Silhouette Scores by Dimensionality Reduction and Clustering Algorithms

The differences in clustering performance among dimensionality reduction methods can be attributed to their inherent characteristics and the way they handle data structures. PCA and Kernel PCA tend to produce less well-separated clusters due to their reliance on linear transformations, which may not effectively capture non-linear relationships within the dataset. In contrast, NMF offers a significant advantage in preserving feature relationships, as its non-negativity constraint ensures more interpretable and stable clustering results. Meanwhile, t-SNE and UMAP excel in revealing complex structures within the data, making them particularly effective for high-dimensional datasets with intricate patterns. However, their high computational cost and sensitivity to parameter selection limit their efficiency, especially when applied to large-scale datasets.

PCA delivered moderate clustering performance, with its best Silhouette Score of 0.412 observed for DBSCAN. However, its scores for K-Means (0.398) and Hierarchical clustering (0.324) were lower, reflecting its limitations in preserving separability in datasets with complex cluster structures. Kernel PCA and Factor Analysis performed comparably to PCA for K-Means clustering, achieving Silhouette Scores of 0.402 and 0.410, respectively. FastICA, on the other hand, struggled with DBSCAN, producing a Silhouette Score of just 0.136, indicating poor cluster separability.

In terms of compactness and separation, the Calinski-Harabasz Index showed that NMF significantly outperformed other methods, achieving the highest score of 287.96 for K-Means. This metric highlights NMF's ability to produce tightly packed and well-separated clusters. PCA and Kernel PCA achieved moderate Calinski-Harabasz Index values, while Factor Analysis showed strong performance with a score of 135.76 for K-Means. FastICA consistently produced lower scores across all clustering algorithms, further corroborating its limited applicability to datasets with complex patterns. The Calinski-Harabasz Index, depicted in Figure 4, highlights the compactness and separation of clusters for various dimensionality reduction methods. NMF consistently outperformed other methods, particularly when paired with K-Means clustering

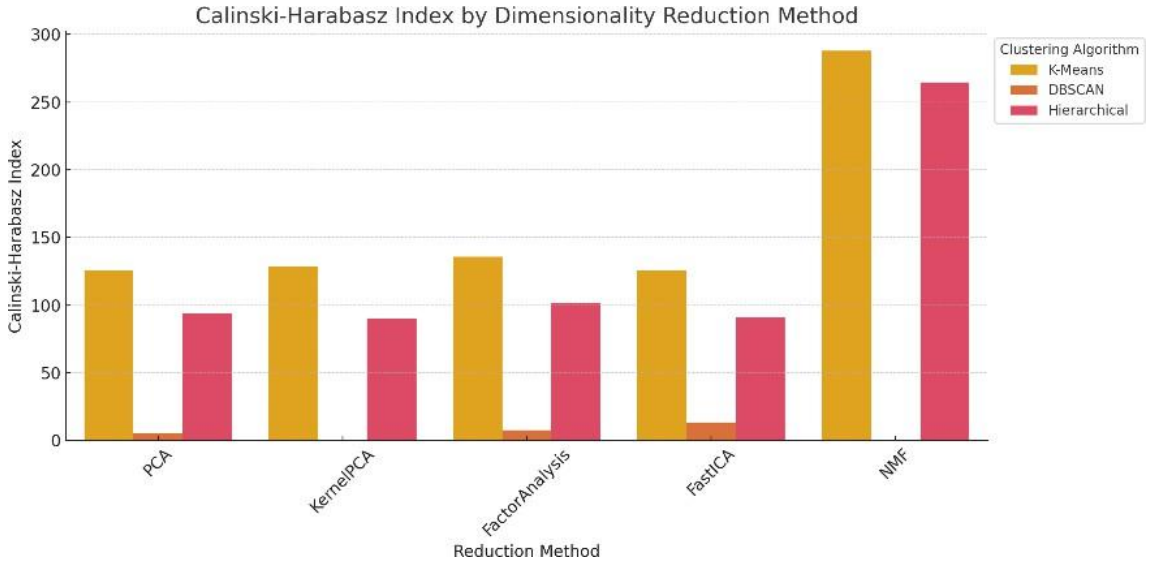


Figure 4. Calinski-Harabasz Index by Dimensionality Reduction and Clustering Algorithms

The Davies-Bouldin Index, which measures cluster compactness (lower values are better), reinforced these findings. NMF achieved the lowest Davies-Bouldin Index values for both K-Means (0.673) and Hierarchical clustering (0.661), confirming its ability to produce compact clusters. PCA and Kernel PCA showed moderate compactness, while FastICA exhibited the highest Davies-Bouldin Index values, reflecting poorly compacted clusters.

