CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 616-629
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Optimizing K-Means Clustering through Distance Metric Simulation for Strategic Enrollment Segmentation in Private Universities

Regita Putri Permata*, Amalia Nur Alifah, and I Made Wisnu Adi Sanjaya

Data Science, Telkom University, Surabaya Campus, Ketintang 156, 60231, Indonesia

Abstract

K-Means clustering is a widely used unsupervised learning technique to identify patterns and group data based on feature similarities. However, the effectiveness of K-Means significantly depends on the choice of distance metric. This study conducts a comprehensive simulation to evaluate and compare the performance of four distance metrics: Euclidean, Cityblock (Manhattan), Canberra, and Mahalanobis in the context of strategic market segmentation for private universities. The dataset includes simulated and institutional data incorporating variables such as account creation, registration, graduation, student performance (social, science, and scholastic scores), income, and geographic distance. The results indicate that Euclidean and Cityblock distances yield efficient and interpretable clusters with low computational costs, whereas the Mahalanobis distance, despite its capacity to model covariance, introduces computational overhead without proportional improvement in segmentation quality. Interestingly, Canberra distance produces compact clusters but does not offer significant gain in separability. From the resulting segmentation, two clusters emerge as high-potential targets for marketing strategies: Cluster 0 (high-income and distant students) and Cluster 1 (diverse academic and socioeconomic profiles). The findings highlight the importance of aligning distance metric selection with specific clustering objectives and offer practical insights for data-driven strategic enrollment planning in private higher education institutions.

Keywords: Distance Metrics, K-Means Clustering, Market Segmentation, Private Universities, Strategic Enrollment.

1 Introduction

Clustering of K-Means is one of the most popular data analysis methods used to group data based on similarities of characteristics [1], [2], [3]. In this method, the selection of appropriate distance metrics plays a crucial role in determining effective clustering results. Different distance metrics such as Euclidean, Manhattan, Canberra, Mahalanobis, and many more have different characteristics and can significantly affect clustering results [4], [5], [6]. The main problem in applying K-Means clustering is that the search results often do not show group members that meet the desired criteria. Several variables can affect the results of the experiment, including the type of data used, the amount of data, the type of distance measure applied, and the number of

*Corresponding author. E-mail: regitapermata@telkomuniversity.ac.id

Submitted: May 6, 2025 Reviewed: June 3, 2025 Accepted: June 24, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.33089

clusters considered most appropriate to simplify the information. An inappropriate choice of distance measure can result in the loss of relevant information or misgrouping of data, which in turn can reduce the effectiveness of marketing strategies developed based on the clustering results [7].

In the context of private universities, the use of varying distance measures in clustering can affect the market segmentation of prospective students. For example, the Euclidean distance measure may be more suitable for normally distributed data [8], while the Mahalanobis distance measure may provide more accurate results for data with varying variances [9]. The results of this study show that the use of Mahalanobis metrics is more effective with an increasing number of iterations [8]. Differences in clustering results can affect the identification of school groups that should be prioritized in marketing strategies. Euclidean distance performs less well than Mahalanobis and Canberra in the two tests mentioned in the previous study with average performance for smaller sample sizes [5], [8]. This affects the clustering results because Euclidean cannot recognize data with the same criteria. Research on [10] also compared the performance of Canberra, and the best cluster was found using SSE (Sum Square Error) and Silhouette Index. In further research, Manhattan distance is often used for facial recognition with a similarity level of up to 70% [11], and based on the resulting clusters, there are data that should not be in that cluster.

Despite the widespread application of K-Means clustering in data segmentation, previous studies have not sufficiently explored how different distance metrics influence clustering results in the context of higher education marketing [12]. Most existing research focuses on general-purpose clustering performance or technical domains, leaving a gap in understanding how the choice of distance metric affects the precision and strategic relevance of market segmentation for private universities. This study addresses that gap by systematically comparing the clustering outcomes of various distance metrics—Euclidean, Manhattan, Canberra, and Mahalanobis—using real-world prospective student data. The novelty of this research lies in its contextual application, evaluating clustering performance not only through traditional metrics such as SSE and Silhouette Index but also in terms of its practical implications for marketing decisions in private university settings. By identifying the most suitable distance metric for segmenting prospective students, this study contributes new insights that can support the development of more precise, data-driven marketing strategies to enhance institutional competitiveness in today's dynamic higher education landscape [12].

2 Methods

To achieve the research objective—comparing the effectiveness of various distance metrics in K-Means clustering—this section is organized into clearly defined subsections: the role of distance metrics in clustering, descriptions of the distance metrics used, and how these metrics are integrated into the K-Means algorithm for experimental analysis.

In clustering tasks, distance metrics are essential in determining the similarity or dissimilarity between observations. The selection of an appropriate metric significantly influences the formation of clusters, as it defines how data points relate to one another. The primary purpose of applying a distance or similarity function is to discover meaningful groupings based on the intrinsic structure of the data.

