Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

A Comparative Analysis of Clustering Algorithms for Expedia's Travel Dataset

Gregorius Airlangga
Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
gregorius.airlangga@atmajaya.ac.id

Submitted :Dec 14, 2024 | **Accepted** : Jan 22, 2025 | **Published** : Feb 9, 2025

Abstract: The effective segmentation of travel data is crucial for deriving actionable insights in the tourism and hospitality sectors. This study conducts a comprehensive evaluation of four clustering algorithms Agglomerative Clustering, DBSCAN, Gaussian Mixture Models (GMM), and KMeans on a travel dataset, using three widely recognized metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. The dataset was preprocessed through standardization and dimensionality reduction via Principal Component Analysis (PCA) to facilitate visualization and ensure computational efficiency. The results highlight significant differences in the performance of these algorithms. Agglomerative Clustering achieved the highest Silhouette Score, indicating superior cluster cohesion and separation, while KMeans recorded the highest Calinski-Harabasz Score, demonstrating strong inter-cluster variance. In contrast, DBSCAN performed poorly, producing low scores across all metrics, attributed to sensitivity to parameter selection and density irregularities in the dataset. Gaussian Mixture Models exhibited moderate performance but struggled with overlapping clusters due to limitations in modeling non-Gaussian data distributions. Visualization of clustering results confirmed these findings, revealing compact clusters for Agglomerative and KMeans, while DBSCAN and GMM showed less defined structures. This study underscores the importance of selecting clustering algorithms based on dataset characteristics and analysis objectives.
Keywords: Clustering Algorithms, Travel Data Analytics, Agglomerative, Clustering, KMeans, DBSCAN

INTRODUCTION

The integration of advanced clustering techniques into travel-related datasets offers promising opportunities to extract actionable insights for the tourism and hospitality industries (Aderline, Ting, & Atanda, 2024; Cherenkov et al., 2024; Semwal et al., 2023). As the global tourism sector becomes increasingly data-driven, understanding customer behaviors, preferences, and travel patterns is crucial for optimizing marketing strategies, enhancing user experiences, and driving economic growth (Aljizawi, 2024; Cherenkov et al., 2024; Guo, Mu, & Lou, 2024). Despite extensive research on travel data analytics, the complex and heterogeneous nature of such datasets presents persistent challenges in uncovering meaningful patterns, especially when dealing with high-dimensional data (Hamdi et al., 2022). Clustering, as an unsupervised machine learning technique, has emerged as a pivotal approach for segmenting data into groups of similar characteristics (Wang & Biljecki, 2022). It serves as a foundation for applications ranging from recommendation systems to market segmentation (Zangerle & Bauer, 2022). Traditional clustering algorithms, such as KMeans and Agglomerative Clustering, have been extensively employed due to their simplicity and interpretability (Ikotun, Ezugwu, Abualigah, Abuhaija, & Heming, 2023). However, the rise of more sophisticated methods, such as Gaussian Mixture Models and density-based clustering like DBSCAN, provides new avenues for handling noise, outliers, and varying cluster densities (Campello, Kröger, Sander, & Zimek, 2020). These methods promise to address limitations in existing techniques but require comprehensive evaluation to validate their effectiveness in real-world scenarios(Sarker, 2021).

The Expedia travel dataset, a large-scale collection of booking and user interaction data, represents a unique resource for exploring the utility of clustering models (Tian et al., 2021). This dataset encompasses diverse numerical features, making it a compelling testbed for clustering algorithms. However, prior studies utilizing similar datasets often lack rigorous evaluation frameworks, particularly cross-validation techniques that assess model robustness across varying data splits (Yates, Aandahl, Richards, & Brook, 2023). Moreover, comparisons of clustering performance metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score remain underexplored in the domain of travel data analysis (Bolaños-Martinez, Bermudez-Edo, & Garrido, 2024).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

This study addresses the above gaps by systematically comparing the performance of multiple clustering algorithms, including KMeans, Agglomerative Clustering, DBSCAN, and Gaussian Mixture Models, on a representative subsample of the Expedia travel dataset (Ferretti, 2022). Employing a robust cross-validation framework, we evaluate the stability and reliability of each algorithm using three widely recognized metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. Furthermore, the study incorporates dimensionality reduction via Principal Component Analysis (PCA) to visualize clustering results, providing an intuitive understanding of the underlying data structures.

