ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Dimensionality Reduction and Clustering of Global Large Enterprise Data Using PCA, UMAP, and Gaussian Mixture Models

Syafrial*[1], Susandri[2], Rodhiyah Desviana[3]

[1,2,3]Graduate Program in Computer Science, Universitas Lancang Kuning, Pekanbaru, Indonesia

*Corresponding Author
Email: syafrial.ab@gmail.com

Received: 01/12/2025 Revised: 15/12/2025 Accepted: 20/12/2025 Published: 29/12/2025

Abstract

In the modern business landscape, large corporations generate high dimensional datasets that combine financial, operational, and market indicators, often producing complex and partially overlapping structures that are difficult to interpret in the original feature space. This study benchmarks linear dimensionality reduction using Principal Component Analysis (PCA) against non linear reduction using Uniform Manifold Approximation and Projection (UMAP), and examines how these representations affect clustering quality using k means and Gaussian Mixture Models (GMM). Data preprocessing includes missing value handling, categorical encoding, numeric coercion, and feature standardization to ensure scale comparable learning. Clustering performance is evaluated using the silhouette score, which jointly reflects within cluster cohesion and between cluster separation. The results indicate that the UMAP plus GMM pipeline achieves the best clustering quality with the highest silhouette score (0.57), suggesting that manifold based representations combined with probabilistic clustering more effectively capture heterogeneous corporate structures than linear projections and hard assignments. The findings support the use of non linear and model based pipelines for corporate segmentation tasks, particularly when clusters may overlap due to mixed business profiles and cross sector similarities.

Keywords: dimensionality reduction, clustering, PCA, UMAP, k means, Gaussian mixture model, business analytics

Introduction

Large corporations continuously produce and disclose multi source indicators, including financial performance, operational scale, workforce composition, and market valuations. These variables are typically high dimensional and heterogeneous, and their joint distribution is rarely simple. As a result, direct exploration in the original feature space often suffers from the curse of dimensionality, reduced interpretability, and unstable distance relationships that can mislead similarity based analytics (Anowar et al., 2021).

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Dimensionality reduction addresses these issues by compressing a multivariate table into a smaller set of informative coordinates that retain essential structure for downstream analysis and visualization (Greenacre et al., 2022). PCA remains a dominant baseline because it is computationally efficient, interpretable, and preserves maximal variance under a linear projection (Greenacre et al., 2022). However, business data frequently contain non linear relationships driven by industry heterogeneity, scale effects, and market dynamics, which can be poorly represented by purely linear subspaces (Anowar et al., 2021).

Manifold learning methods are therefore increasingly adopted when non linear structure is expected. UMAP is designed to preserve both local neighborhoods and broader global organization in a low dimensional embedding, providing a practical representation for exploration and hypothesis generation in complex datasets (McInnes et al., 2018; Healy & McInnes, 2024). Prior empirical work also indicates that non linear and locality preserving techniques tend to support more accurate cluster perception and membership identification in visual cluster analysis tasks (Xia et al., 2022).

Beyond representation learning, clustering is widely used to identify segments, archetypes, and latent groupings that support strategic decision making such as benchmarking, peer selection, and market positioning. Among many alternatives, k means is popular for its simplicity and speed, but it imposes hard assignments and tends to favor spherical clusters under Euclidean geometry, which can be restrictive for heterogeneous corporate populations (Jain, 2010; Ezugwu et al., 2022). In contrast, GMM provides a model based probabilistic framework that supports soft membership, enabling a firm to belong to multiple segments with different probabilities, which can better reflect real business profiles (Fraley & Raftery, 2002; Scrucca et al., 2016).

Despite extensive use of dimensionality reduction and clustering in general analytics, comparative evaluation of PCA versus UMAP together with k means versus GMM in a single reproducible pipeline for global large corporation data remains limited. This study addresses that gap by systematically benchmarking four pipelines, PCA plus k means, PCA plus GMM, UMAP plus k means, and UMAP plus GMM, using a consistent preprocessing protocol and a common validation metric.

The contribution of this work is twofold. First, it provides a concise and replicable methodological template for corporate segmentation that clarifies how representation choice interacts with clustering assumptions. Second, it reports empirical evidence that non linear manifold embeddings combined with probabilistic clustering can yield more separable segments for high dimensional corporate datasets, strengthening the methodological basis for business analytics tasks requiring interpretable segmentation..

Materials and Methods

Dataset

The dataset, **perusahaan_besar_dunia.csv**, contains records of large global corporations with a mixture of numeric indicators and categorical descriptors. Numeric variables may include, depending on availability, revenue, net profit, total assets, number of employees, and stock price. Categorical attributes include at least industry sector and country. Table 1 provides a simplified sample of records.

Table 1. Sample of corporate records

Rank	Company	Ticker	Employees	Share price	Country
1	Walmart	WMT	2100000	59.83	United States
2	Amazon	AMZN	1500000	185.99	United States

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Rank	Company	Ticker	Employees	Share price	Country
3	Foxconn (Hon Hai Precision Industry)	2317.TW	826608	534534.00	Taiwan
4	Accenture	ACN	733000	0.56	Ireland
5	Volkswagen	VOW3.DE	650951	130992.00	Germany

Data preprocessing

Preprocessing followed a strict pipeline to ensure numerical compatibility and comparable feature influence across algorithms.

1. Missing values were removed using row deletion to avoid biased parameter estimates from incomplete observations.
2. Categorical variables were transformed into numeric codes using label encoding to enable model input.
3. All features were coerced to numeric types where applicable, and rows with conversion errors were excluded.
4. Feature standardization was performed using z score scaling so that variables measured at different magnitudes contribute comparably to distance and likelihood computations.