2.1 Role of Distance Metrics in Clustering

In clustering tasks, distance metrics are essential in determining the similarity or dissimilarity between observations. The selection of an appropriate metric significantly influences the formation of clusters, as it defines how data points relate to one another. The primary purpose of applying a distance or similarity function is to discover meaningful groupings based on the intrinsic structure of the data.

2.2 Distance Metrics Used in the Study

This study considers four well-known distance metrics, each with distinct characteristics and sensitivity to data scale, distribution, and feature correlation:

2.2.1 Euclidean Distance

Euclidean Distance measures the straight-line distance between two points in n-dimensional space. It is suitable for normally distributed, scaled data [13] be shown in Eq. 1.

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

where

- $d_E(x,y)$: Euclidean distance between points x and y
- x_i, y_i: The i-th component of vectors x and y
- n: Number of dimensions (features)

2.2.2 Manhattan Distance

Manhattan Distance calculates the absolute differences along each dimension, often used for non-normally distributed data or grid-based structures [14] can be shown in Eq. 2.

$$d_M(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

where

- $d_M(x,y)$: Manhattan distance between points x and y
- $|x_i - y_i|$: Absolute difference of the i-th component
- n: Number of dimensions (features)

2.2.3 Canberra Distance

Canberra Distance emphasizes smaller values by normalizing each dimension, making it sensitive to small variations and appropriate for sparse data [15] in Eq. 3.

$$d_C(x,y) = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{3}$$

where

- $d_C(x,y)$: Canberra distance between points x and y
- $|x_i - y_i|$: Absolute difference of the i-th component
- $|x_i| + |y_i|$: Sum of absolute values of the i-th components of x and y
- n: Number of dimensions (features)

2.2.4 Mahalanobis Distance

Mahalanobis Distance accounts for correlations among variables and is ideal for datasets with non-uniform variance and correlated features [16] shown in Eq. 4.

$$d_{Mah}(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)} \tag{4}$$

where

- $d_{Mah}(x, y)$: Mahalanobis distance between vectors x and y
- x, y: Data vectors (points)
- S: Covariance matrix of the dataset
- S^{-1}: Inverse of the covariance matrix
- $(x - y)^T$: Transpose of the difference vector

2.3 K-Means Clustering

K-Means is a popular and widely used clustering algorithm in data analysis [6]. The basic procedure of the K-Means algorithm [8] is as follows:

1. Initialization: Randomly select k initial centroids, where k is the number of clusters.
2. Assignment Step: Assign each data point to the nearest centroid based on the chosen distance metric, forming k clusters.
3. Update Step: Recalculate the centroids by computing the mean of all data points in each cluster.
4. Iteration: Repeat the assignment and update steps until the centroids no longer change significantly or a predefined number of iterations is reached (convergence).

2.4 Cluster Evaluation

2.4.1 Average Within Cluster Distance (AWCD)

The Average Within-Cluster Distance is a metric used to evaluate the compactness of clusters. It measures the average distance between all pairs of data points within the same cluster. Although Euclidean distance is most commonly used, other distance metrics may also be applied depending on the context. Mathematically, the AWCD for a cluster C_k is defined as shown in Eq. 5:

$$\text{AWCD}(C_k) = \frac{1}{|C_k|} \sum_{x_i, x_j \in C_k} d(x_i, x_j) \tag{5}$$

where $d(x_i, x_j)$ represents the distance between two data points within cluster C_k.

2.4.2 Dunn Index

Cluster evaluation is used to determine the quality of clustering results in the K-Means algorithm [17]–[20]. One such evaluation metric is the *Dunn Index*, which aims to identify well-separated and compact clusters. A higher Dunn Index indicates more optimal clustering results [21]. The Dunn Index is defined by Eq. 6:

$$D = \frac{\min\limits_{1 \le i < j \le k} d(C_i, C_j)}{\max\limits_{1 \le l \le k} d(C_l)} \tag{6}$$

where:

- $d(C_i, C_j)$: The inter-cluster distance between clusters i and j.
- $d(C_l)$: The intra-cluster distance within cluster l.

We also adopt a scalable version of this index called the *Scalable Dunn Index (S-DI)*, which implements a divide-and-conquer approach using the Spark framework to support large-scale data processing [20].

2.4.3 Silhouette Index

The *Silhouette Index* is used to evaluate the strength and quality of clusters by measuring how similar a data point is to its own cluster compared to other clusters [22]. The Silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters. The general formula for the Silhouette Coefficient is shown in Eq. 7:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{7}$$

where:

- $a(i)$: The average distance between data point i and all other points in the same cluster.
- $b(i)$: The average distance between data point i and all points in the nearest neighboring cluster.
- $s(i)$: The Silhouette score for data point i.

The higher the average Silhouette score across all data points, the better the clustering structure and separation between clusters.

2.5 Integration with K-Means Clustering

The selected distance metrics are integrated into the K-Means algorithm following a consistent experimental procedure to evaluate their effect on clustering quality. The overall workflow of the experimental process is illustrated in Figure 1, which outlines the steps from data initialization to the evaluation and visualization of clustering results.