The main contribution of this research lies in its application of a systematic cross-validation framework to cluster algorithms, ensuring reproducibility and reliability in performance evaluations. By leveraging advanced clustering techniques and incorporating visualization strategies, this study provides a comprehensive perspective on the applicability and limitations of these models in travel data analytics. Our findings offer actionable insights for practitioners in the tourism industry and contribute to the broader literature on unsupervised learning in high-dimensional, real-world datasets. The remainder of this article is structured as follows. The next section presents a detailed literature survey, highlighting existing methodologies and their limitations in travel data clustering. Following this, the materials and methods section outlines the dataset preprocessing steps, clustering algorithms, and evaluation metrics. The results section presents empirical findings, supported by visualizations and cross-validation analysis. Finally, the discussion interprets the results in the context of existing literature, and the conclusion emphasizes the implications of the study, along with directions for future research.

LITERATURE REVIEW

The application of clustering techniques in analyzing travel data has garnered significant attention in recent years due to its potential to uncover latent patterns and insights (Chaudhry et al., 2023). This has proven especially relevant in the tourism and hospitality industries, where understanding customer behaviors, preferences, and travel patterns is critical for optimizing marketing strategies, enhancing user experiences, and driving economic growth (Aljizawi, 2024). Clustering, as an unsupervised machine learning technique, has been a cornerstone in segmenting data into groups of similar characteristics, enabling its application in areas such as recommendation systems and market segmentation (Reuvers, 2021). Among the commonly used clustering methods, KMeans clustering has been widely adopted due to its computational efficiency and straightforward implementation (Ikotun et al., 2023). Research has demonstrated its effectiveness in segmenting tourists based on preferences and travel behaviors, providing valuable insights for personalized services (Li & Cao, 2022). However, the method's reliance on spherical cluster assumptions limits its application to datasets with non-convex clusters or varying densities, common characteristics in real-world travel datasets (Wegmann, Zipperling, Hillenbrand, & Fleischer, 2021). Hierarchical clustering techniques, including both agglomerative and divisive approaches, have been employed to provide hierarchical representations of data (Ran, Xi, Lu, Wang, & Lu, 2023). These methods offer flexibility in determining the number of clusters and have been successfully used in categorizing destinations and user profiles. Despite their usefulness, hierarchical methods are computationally intensive for large datasets and exhibit sensitivity to noise and outliers (Shukla & Sengupta, 2020).

Density-based methods, such as DBSCAN, have emerged as alternatives due to their ability to identify clusters of arbitrary shapes and manage noise effectively (Hajihosseinlou, Maghsoudi, & Ghezelbash, 2024). These methods have been particularly valuable in detecting anomalous behaviors and grouping tourists with uncommon patterns (Banerjee & George, 2024). Nevertheless, their performance is highly dependent on parameter selection, and their sensitivity to these parameters can lead to suboptimal clustering in complex datasets. Gaussian Mixture Models (GMMs) offer another perspective by assuming that data points are generated from a mixture of Gaussian distributions (Alqahtani & Kalantan, 2020). Their probabilistic nature allows for flexibility in cluster modeling, and they have been applied effectively in segmenting airline passengers based on preferences and demographics (Chen, Zhang, & Gao, 2024). However, GMMs are constrained by their assumption of Gaussian distributions, which may not align with real-world data characteristics (Pesce, Krzakala, Loureiro, & Stephan, 2023). Evaluating the performance of clustering models is crucial to ensure their reliability and applicability. Metrics such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score are widely recognized for this purpose (Geng, 2024). The Silhouette Score measures the cohesion and separation of clusters, providing insights into the quality of clustering in datasets with well-separated groups (Chaudhry et al., 2023). The Davies-Bouldin Index evaluates the average similarity ratio between clusters, where lower values indicate better performance, and has proven effective in comparing clustering models in tourism data (Herrera, Arroyo, Jiménez, & Herrero, 2024). The Calinski-Harabasz Score, or Variance Ratio Criterion, measures the dispersion of data points within clusters relative to the dispersion between clusters, offering a robust evaluation of clustering performance, particularly in high-dimensional datasets (Solanki, 2021).