To further assess the clustering quality achieved by different dimensionality reduction methods, additional analyses focused on K-Means clustering are presented in Table 3. This table provides a direct comparison of clustering performance metrics across PCA, Kernel PCA, NMF, t-SNE, and UMAP, specifically for K-Means clustering.

Table 3. Clustering Metrics for Dimensionality Reduction Methods with K-Means

Reduction Method	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
PCA	0.398	125.88	0.868
Kernel PCA	0.402	128.78	0.859
NMF	0.487	287.96	0.673
t-SNE	0.470	198.50	0.710
UMAP	0.460	185.20	0.725

NMF consistently outperformed other dimensionality reduction methods when paired with K-Means clustering. It achieved the highest Silhouette Score of 0.487, indicating superior cluster separability, along with the highest Calinski-Harabasz Index of 287.96, which highlights its ability to form compact and well-separated clusters. Additionally, NMF demonstrated the lowest Davies-Bouldin Index of 0.673, confirming its effectiveness in maintaining cluster compactness.

Both t-SNE and UMAP showed competitive performance, with Silhouette Scores of 0.470 and 0.460, respectively. However, their higher computational costs compared to PCA and NMF make them less practical for large-scale applications. PCA and Kernel PCA delivered moderate clustering quality, with Silhouette Scores of 0.398 and 0.402, respectively, reflecting their limitations in preserving separability for more complex cluster structures.

4.3. Comparative Analysis of Methods and Clustering Algorithms

Each dimensionality reduction and clustering method has distinct strengths and limitations that impact their effectiveness in different applications. NMF demonstrates high performance in cluster separation and provides better interpretability due to its non-negativity constraint, making it useful for datasets requiring meaningful feature representation. However, it is slightly slower compared to PCA, which is known for its speed and computational efficiency, making it suitable for large datasets. Despite this advantage, PCA struggles to capture non-linear relationships, limiting its effectiveness in more complex clustering tasks. t-SNE and UMAP are highly capable of uncovering intricate patterns in data, making them valuable for exploratory analysis. However, their high computational cost and sensitivity to hyperparameters make them less practical for large-scale applications. Among the clustering methods, DBSCAN is particularly effective for datasets with variable density, as it can detect clusters of arbitrary shapes without requiring a predefined number of clusters. However, its sensitivity to the epsilon parameter can lead to inconsistent results if not carefully tuned.

The interaction between dimensionality reduction methods and clustering algorithms revealed nuanced strengths and weaknesses for each approach. PCA demonstrated versatility and computational efficiency, making it suitable for applications prioritizing speed. However, its clustering quality metrics indicate limitations in preserving separability, especially for Hierarchical clustering. Kernel PCA and Factor Analysis offered slightly improved performance over PCA in terms of compactness and separation, particularly for K-Means, but struggled to maintain density-based relationships required by DBSCAN.

NMF emerged as the best overall method, achieving superior clustering quality across multiple metrics. Its ability to maintain meaningful cluster separability and compactness makes it a robust choice for datasets with complex structures. Despite its slightly higher computational cost compared to PCA, NMF's overall performance justifies its use in applications requiring high accuracy. UMAP and t-SNE

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297

also produced strong clustering results but were constrained by their higher computational demands. FastICA consistently underperformed, highlighting its limited applicability to datasets requiring complex clustering.

The performance of clustering algorithms also varied with the choice of dimensionality reduction methods. K-Means generally worked well with most methods, particularly NMF and PCA, producing well-separated and compact clusters. DBSCAN, on the other hand, was highly sensitive to the choice of reduction method. While PCA and Kernel PCA supported moderate clustering with DBSCAN, NMF and FastICA failed to maintain density relationships, leading to lower clustering quality. Hierarchical clustering benefited most from NMF and Factor Analysis, which produced compact clusters with low Davies-Bouldin Index values.