1. Initialization: Generate or prepare the dataset to be clustered. Initialize required variables and define the number of iterations (e.g., $n = 100$) to ensure robustness of results. Determine the number of clusters k using the Elbow method, which identifies the optimal number of clusters by locating the point of inflection in the Sum of Squared Errors (SSE) curve.
2. Distance Calculation: For each selected distance metric (Euclidean, Manhattan, Canberra, and Mahalanobis), calculate the distance between each data point and the current cluster centroids.
3. K-Means Clustering: Apply the K-Means algorithm with $k = 4$ clusters using the respective distance metric. Assign each data point to the nearest centroid based on the calculated distance and update the centroid positions. Repeat this process until convergence is achieved.
4. Evaluation Metrics: After clustering, calculate evaluation metrics to assess the performance of each distance metric:
    - AWCD (Average Within-Cluster Distance): Measures compactness of clusters.
    - Dunn Index: Assesses the separation and compactness of clusters.
    - Execution Time: Measures the computational efficiency of the algorithm for each metric.
    - SSE (Sum of Squared Errors): Evaluates total intra-cluster variance.
    - Silhouette Index: Measures how similar an object is to its own cluster compared to other clusters.
5. Visualization: Present the results of the clustering process using appropriate visual tools to compare the performance of different distance metrics.

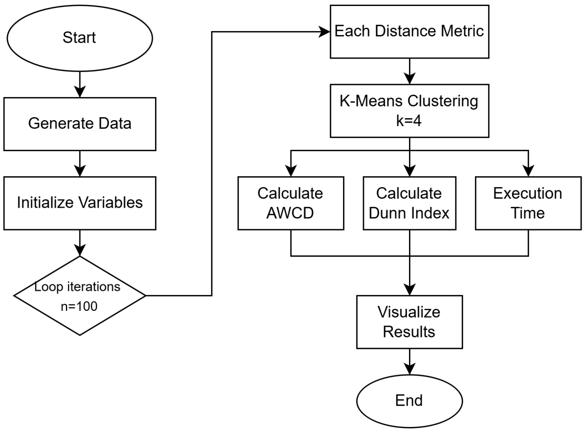


Figure 1: Flowchart K-Means Clustering

3 Results and Discussion

The results and discussion in this study compare four different distance measures based on the Average within Cluster Distance (AWCD) value, Dunn Index, and time execution using simulated data. The simulated data was created using a 150-row random normal.

```
x = np.random.normal(loc=np.repeat([1, 2, 3, 4], 4), scale=0.2, size=150)
y = np.random.normal(loc=np.repeat([3, 2, 1, 2], 4), scale=0.2, size=150)
```

The visualization of the simulation data is shown in Figure 2.

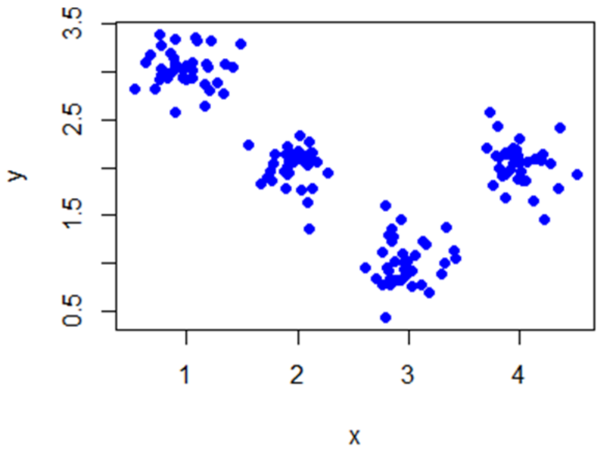


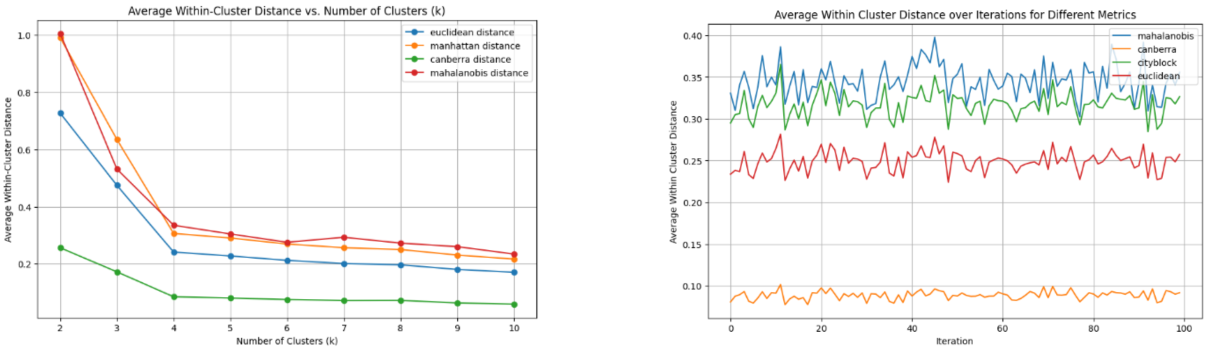
Figure 2: Visualize of Data Visualization