High-dimensional data presents unique challenges for clustering, as it often includes diverse numerical, categorical, and temporal features (Thudumu, Branch, Jin, & Singh, 2020). Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and t-SNE, are frequently employed to address these challenges

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

(Anowar, Sadaoui, & Selim, 2021). PCA, for instance, has demonstrated effectiveness in improving clustering outcomes by reducing data complexity, although concerns remain regarding the loss of information during dimensionality reduction (Jia, Sun, Lian, & Hou, 2022). Recent advancements in clustering algorithms aim to address these limitations. Spectral clustering, which utilizes eigenvalues of similarity matrices, has shown promise in identifying clusters in non-linear data structures (Rouhi, Bouyer, Arasteh, & Liu, 2024). Hybrid approaches that combine clustering algorithms with ensemble techniques have also been proposed to enhance robustness and accuracy, demonstrating improved performance in segmenting complex datasets (Yang, Zhang, Wang, Deng, & Li, 2021). Despite these advancements, significant gaps persist in literature. Many studies focus on applying individual clustering algorithms without systematically comparing their performance across varying conditions (Ezugwu et al., 2021; Ikotun et al., 2023). Furthermore, evaluation frameworks often lack rigor, particularly in the absence of cross-validation techniques to assess the robustness of models. While visualization methods such as PCA are acknowledged as valuable tools for understanding clustering outcomes, they are rarely integrated into performance evaluations, leaving a gap in the interpretability of results.

This review highlights that while clustering techniques have been successfully applied to travel data, limitations such as sensitivity to noise, reliance on assumptions, and inadequate evaluation frameworks hinder their broader applicability. To address these gaps, this study adopts a comprehensive approach, systematically comparing multiple clustering algorithms using a robust cross-validation framework and evaluating their performance with established metrics. By incorporating PCA for visualization, this research bridges the gap between cluster analysis and interpretability, providing a meaningful contribution to the field of travel data analytics. This survey establishes the context for the present research and underscores its importance in advancing the state of the art in clustering techniques for real-world applications.

METHOD

This section details the dataset preprocessing steps, clustering algorithms, and evaluation metrics employed in this study. The methodology emphasizes mathematical rigor to ensure clarity and reproducibility, providing a robust foundation for analyzing clustering performance.

Data Preprocessing

The dataset used in this study, denoted as $X = \{x_1, x_2, \dots, x_n\}$, consists of n data points, each represented as a vector $x_i \in R^d$ in d-dimensional space. To manage computational complexity and ensure data quality, the dataset was subjected to several preprocessing steps. First, a random subsample of size N was selected from the dataset to facilitate efficient computation, ensuring that the sample retained the overall structure of the original data. Missing values, which could potentially bias the clustering results, were removed entirely to produce a clean dataset, X', where all features are complete. To standardize the dataset, each feature x_j was normalized to have zero mean and unit variance. This transformation ensures that all features contribute equally to the clustering process, avoiding domination by features with larger numerical ranges. The standardized dataset, denoted as Z, is computed as $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, where μ_j and σ_j and are the mean and standard deviation of feature j, respectively. This preprocessing step guarantees that the clustering algorithms are not affected by differences in feature scales.

In addition to standardization, Principal Component Analysis (PCA) was optionally applied to reduce the dimensionality of the dataset. Dimensionality reduction aims to project the data onto a lower-dimensional space while retaining the maximum amount of variance. Mathematically, this involves computing the projection matrix $W \in R^{k \times d}$, where k is the number of principal components, and projecting the dataset as $Y = WZ$. This step aids in visualization and improves computational efficiency, especially for high-dimensional datasets.