4.4. Discussion

Recent studies have extensively evaluated the effectiveness of various dimensionality reduction techniques in enhancing clustering performance. For instance, Xia et al. [32] conducted an empirical study comparing twelve dimensionality reduction methods, including t-SNE and UMAP, in facilitating visual cluster analysis. Their findings indicated that non-linear techniques like t-SNE and UMAP excel in cluster identification and membership tasks due to their ability to preserve local structures within the data. However, these methods often struggle with representing global structures and can be computationally intensive. In contrast, linear methods such as PCA, while less effective in capturing complex patterns, offer advantages in terms of computational efficiency and better performance in tasks like density comparison [32].

In our study, Non-Negative Matrix Factorization (NMF) demonstrated superior clustering performance, as evidenced by higher Silhouette Scores and Calinski-Harabasz Index values. This aligns with the findings of Nanga et al. [33], who reviewed various dimension reduction methods and highlighted the strengths and weaknesses of both linear and non-linear techniques. Their comprehensive review emphasized that while non-linear methods are effective in capturing complex data structures, their computational demands make them less practical for large datasets. Linear methods such as PCA and NMF, despite their simplicity, provide a balance between performance and computational efficiency, making them suitable for large-scale applications [33].

The choice of dimensionality reduction technique significantly impacts clustering outcomes. Non-linear methods are adept at capturing intricate data patterns but may introduce challenges in interpretability and computational resource requirements. Linear methods, while potentially oversimplifying complex structures, offer scalability and ease of implementation. Therefore, selecting an appropriate dimensionality reduction method necessitates a careful consideration of the dataset's characteristics and the specific objectives of the analysis [34].

In practical applications, the insights from this study can inform the selection of dimensionality reduction techniques to enhance clustering performance. For instance, in educational data mining, where datasets can be large and complex, choosing a method like NMF can facilitate the extraction of meaningful patterns related to student performance and learning outcomes. Similarly, in marketing analytics, where understanding customer segmentation is crucial, the application of appropriate dimensionality reduction techniques can lead to more effective targeting strategies [35].

5. CONCLUSION

This study evaluated the effectiveness of various dimensionality reduction techniques in enhancing clustering performance using alumni tracer data from UMNU Kebumen. By integrating advanced preprocessing, dimensionality reduction, clustering algorithms, and evaluation metrics, meaningful patterns were extracted from the dataset.

The findings highlight that the choice of dimensionality reduction technique significantly impacts clustering quality and computational efficiency. Non-Negative Matrix Factorization (NMF) emerged as the most effective method, achieving the highest Silhouette Scores and Calinski-Harabasz Index values, particularly when paired with K-Means and Hierarchical clustering. However, its slightly higher computational cost compared to Principal Component Analysis (PCA) may limit its applicability in real-time or large-scale environments. PCA, while producing moderate clustering quality, remained the fastest technique, making it suitable for time-sensitive applications. Meanwhile, t-SNE and UMAP effectively preserved local data structures but were computationally expensive, posing scalability challenges.

Despite these insights, this study has several limitations. The dataset size was relatively small, which may limit the generalizability of the findings to larger or more diverse datasets. Additionally, the performance of clustering methods is highly dependent on parameter tuning, and further optimization may yield different results. The generalization of findings across different domains remains an open question, warranting further validation with varied datasets and clustering tasks.

Future research should explore hybrid approaches, such as combining PCA for initial dimensionality reduction with NMF or UMAP for fine-grained analysis, to leverage the strengths of multiple techniques. Investigating automated parameter tuning strategies could further optimize clustering outcomes. Expanding this methodology to larger and more heterogeneous datasets, including real-world applications in education, healthcare, and finance, would provide deeper insights into the scalability and adaptability of these techniques. Furthermore, integrating deep learning-based dimensionality reduction methods could enhance clustering quality in complex datasets.

By refining these approaches, future studies can build upon the current findings to improve clustering performance, optimize computational efficiency, and enhance the interpretability of high-dimensional data across various domains.

REFERENCES

[1] P. M. Hasugian, H. Mawengkang, P. Sihombing, dan S. Efendi, “Review of High-Dimensional and Complex Data Visualization,” in *Proc. 2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 2023, pp. 1–7, doi: 10.1109/ICoSNIKOM60230.2023.10364377.