3.1 Comparison of Average Within Cluster Distance

The Average Within-Cluster Distance (AWCD) is a critical metric in K-Means clustering that measures the compactness of clusters by calculating the average distance between each point within a cluster and the cluster's centroid. A lower AWCD value indicates more compact clusters, meaning that the data points within each cluster are closely grouped around the centroid, which is desirable in most clustering applications. For each distance metric (Euclidean, Manhattan, Canberra, Mahalanobis), the AWCD will be calculated for each number of clusters k from 2 to 10. According to Figure 3a., it can be seen that the four distances show the sharpest decrease (elbow) at k=4, so the number of clusters set on the simulation data is k=4.

Figure 3b shows a comparison between four different distances with 100 iterations for AWCD calculation. Based on Average Within Cluster Distance (AWCD), the most effective metric is

Canberra, as it has the lowest and most stable average distance within a cluster compared to other metrics. A lower distance indicates that the data within a cluster is more homogeneous, resulting in better clusters. Mahalanobis (blue) and Cityblock (Manhattan) although used, show more volatile results and higher distances, making them less effective in maintaining cluster homogeneity.



(a) **(b)**

Figure 3: (a) Result of Average Within Cluster Distance (AWCD); (b) The comparison of AWCD by Distance

Based on the AWCD values obtained from different distance metrics (Euclidean, Manhattan, Canberra, Mahalanobis) on Table 1, Canberra Distance consistently produced the lowest AWCD values across various numbers of iterations. This indicates that using the Canberra Distance in K-Means clustering results in the most compact clusters, where data points are more tightly grouped around their centroids compared to other metrics.

Table 1: Average Within Cluster Distance (AWCD) for Different Metrics

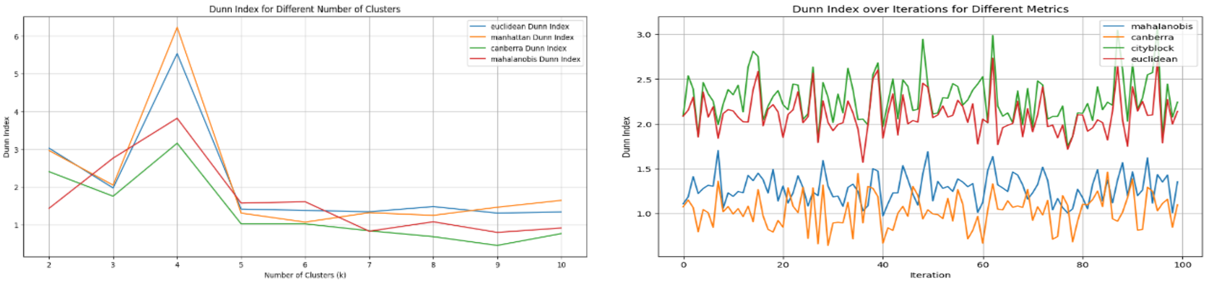
Metric	**AWCD**
Canberra	0.0880
Cityblock (Manhattan)	0.3158
Euclidean	0.2482
Mahalanobis	0.3446

3.2 Comparison of Dunn Index

The Dunn Index is used to evaluate the quality of clusters by measuring the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. Higher Dunn Index values generally indicate better clustering, as they suggest well-separated clusters with less variation within each cluster. Comparison of k=2 - 10, shows the maximum Dunn index is at k=4 for all distances shown in Figure 4a. Figure 4b shows a comparison between four different distances with 100 iterations for Dunn Index calculation.

Based on Figure 4b, the Cityblock (Manhattan) distance metric consistently achieves the highest Dunn Index values, indicating that it produces clusters that are both well-separated and compact. Euclidean distance also performs relatively well, though slightly less consistently than Cityblock. The Mahalanobis distance shows significant variability with generally lower Dunn Index values, suggesting that it may not be as effective for clustering in this context. The Canberra distance consistently results in the lowest Dunn index values, indicating that it is the least effective for achieving distinct and compact clusters. Comparison of average dunn index are presented in Table 2.

Optimum Distance Metric: Based on the Dunn Index in Table 2, Cityblock (Manhattan) Distance provides the highest Dunn Index. This suggests that Manhattan distance is most effective at separating clusters and minimizing intra-cluster distance at this number of clusters.



(a) **(b)**

Figure 4: (a) Dunn Index for Different Number of Clusters; (b) Dunn Index over Iterations for Different Metrics

Table 2: Dunn Index for Different Distance Metrics

Metric	Dunn Index
Canberra	1.0291
Cityblock (Manhattan)	2.2907
Euclidean	2.1045
Mahalanobis	1.2828

3.3 Silhouette Index

The graph shows the Silhouette Index over 100 iterations for four different distance metrics: Mahalanobis, Canberra, Cityblock (Manhattan), and Euclidean. The silhouette index measures how well separated the clusters are; higher values indicate better-defined clusters.