Clustering Algorithms

Four clustering algorithms were employed in this study: K-Means, Agglomerative Clustering, DBSCAN, and Gaussian Mixture Models (GMM). The general goal of clustering is to partition the dataset Z into C clusters, denoted as $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_\mathcal{C}\}$, such that the union of all clusters forms the dataset, and each data point belongs to exactly one cluster. Each algorithm approaches this problem differently, as described below. The K-Means algorithm aims to minimize the within-cluster sum of squares (WCSS), defined as $J = \sum_{i=1}^{C} \sum_{z \in \mathcal{C}_i} |z - \mu_i|^2$ where μ_i is the centroid of cluster $\mathcal{C}_i$. By iteratively updating cluster assignments and centroids, K-Means seeks to converge to a solution that minimizes the intra-cluster variability. However, its reliance on spherical clusters can limit its effectiveness in datasets with non-convex or irregularly shaped clusters.

Agglomerative Clustering, a hierarchical method, constructs a dendrogram by iteratively merging the two closest clusters. The proximity between clusters is determined by a linkage criterion, such as single linkage, which considers the minimum distance between any two points in different clusters, or complete linkage, which considers the maximum distance. This approach is flexible and interpretable but can be computationally intensive for large datasets. Next, DBSCAN, or Density-Based Spatial Clustering of Applications with Noise, clusters data points

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

based on their density. A point is classified as a core point if its neighborhood, defined by a radius parameter ϵ, contains at least a minimum number of points. Non-core points are classified as noise or assigned to clusters depending on their proximity to core points. This algorithm is well-suited for datasets with clusters of arbitrary shapes and is robust to noise, although its performance heavily depends on the choice of parameters ϵ and the minimum number of points.

Lastly, Gaussian Mixture Models (GMMs) approach clustering probabilistically, assuming that the data is generated from a mixture of Gaussian distributions. The probability density function for each cluster is modeled as $p(z) = \sum_{i=1}^{C} \pi_i \mathcal{N}(z|\mu_i, \Sigma_i)$, where π_i is the weight of the i-th component, $\mathcal{N}(\cdot)$ is the Gaussian distribution, and μ_i and Σ_i are the mean and covariance matrix of the i-th Gaussian. The parameters are estimated using the Expectation-Maximization (EM) algorithm, which iteratively refines the likelihood of the data given the model.

Evaluation Metrics

The performance of the clustering algorithms was evaluated using three metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. These metrics provide a comprehensive evaluation of clustering quality by considering both intra-cluster cohesion and inter-cluster separation. The Silhouette Score measures the quality of clustering by comparing the average distance between points in the same cluster, $a(z_i)$, with the average distance to points in the nearest cluster, $b(z_i)$. The score for each point is computed as: $S(z_i) = \frac{b(z_i)-a(z_i)}{\max(a(z_i),b(z_i))}$, and the overall score is the mean across all points. Higher scores indicate better-defined clusters.

The Davies-Bouldin Index quantifies the similarity between clusters by computing the ratio of intra-cluster distance to inter-cluster distance. For clusters $\mathcal{C}_i$ and $\mathcal{C}_j$, the index is defined as $DB = \frac{1}{C}\sum_{i=1}^{C} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)}$, where σ_i is the average distance within cluster $\mathcal{C}_i$, and $d(\mu_i, \mu_j)$ is the distance between cluster centroids. Lower values indicate better clustering. The Calinski-Harabasz Score evaluates clustering performance based on the ratio of between-cluster dispersion to within-cluster dispersion. It is defined as $CH = \frac{\text{trace}(B)/(C-1)}{\text{trace}(W)/(n-C)}$, where B and W are the between-cluster and within-cluster scatter matrices, respectively. Higher values indicate better separation between clusters. This methodological framework ensures a rigorous evaluation of clustering algorithms and provides robust insights into their performance across diverse metrics.

RESULT

This section presents the results of the clustering experiments using four algorithms: Agglomerative Clustering, DBSCAN, Gaussian Mixture Models (GMM), and KMeans. The evaluation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, they are analyzed to assess clustering performance. Visualizations of the clustering outcomes, using PCA for dimensionality reduction, further support the quantitative results and provide insights into the clustering structures. Agglomerative Clustering demonstrated the highest Silhouette Score of 0.1100, indicating relatively better cohesion and separation of clusters compared to the other algorithms. However, its Davies-Bouldin Index of 2.5459 suggests that there is still a substantial overlap between clusters, and the Calinski-Harabasz Score of 276.1107 points to moderate inter-cluster separation.