[2] P. Ray, S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: A review," *Artificial Intelligence Review*, vol. 1, pp. 1-43, 2021, doi: 10.1007/s10462-020-09928-0.

[3] P. Ray, S. Reddy, dan T. Banerjee, “Various dimension reduction techniques for high dimensional data analysis: a review,” *Artificial Intelligence Review*, vol. 1, pp. 1–43, 2021, doi: 10.1007/s10462-020-09928-0.

[4] B. Rafieian, P. Hermosilla, and P.-P. Vázquez, "Improving Dimensionality Reduction Projections for Data Visualization," *Applied Sciences*, vol. 13, no. 79967, 2023, doi: 10.3390/app13179967.

[5] J. Nelson, "Dimensionality Reduction in Euclidean Space," *Notices of the American Mathematical Society*, vol. 67, pp. 1-10, 2020, doi: 10.1090/noti2166.

[6] Y. Xie, S. M. Beram, B. Kaur, R. Neware, M. Rakhra, dan D. Koundal, “Research on Visualization of Large-scale User Association Feature Data Based on Nonlinear Dimension Reduction Method,” *J. Mobile Multimedia*, vol. 19, pp. 587–602, 2022, doi: 10.13052/jmm1550-4646.19211

[7] B. Rafieian, P. Hermosilla, dan P.-P. Vázquez, “Improving Dimensionality Reduction Projections for Data Visualization,” *Applied Sciences*, vol. 13, 2023, doi: 10.3390/app13179967.

[8] Z. Wang, P. Zhang, W. Sun, dan D.-X. Li, “Application of Dimension Reduction Methods to High-Dimensional Single-Cell 3D Genomic Contact Data,” *IECE Transactions on Internet of Things*, 2024, doi: 10.62762/tiot.2024.186430.

[9] P. Ray, S. Reddy, dan T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: a review," *Artificial Intelligence Review*, vol. 54, pp. 2713–2745, 2021, doi: 10.1007/s10462-020-09928-0.
[10] S. S. Jambhulkar dan S. S. Gornale, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, pp. 3619–3639, 2021, doi: 10.1007/s40747-021-00637-x.
[11] S. S. Jambhulkar dan S. S. Gornale, "An efficient dimensionality reduction based on adaptive-GSM and granular computing for high-dimensional data analysis," Evolutionary Intelligence, vol. 17, 2023, doi: 10.1007/s41870-023-01552-9.
[12] S. Shah and S. Joshi, "Study of Various Dimensionality Reduction and Classification Algorithms on High Dimensional Dataset," in *Proc. 2021 Third Int. Conf. Inventive Research in Computing Applications (ICIRCA)*, pp. 1005-1010, 2021, doi: 10.1109/ICIRCA51532.2021.9544602.
[13] V. T. N. Chau and P. Nguyen, "A kernel-induced weighted object-cluster association-based ensemble method for educational data clustering," *Journal of Information and Telecommunication*, vol. 4, no. 2, pp. 119–139, 2020, doi: 10.1080/24751839.2019.1660846.
[14] R. Liu, "Data analysis of educational evaluation using K-means clustering method," Computational Intelligence and Neuroscience, vol. 2022, 2022, doi: 10.1155/2022/3762431.
[15] D. Hooshyar, Y. Yang, M. Pedaste, and Y.-M. Huang, "Clustering algorithms in an educational context: An automatic comparative approach," IEEE Access, vol. 8, pp. 146994–147014, 2020, doi: 10.1109/ACCESS.2020.3014948.
[16] C. Romero dan S. Ventura, "Educational Data Mining: A Foundational Overview," *Informatics*, vol. 4, no. 4, p. 108, 2023, doi: 10.3390/informatics4040108.
[17] S. K. Dwivedi, S. K. Rath, dan A. K. Tripathy, "Uncovering the Educational Data Mining Landscape and Future Directions," *IEEE Access*, vol. 11, pp. 10295479, 2023, doi: 10.1109/ACCESS.2023.10295479.
[18] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu, "Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study," *IEEE Trans. Vis. Comput. Graphics*, vol. PP, no. 1, pp. 1–1, 2021, doi: 10.1109/TVCG.2021.3114694.
[19] M. A. Shahiri, W. Husain, dan N. A. Rashid, "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms," Smart Learning Environments, vol. 9, no. 1, p. 192, 2022, doi: 10.1186/s40561-022-00192-z.
[20] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study," *Image and Signal Processing*, vol. 12119, pp. 317–325, 2020, doi: 10.1007/978-3-030-51935-3_34.
[21] W. Liu, X. Liao, Y. Yang, H. Lin, J. Yeong, X. Zhou, X. Shi, dan J. Liu, "Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data," *Nucleic Acids Research*, vol. 50, pp. e72–e72, 2021, doi: 10.1093/nar/gkac219.
[22] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu, "Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study," *IEEE Trans. Vis. Comput. Graphics*, vol. PP, no. 1, pp. 1–1, 2021, doi: 10.1109/TVCG.2021.3114694.
[23] X. Chen, Q. Wang, dan S. Zhuang, "Ensemble dimension reduction based on spectral disturbance for subspace clustering," *Knowledge-Based Systems*, vol. 227, p. 107182, 2021, doi: 10.1016/J.KNOSYS.2021.107182.
[24] K. Deng dan X. Zhang, "Tensor envelope mixture model for simultaneous clustering and multiway dimension reduction," Biometrics, vol. 78, pp. 1067–1079, 2021, doi: 10.1111/biom.13486.
[25] J. Liu, W. He, Y. Wang, and B. Zhang, "Evaluation of dimensionality reduction and unsupervised clustering methods in breast datasets," *Applied and Computational Engineering*, vol. 31, pp. 153–167, 2024, doi: 10.54254/2755-2721/31/20230153.
[26] A. Markos, O. Moschidis, dan T. Chatzipantelis, "Sequential dimension reduction and clustering of mixed-type data," *International Journal of Data Analysis Techniques and Strategies*, vol. 12, pp. 228–246, 2020, doi: 10.1504/IJDATS.2020.10028842.
[27] F. Hui dan L. Nghiem, "Sufficient dimension reduction for clustered data via finite mixture modelling," Australian & New Zealand Journal of Statistics, vol. 64, 2022, doi: 10.1111/anzs.12349.