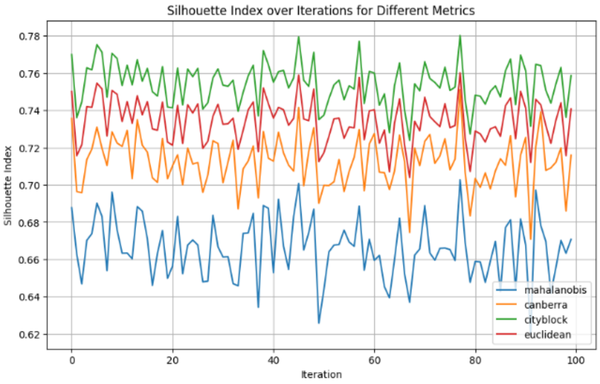


Figure 5: Silhoutte Index Over Iterations for Different Metric

Based on Figure 5, The Silhouette Index analysis reveals that the Cityblock (Manhattan) distance metric consistently produces the most effective clustering results, followed closely by the Euclidean metric.

Table 3: Average Iteration of Silhouette Index Results

Metric	Silhouette Index
Canberra	0.7126
Cityblock (Manhattan)	0.7539
Euclidean	0.7338
Mahalanobis	0.6654

If a comparison is made through the average Silhoutte value in the Table 3, it shows that the average value of the Cityblock (Manhattan) distance is higher than the other distances. Overall, Cityblock emerges as the optimal metric for clustering in this scenario.

3.4 Comparison of Execution Time

When evaluating the efficiency of the clustering algorithm, execution time is a critical factor, especially with large datasets or in real-time processing scenarios. The difference distant metric used within these algorithm can significantly impact the overall execution time, directly affecting their practicality and performance. Comparison of execution time shown in Figure 6.

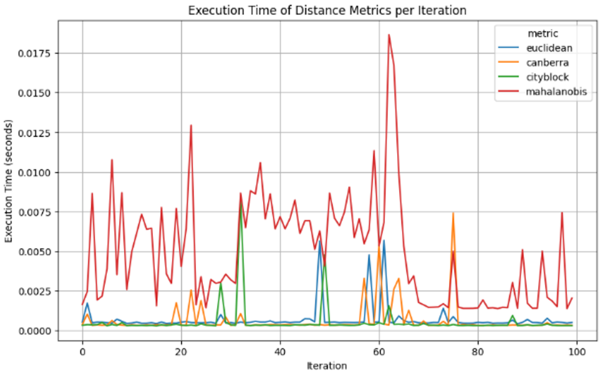


Figure 6: Execution Time of Distance Metrics per Iteration

The graphical from Figure 6 representation of execution times further supports the findings from the Table 4, the Mahalanobis distance consistently shows higher spikes in execution time across multiple iterations, reflecting its computational intensity. Meanwhile, the Canberra, Cityblock (Manhattan), and Euclidean metrics remain relatively stable and efficient throughout the iterations, with Cityblock (Manhattan) being the most efficient.

Table 4: Average Iteration of Execution Time Results

Metric	Execution Time (Second)
Canberra	0.000643
Cityblock (Manhattan)	0.000519
Euclidean	0.000700
Mahalanobis	0.004913

Based on average of Execution time in Table 4, shows that Mahalanobis distance stands out as the slowest, with an average execution time of 0.004913 seconds. This higher execution time is likely due to the complex computation involved in calculating the covariance matrix, which is a key component of the Mahalanobis metric. Overall cityblock (Manhattan) is the distance with the most efficient execution time. Figure 7 is a visualization of K-means clustering with Manhattan distance on simulated data.

Result of the analysis of clustering metrics reveals that the Euclidean and Cityblock distances provide high clustering quality, with strong Dunn and Silhouette Index scores, and efficient execution times. The Canberra Distance, while yielding the most compact clusters with the lowest AWCD values, struggles with separation and cohesion, and the Mahalanobis distance, despite its ability to handle data correlations, incurs high computational costs. Overall, Cityblock (Manhattan) and Euclidean metrics offer a practical balance of performance and efficiency, making them preferable for most clustering tasks.