Table 1. Clustering Performance

Index	Avg Silhouette	Avg Davies-Bouldin	Avg Calinski-Harabasz
Agglomerative	0.11001888897435683	2.5458778716637576	276.1106679906019
DBSCAN	-0.23804914085346957	2.551936802616206	7.724116820808881
GaussianMixture	0.06893843898723348	3.039991823536781	250.56341870991292
KMeans	0.0810684867077573	2.661619493607059	300.50806492233863

DBSCAN produced the lowest Silhouette Score at -0.2380, indicating poor clustering with overlapping or undefined clusters. This is further corroborated by its extremely low Calinski-Harabasz Score of 7.7241, which reflects poor cluster compactness and separation. DBSCAN's Davies-Bouldin Index of 2.5519, though slightly better than KMeans and GMM, still highlights significant inter-cluster similarity. The poor performance of DBSCAN can be attributed to the choice of parameters ϵ=1.5 and minimum samples, which may not have been well-suited for this dataset. Gaussian Mixture Models (GMM) achieved a Silhouette Score of 0.0689, slightly higher than DBSCAN but lower than Agglomerative and KMeans. The Davies-Bouldin Index of 3.0400 suggests poor inter-cluster separation, and its Calinski-Harabasz Score of 250.5634 indicates that the clusters were not well-separated in terms of variance. This performance suggests that GMM struggled to model the dataset's inherent structure, possibly due to the complexity of the data distribution and the assumption of Gaussian clusters.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