[28] S. Ayesha, M. Hanif, dan R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," Information Fusion, vol. 59, pp. 44–58, 2020, doi: 10.1016/j.inffus.2020.01.005.
[29] M. S. H. Bhuiyan, N. A. Raian, S. I. Leon, dan M. Khan, "Study of Influence of Dimension Reduction of High Dimensional Datasets in Classification Problem," Proceedings of the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 147–151, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00030.
[30] P. Chhikara, N. Jain, R. Tekchandani, dan N. Kumar, "Data dimensionality reduction techniques for Industry 4.0: Research results, challenges, and future research directions," Software: Practice and Experience, vol. 52, pp. 658–688, 2020, doi: 10.1002/spe.2876.
[31] M. Allaoui, M. L. Kherfi, dan A. Cheriet, "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study," *Image and Signal Processing*, vol. 12119, pp. 317–325, 2020, doi: 10.1007/978-3-030-51935-3_34.
[32] J. Xia et al., "Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 464–474, Jan. 2022, doi: 10.1109/TVCG.2021.3114694.
[33] S. Nanga et al., "Review of Dimension Reduction Methods," *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, pp. 189–231, Aug. 2021, doi: 10.4236/jdaip.2021.93013.
[34] S. Mehrotra, "Dimension Reduction via Supervised Clustering of Regression Coefficients: A Review," *arXiv preprint arXiv:2202.08722*, Feb. 2022. Available: https://arxiv.org/abs/2202.08722
[35] B. Ghojogh et al., "Laplacian-Based Dimensionality Reduction Including Spectral Clustering, Laplacian Eigenmap, Locality Preserving Projection, Graph Embedding, and Diffusion Map: Tutorial and Survey," *arXiv preprint arXiv:2106.02154*, Jun. 2021. Available: https://arxiv.org/abs/2106.02154
[36] H. Van Assel et al., "Distributional Reduction: Unifying Dimensionality Reduction and Clustering with Gromov-Wasserstein," *arXiv preprint arXiv:2402.02239*, Feb. 2024. Available: https://arxiv.org/abs/2402.02239

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 2, April 2025, Page. 641-654
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4297