3.5 Strategic Market Segmentation Private Universities Data

For university segmentation strategy, employed with using K-Means utilizing the Manhattan distance metric. In this case, K value of 4 was determined for this grouping. This data uses quantity factors (accounts, registrations, graduates, and PINs), quality factors (social, science,

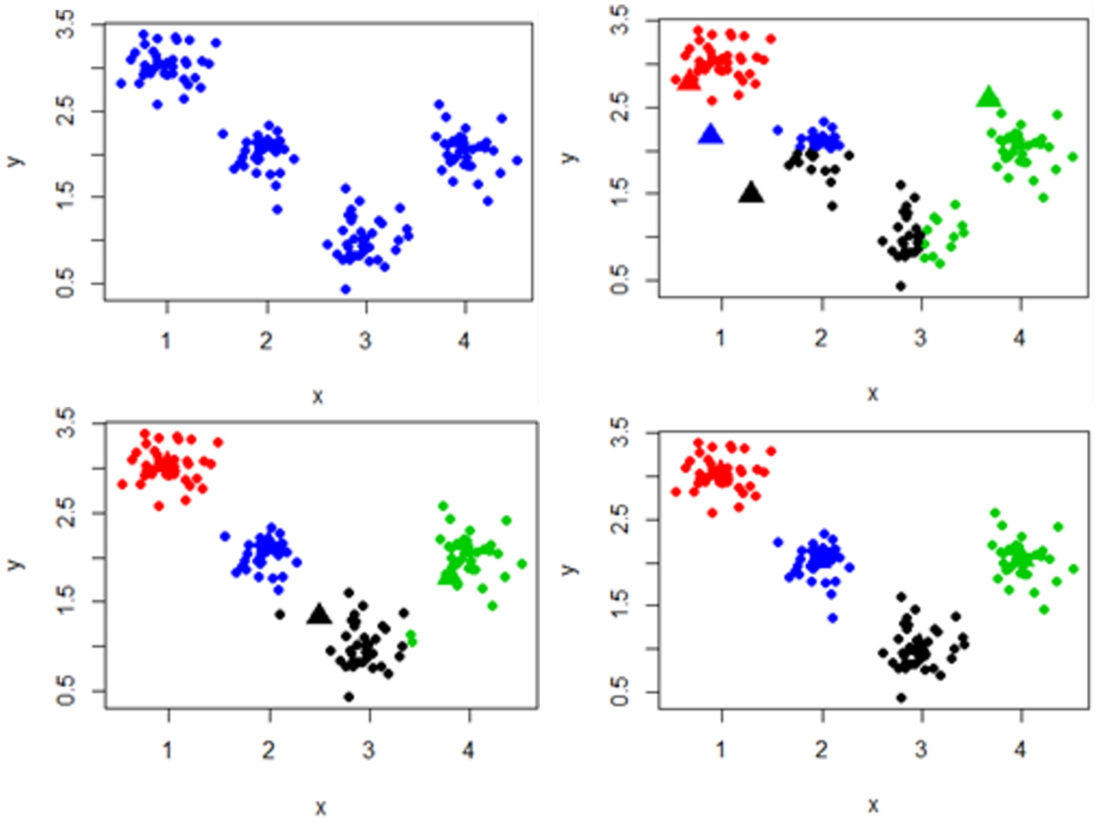


Figure 7: visualization of K-means clustering with Manhattan distance on simulated data.

and scholastic values), and additional components, namely income and distance. Table 5 presents the calculated centroid for this K-Means analysis.

Table 5: Average Statistics per Cluster

Cluster	**0**	**1**	**2**	**3**
Income (mean)	1.61	2.04	1.61	2.07
Account	11.79	5.96	34.13	7.37
Distance	286.01	772.90	116.42	466.81
Passed	2.55	0.88	8.55	1.58
PIN	1.15	0.53	5.84	0.77
Register	519.65	514.16	515.84	514.77
Science and Technology	554.88	543.18	543.36	540.67
Scholastic	543.14	543.23	534.83	536.50

The calculated centroids from the K-Means analysis, resulted in the groupings visualized in Figure 8. This figure, a boxplot representation, illustrates the distribution of each variable within the formed clusters.

Cluster 0 has the highest number of accounts and registrations, with a moderate average income and high variation in PIN values. Cluster 1 shows a higher average income but lower accounts and registrations compared to Cluster 0. Cluster 2 has the highest median distance and low accounts and registrations, indicating a geographically dispersed group with less activity. Cluster 3 has the lowest average income and low PIN values, suggesting a segment that may be less economically advantaged.