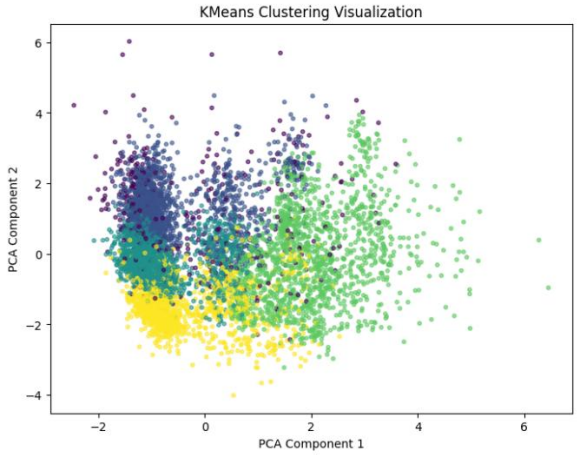

Fig. 1 KMeans Clustering Result

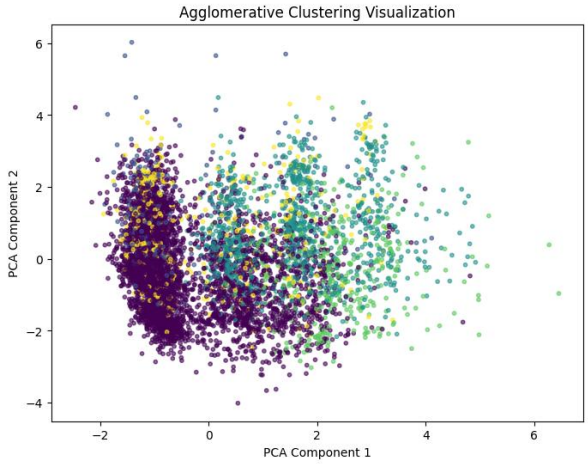

Fig. 2 Agglomerative Clustering Result

KMeans achieved the highest Calinski-Harabasz Score of 300.5081, indicating that it provided the best-defined clusters with high between-cluster variance relative to within-cluster variance. However, its Silhouette Score of 0.0811 and Davies-Bouldin Index of 2.6616 indicate moderate performance in terms of cluster cohesion and separation. While KMeans outperformed the other algorithms in certain metrics, it also exhibited limitations, particularly in handling irregularly shaped clusters. The visualizations of the clustering outcomes after applying PCA to reduce the dataset to two dimensions provide additional insights into the clustering structures and help interpret the quantitative results. Agglomerative Clustering displayed clusters that were moderately well-separated, consistent with its relatively high Silhouette and Calinski-Harabasz Scores. However, there is still noticeable overlap between clusters, especially in denser regions of the dataset. DBSCAN struggled to identify meaningful clusters, with a significant portion of the data assigned as noise or clustered into one large group. This outcome aligns with the poor quantitative results, particularly the low Calinski-Harabasz Score. Gaussian Mixture Models formed clusters with varying densities, but the overlapping nature of the clusters indicates a lack of clear separation, consistent with its low Silhouette and Calinski-Harabasz Scores. KMeans produced clusters that were compact and well-separated in some regions but showed overlapping boundaries in others. The visualization supports its relatively high Calinski-Harabasz Score, although the moderate Silhouette Score reflects the presence of overlapping points.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2025
DOI : https://doi.org/10.33395/sinkron.v9i1.14343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

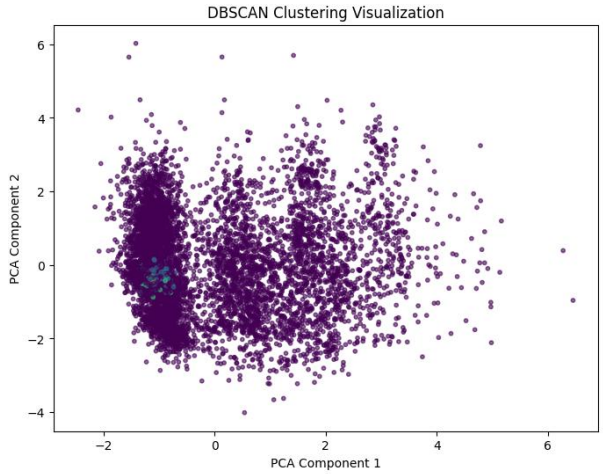

Fig. 3 DBSCAN Clustering Result

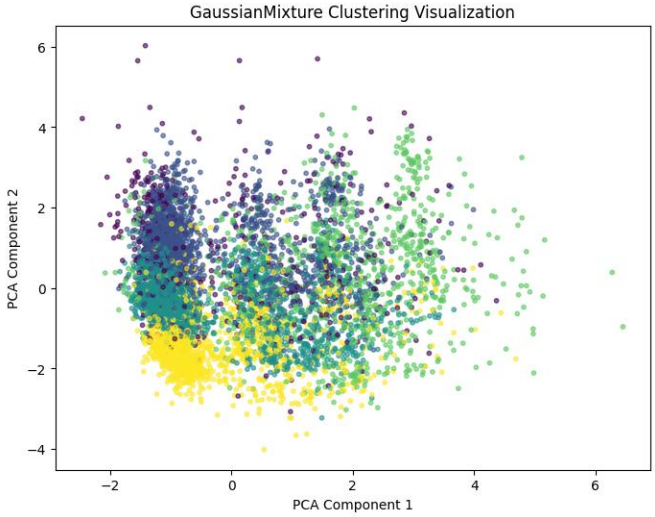

Fig. 4 Gaussian Mixture Clustering Result

DISCUSSIONS

The results highlight the varying performance of clustering algorithms on this dataset, emphasizing the strengths and limitations of each method. Agglomerative Clustering performed well in terms of overall cohesion and separation, making it a suitable choice for exploratory analysis of the dataset. However, its computational complexity may limit its scalability to larger datasets. KMeans showed strong inter-cluster separation, as evidenced by its Calinski-Harabasz Score, but struggled with overlapping clusters due to its reliance on spherical cluster assumptions. DBSCAN, while effective in identifying arbitrarily shaped clusters and handling noise in theory, failed to perform well on this dataset. This poor performance underscores the sensitivity of DBSCAN to parameter selection, suggesting that more careful tuning of ϵ and minimum samples is required to adapt the algorithm to the data's characteristics. Gaussian Mixture Models, although flexible and probabilistic, also underperformed, likely due to the dataset's deviation from Gaussian assumptions. These findings underscore the importance of selecting clustering algorithms based on the specific characteristics of the dataset and the goals of the analysis. While Agglomerative Clustering and KMeans demonstrated relatively better performance, further parameter tuning and hybrid approaches may yield improved results for datasets with complex structures. The incorporation of additional domain knowledge and feature engineering could also enhance the clustering outcomes. Future studies should explore these avenues to build on the findings of this research.