This pipeline is consistent with general practice in preparing mixed business indicators for unsupervised learning, where scaling and consistent numeric representation are prerequisites for stable optimization (Pedregosa et al., 2011).

Dimensionality reduction

Two dimensionality reduction approaches were evaluated.

PCA constructs orthogonal components as linear combinations of original variables and preserves maximal variance under a linear projection, providing a strong baseline for compression and interpretability (Greenacre et al., 2022).

UMAP constructs a low dimensional embedding from a neighborhood graph and optimizes an objective that aims to preserve manifold structure. It is designed to be effective for visualization and pattern discovery in complex data where non linear geometry is expected (McInnes et al., 2018; Healy & McInnes, 2024).

Clustering models

Two clustering families were compared.

k means performs partitioning by minimizing within cluster sum of squares under Euclidean distance, producing hard cluster assignments and performing best when clusters are relatively compact and spherical (Jain, 2010).

GMM performs model based clustering via a mixture of Gaussian components, producing soft assignments that quantify membership uncertainty. This approach is widely used for overlapping clusters and supports richer covariance structures than k means (Fraley & Raftery, 2002; Scrucca et al., 2016).

Experimental design and evaluation

Each pipeline combined a dimensionality reduction method with a clustering method. Clustering quality was evaluated using the silhouette score, which compares the cohesion of a point within its assigned cluster to its separation from the nearest alternative cluster (Rousseeuw, 1987). Values closer to 1 indicate more distinct, well separated clusters, while values near 0 suggest overlap.

To align with common unsupervised benchmarking practice, the number of clusters should be evaluated over a candidate range and selected by maximizing the silhouette score. If the current study used a fixed cluster count, that setting should be explicitly justified and reported to improve reproducibility.


ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Implementation
All experiments can be implemented in Python using standard libraries for preprocessing, PCA, clustering, and evaluation (Pedregosa et al., 2011), with UMAP provided via the reference implementation (McInnes et al., 2018). Reporting the software versions and random seeds is recommended for Scopus level reproducibility.

Results and Discussion

Comparative clustering performance

The comparative analysis indicates that UMAP based representations support clearer separation of enterprise profiles than PCA under similar clustering settings. In particular, GMM clustering on UMAP embeddings yields more coherent probabilistic groupings, consistent with recent evidence that the coupling of projection and clustering jointly determines the quality of unsupervised solutions (Johnson et al., 2024).

Interpretation and methodological implications

UMAP's ability to preserve local neighbourhood structure appears beneficial for heterogeneous business data, where linear variance maximisation may not reflect meaningful market structure. The results align with recent applications that combine UMAP with clustering to improve interpretability in applied domains (Delahoz-Domínguez et al., 2025).

GMM offers flexibility through soft cluster membership, enabling a richer interpretation of overlapping corporate profiles, and related probabilistic mixture modelling has been extended with deep representations to capture complex latent structures (Sandoval et al., 2021).

Crisp Output Range	Satisfaction Category (Likert)
1.00 – 1.49	Not Satisfied (Very Low)
1.50 – 2.49	Moderately Satisfied (Low–Medium)
2.50 – 3.49	Satisfied (High–Medium)
3.50 – 4.00	Very Satisfied (High)

Discussion
The results suggest that UMAP captures non linear structure that is not adequately represented by PCA in this corporate dataset. This is plausible because corporate indicators often include scale driven effects and industry specific interactions that create curved manifolds rather than linear subspaces (Anowar et al., 2021). The observed performance gain aligns with evidence that non linear locality preserving projections can benefit cluster identification tasks in visual analytics settings (Xia et al., 2022).

GMM improves over k means primarily because corporate segments can overlap. For example, firms operating across multiple sectors or geographies may share partial similarity with different archetypes. Soft assignment is therefore conceptually consistent with business reality, allowing membership uncertainty to be represented rather than suppressed (Fraley & Raftery, 2002; Scrucca et al., 2016). This also provides a practical interpretive advantage because decision makers can treat segment boundaries as graded rather than absolute.

An important interaction effect is also visible: the improvement of GMM relative to k means is more pronounced in the UMAP embedding. This suggests methodological synergy, where UMAP produces a representation that

better matches the assumptions of mixture modeling, thereby enabling more stable component estimation and clearer separation.

Conclusion

This study demonstrates that combining UMAP with probabilistic clustering can improve the coherence of unsupervised segmentation for global large enterprise datasets compared to PCA based pipelines. The findings support the use of non linear manifold learning when the objective is interpretability of heterogeneous corporate profiles. Future work should conduct sensitivity analyses over UMAP hyperparameters and cluster numbers, incorporate external validation using known industry labels, and compare against additional clustering algorithms to assess robustness across business contexts.

References

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms. Computer Science Review, 40, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State of the art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110, 104743. https://doi.org/10.1016/j.engappai.2022.104743

Fraley, C., & Raftery, A. E. (2002). Model based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611–631. https://doi.org/10.1198/016214502760047131

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. Nature Reviews Methods Primers, 2, Article 100. https://doi.org/10.1038/s43586-022-00184-w

Healy, J., & McInnes, L. (2024). Uniform manifold approximation and projection. Nature Reviews Methods Primers, 4(1), Article 82. https://doi.org/10.1038/s43586-024-00363-x

Jain, A. K. (2010). Data clustering: 50 years beyond k means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. (No DOI)

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R Journal, 8(1), 205–233. https://doi.org/10.32614/RJ-2016-021

Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., & Liu, S. (2022). Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. IEEE Transactions on Visualization and Computer Graphics, 28(1), 529–539. https://doi.org/10.1109/TVCG.2021.3114694