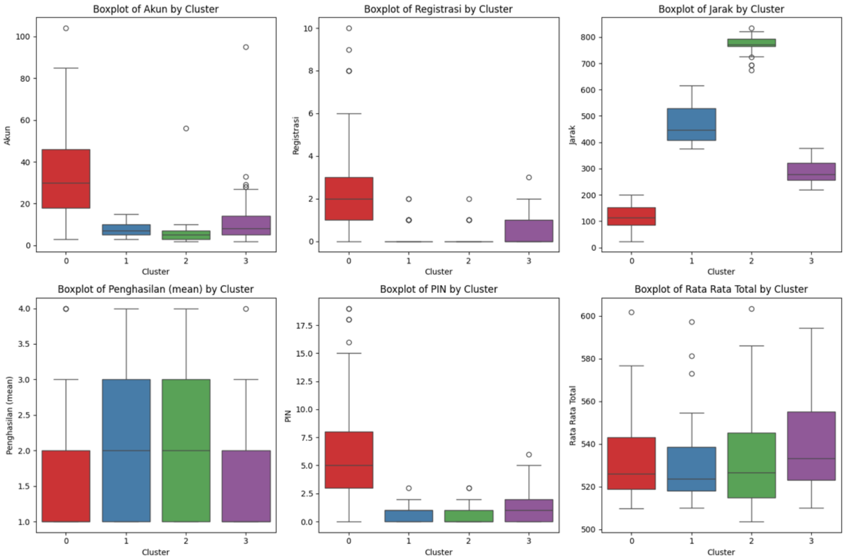


Figure 8: Boxplot of the results of Clustering

3.6 Discussion

The analysis of clustering performance across Euclidean, Cityblock (Manhattan), Canberra, and Mahalanobis provides key insights into their effectiveness and efficiency. The Dunn Index and Silhouette Index both indicate that Euclidean and Cityblock (Manhattan) metrics perform well, with higher values reflecting better-defined clusters that are both compact and well-separated. In contrast, the Mahalanobis distance, despite accounting for data correlations, and Canberra distance tend to show lower Dunn and Silhouette scores, suggesting challenges in maintaining cluster separation and cohesion. Specifically, the Canberra Distance consistently produced the lowest Average Within-Cluster Distance (AWCD) values across various iterations, indicating that it results in the most compact clusters with data points tightly grouped around their centroids. However, the Mahalanobis distance incurs significantly higher execution times due to its complexity, making it less practical for large datasets. In contrast, the Cityblock and Euclidean metrics not only perform well in clustering quality metrics but also maintain lower execution times, thus offering a balanced choice for efficiency and effectiveness. Ultimately, the selection of a distance metric should align with the specific needs of the task, including considerations for computational efficiency, clustering quality, and data characteristics.

To develop a more effective market segmentation strategy for Private Universities, targeting Cluster 0 and Cluster 1 offers the most strategic advantages. Cluster 0 consists of highly engaged prospective students with moderate income and closer proximity to university locations, making them ideal for campaigns focused on accessibility and affordability. Cluster 1 includes a more affluent group with a larger distance from the university, suitable for premium programs that emphasize exclusivity and quality. This cluster also exhibits the most significant variation in accounts and PINs, indicating a diverse student base that can be tapped through tailored marketing strategies. By combining the strengths of both clusters, universities can create holistic segments that appeal to a broad range of potential students, driving both high-volume enrollments and premium offerings.

4 Conclusion

In summary, the analysis of different distance metrics for clustering, Euclidean, Cityblock (Manhattan), Canberra, and Mahalanobis, shows that Euclidean and Cityblock metrics are the most effective, providing well-defined clusters with good separation and low execution times. Although Canberra distance creates the most compact clusters, it does not improve overall clustering quality, and Mahalanobis distance, while accounting for data correlations, is less efficient due to higher computation time.

For Private University market segmentation, Cluster 0 and Cluster 1 are the most promising targets. Cluster 0, with high-income and distant students, is ideal for premium programs, while Cluster 1's diversity offers potential for broader marketing strategies. The choice of distance metric should consider both clustering quality and computational efficiency, tailored to the specific needs of the task.

CRediT Authorship Contribution Statement

Regita Permata: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing–Original Draft Preparation, Writing–Review & Editing, Visualization, Supervision, Funding Acquisition. **Amalia Nur Alifah:** Investigation, Data Curation, Resources, Formal Analysis, Writing–Review & Editing. **Wisnu:** Writing–Review & Editing, Visualization, Layout Editing.

Declaration of Generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

Thank you to LPPM Telkom University provided funding for this research. We are sincerely grateful for the financial support and assistance that they have provided. Activity through Assignment Letter No. KWR4.030/LIT06/PPM-LIT/2024.

Data Availability

The data supporting the findings of this study include simulated data and internal admissions data from Telkom University. Due to institutional confidentiality policies and privacy considerations, these data are not publicly available. However, the data may be made available by the corresponding author upon reasonable request and subject to approval by Telkom University and relevant confidentiality agreements.

References

[1] A. Abdulhafedh, "Incorporating k-means, hierarchical clustering and pca in customer segmentation," *Journal of Computer and Data Sciences*, vol. 3, no. 1, pp. 12–30, Feb. 2021. DOI: 10.12691/jcd-3-1-3.

[2] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 1, 2023. DOI: 10.1016/j.ins.2022.11.139. Available online (visited on 08/20/2024).

[3] S. Abadi, K. S. M. The, B. M. Nasir, *et al.*, "Application model of k-means clustering: Insights into promotion strategy of vocational high school," *International Journal of Engineering & Technology*, vol. 7, no. 2, pp. 182–187, Aug. 22, 2018, Number: 2.27. DOI: 10.14419/ijet.v7i2.11491. Available online (visited on 02/13/2024).

[4] M. Zubair, M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, "An improved k-means clustering algorithm towards an efficient data-driven modeling," *Annals of Data Science*, Jun. 25, 2022. DOI: 10.1007/s40745-022-00428-2. Available online (visited on 08/21/2024).