CONCLUSION

This study systematically evaluated the performance of four clustering algorithms Agglomerative Clustering, DBSCAN, Gaussian Mixture Models (GMM), and KMeans on a subsample of a travel dataset. The evaluation was based on three widely recognized metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score,

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

providing a comprehensive assessment of clustering quality. The results revealed significant differences in the algorithms' ability to identify meaningful clusters, highlighting the importance of algorithm selection based on dataset characteristics. Agglomerative Clustering demonstrated the best overall performance in terms of cluster cohesion and separation, as evidenced by its superior Silhouette Score. However, its computational complexity limits its application to larger datasets. KMeans excelled in terms of inter-cluster variance, achieving the highest Calinski-Harabasz Score, though its reliance on spherical cluster assumptions resulted in overlapping clusters in certain areas. Gaussian Mixture Models, while offering flexibility through probabilistic modeling, struggled to capture the complex structure of the data, likely due to the limitations of Gaussian assumptions. DBSCAN, despite its theoretical ability to handle noise and irregularly shaped clusters, performed poorly on this dataset, emphasizing the sensitivity of the algorithm to parameter selection and the dataset's density distribution.

The study highlights the necessity of considering the strengths and limitations of clustering algorithms in the context of specific datasets. While Agglomerative Clustering and KMeans showed relatively better performance, their limitations suggest that further parameter optimization and the incorporation of hybrid approaches could improve clustering outcomes. The poor performance of DBSCAN and Gaussian Mixture Models underscores the need for careful tuning and, potentially, the use of adaptive algorithms to address the challenges posed by real-world datasets. In conclusion, the findings of this study provide valuable insights into the applicability of different clustering algorithms for travel data analysis. Future work should focus on exploring hybrid clustering methods, integrating domain knowledge, and incorporating feature engineering to enhance clustering performance. Additionally, extending the evaluation framework to include other datasets and metrics could provide a broader perspective on clustering algorithm performance, further advancing the state of the art in data-driven travel analytics.

REFERENCES

Aderline, S. K. X., Ting, H. Y., & Atanda, A. F. (2024). Trends in tourism recommendation systems: a review/Aderline Song Ke Xin, Ting Huong Yong and Abdulwahab Funsho Atanda. *Journal of Computing Research and Innovation (JCRINN)*, *9*(2), 85–107.

Aljizawi, J. (2024). *Personalized Travel Recommendations and Marketing Automation for Saudi Arabia: Harnessing AI for Enhanced User Experience and Business Growth*.

Alqahtani, N. A., & Kalantan, Z. I. (2020). Gaussian mixture models based on principal components and applications. *Mathematical Problems in Engineering*, *2020*(1), 1202307.

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, *40*, 100378.

Banerjee, S., & George, A. (2024). Identifying overtourism \& spill-over tourism using ST-DBSCAN analysis for sustainable management of tourism. *Current Issues in Tourism*, 1–21.

Bolaños-Martinez, D., Bermudez-Edo, M., & Garrido, J. L. (2024). Clustering pipeline for vehicle behavior in smart villages. *Information Fusion*, *104*, 102164.

Campello, R. J. G. B., Kröger, P., Sander, J., & Zimek, A. (2020). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(2), e1343.

Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry*, *15*(9), 1679.

Chen, P., Zhang, X., & Gao, D. (2024). Preference heterogeneity analysis on train choice behaviour of high-speed railway passengers: A case study in China. *Transportation Research Part A: Policy and Practice*, *188*, 104198.

Cherenkov, E., Benga, V., Lee, M., Nandwani, N., Raguin, K., Sueur, M. C., & Sun, G. (2024). From Machine Learning Algorithms to Superior Customer Experience: Business Implications of Machine Learning-Driven Data Analytics in the Hospitality Industry. *Journal of Smart Tourism*, *4*(2), 5–14.

Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-Garc\'\ia, A., & Agushaka, J. O. (2021). Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, *33*, 6247–6306.

Ferretti, J. (2022). *Expedia Travel Dataset*. Retrieved from https://www.kaggle.com/datasets/jacopoferretti/expedia-travel-dataset/data

Geng, S. (2024). Analysis of the Different Statistical Metrics in Machine Learning. *Highlights in Science, Engineering and Technology*, *88*, 350–356.

Guo, Q., Mu, L., & Lou, S. (2024). Revolutionizing travel experiences: An in-depth analysis of intelligent booking systems and behavioral patterns. *Intelligent Decision Technologies*, *18*(2), 1477–1494.

Hajihosseinlou, M., Maghsoudi, A., & Ghezelbash, R. (2024). Intelligent mapping of geochemical anomalies: Adaptation of DBSCAN and mean-shift clustering approaches. *Journal of Geochemical Exploration*, 107393.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Hamdi, A., Shaban, K., Erradi, A., Mohamed, A., Rumi, S. K., & Salim, F. D. (2022). Spatiotemporal data mining: a survey on challenges and open problems. *Artificial Intelligence Review*, 1–48.

Herrera, A., Arroyo, Á., Jiménez, A., & Herrero, Á. (2024). Exploratory techniques to analyse Ecuador's tourism industry. *Logic Journal of the IGPL*, jzae040.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210.

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex \& Intelligent Systems*, *8*(3), 2663–2693.

Li, J., & Cao, B. (2022). Study on tourism consumer behavior and countermeasures based on big data. *Computational Intelligence and Neuroscience*, *2022*(1), 6120511.

Pesce, L., Krzakala, F., Loureiro, B., & Stephan, L. (2023). Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. *International Conference on Machine Learning*, 27680–27708.

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, *56*(8), 8219–8264.

Reuvers, S. (2021). *Discovering customer clusters using unsupervised machine learning to aid the marketing strategy: a case study with an online retail webshop SME*. University of Twente.

Rouhi, A., Bouyer, A., Arasteh, B., & Liu, X. (2024). Two-pronged feature reduction in spectral clustering with optimized landmark selection. *Applied Soft Computing*, *161*, 111775.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Semwal, R., Ranjan, S., Dhama, A., Chauhan, A., Bairwa, M. K., & Madhav, R. C. (2023). Conceptual Framework: Leveraging Artificial Intelligence for Enhanced Travel Review Analysis and Insights. *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, *6*, 2176–2181.

Shukla, R. M., & Sengupta, S. (2020). Scalable and robust outlier detector using hierarchical clustering and long short-term memory (lstm) neural network for the internet of things. *Internet of Things*, *9*, 100167.

Solanki, A. (2021). *Classification vs Clustering for Study Selection in Systematic Literature*.

Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, *7*, 1–30.

Tian, H., Presa-Reyes, M., Tao, Y., Wang, T., Pouyanfar, S., Miguel, A., … Iyengar, S. S. (2021). Data analytics for air travel data: a survey and new perspectives. *ACM Computing Surveys (CSUR)*, *54*(8), 1–35.

Wang, J., & Biljecki, F. (2022). Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, *129*, 103925.

Wegmann, M., Zipperling, D., Hillenbrand, J., & Fleischer, J. (2021). A review of systematic selection of clustering algorithms and their evaluation. *ArXiv Preprint ArXiv:2106.12792*.

Yang, W., Zhang, Y., Wang, H., Deng, P., & Li, T. (2021). Hybrid genetic model for clustering ensemble. *Knowledge-Based Systems*, *231*, 107457.

Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: a review with examples from ecology. *Ecological Monographs*, *93*(1), e1557.

Zangerle, E., & Bauer, C. (2022). Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, *55*(8), 1–38.

*name of corresponding author