[5] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative analysis of inter-centroid k-means performance using euclidean distance, canberra distance and manhattan distance," *Journal of Physics: Conference Series*, vol. 1566, no. 1, p. 012 112, Jun. 2020, Publisher: IOP Publishing. DOI: 10.1088/1742-6596/1566/1/012112. Available online (visited on 08/21/2024).

[6] S. Suraya, M. Sholeh, and D. Andayati, "Comparison of distance metric in k-mean algorithm for clustering wheat grain datasheet," *Jurnal Teknik Informatika C.I.T Medicom*, vol. 15, no. 2, pp. 73–83, May 31, 2023, Number: 2. DOI: 10.35335/cit.Vol15.2023.408.pp73-83. Available online (visited on 08/21/2024).

[7] D. Jollyta, P. Prihandoko, D. Priyanto, A. Hajjah, and Y. N. Marlim, "Comparison of distance measurements based on k-numbers and its influence to clustering," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 1, pp. 93–102, Nov. 2023. DOI: 10.30812/matrik.v23i1.3078. Available online.

[8] A. Farid, E. Pavel, G. Marina, P. Irina, and M. Sergey, "Clustering of k-means based on euclidean distance metric and mahalanobis metric," 2024. DOI: https://doi.org/10.1051/e3sconf/202453103002. Available online.

[9] H. Ghorbani, "Mahalanobis distance and its application for detecting multivariate outliers," *Facta Universitatis, Series: Mathematics and Informatics*, vol. 34, no. 3, pp. 583–595, 2019. DOI: 10.22190/FUMI1903583G. Available online.

[10] M. Raeisi and A. B. Sesay, "A distance metric for uneven clusters of unsupervised k-means clustering algorithm," *IEEE Access*, vol. 10, pp. 86 286–86 297, 2022, Conference Name: IEEE Access. DOI: 10.1109/ACCESS.2022.3198992. Available online (visited on 08/21/2024).

[11] Sunardi, A. Fadlil, and N. Tristanti, "Comparative analysis of euclidean, manhattan, canberra, and squared chord methods in face recognition," *Revue d'Intelligence Artificielle*, vol. 37, no. 3, pp. 593–599, 2023. DOI: 10.18280/ria.370308. Available online.

[12] M. Marhayati, A. M. Fa'ani, S. U. Ruhmanasari, and S. Faridah, "Application of k-means cluster analysis for grouping state islamic university in indonesia based on the readiness indicators for world class university (wcu)," *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 8, no. 2, pp. 30–48, 2023, Licensed under CC-BY-SA 4.0; Published online 2023. DOI: 10.18860/ca.v8i2.18046. Available online.

[13] T. S. Madhulatha, *An overview on clustering methods*, arXiv:1205.1117, 2012. DOI: 10.48550/arXiv.1205.1117. arXiv: 1205.1117 [cs.DS]. Available online.

[14] S. A. Afghani and W. Y. M. Putra, "Clustering with euclidean distance, manhattan - distance, mahalanobis - euclidean distance, and chebyshev distance with their accuracy," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 369–376, 2021. DOI: 10.29244/ijsa.v5i2p369-376. Available online.

[15] A. F. Pulungan, M. Zarlis, and S. Suwilo, "Analysis of braycurtis, canberra and euclidean distance in knn algorithm," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 4, no. 1, pp. 74–77, 2019. DOI: 10.33395/sinkron.v4i1.10207. Available online.

[16] B. H. Russell and L. R. Lines, "Mahalanobis clustering, with applications to AVO classification and seismic reservoir parameter estimation," vol. 15, 2003. Available online.

[17] R. Mohemad, N. N. M. Muhait, N. M. M. Noor, and Z. A. Othman, "Performance analysis in text clustering using k-means and k-medoids algorithms for malay crime documents," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 5014–5026, Oct. 1, 2022, Number: 5. DOI: 10.11591/ijece.v12i5.pp5014-5026. Available online (visited on 08/21/2024).

[18] S. Suraya, M. Sholeh, and U. Lestari, "Evaluation of data clustering accuracy using k-means algorithm," *International Journal of Multidisciplinary Approach Research and Science*, vol. 2, no. 1, pp. 385–396, 2024, Number: 01. DOI: 10.59653/ijmars.v2i01.504. Available online (visited on 08/21/2024).

[19] A. O. Oluwatobi, V. A. Alli-Johnson, C. A. Ayedun, and O. A. Akinjare, "Assessment of the effectiveness of maintenance management systems in delivering quality maintenance services in higher institutions," in *Journal of Physics: Conference Series*, vol. 1299, IOP Publishing, 2019, p. 012015. DOI: 10.1088/1742-6596/1299/1/012015.

[20] M. Paramadina, S. Sudarmin, and M. K. Aidid, "Perbandingan analisis cluster metode average linkage dan metode ward (kasus: Ipm provinsi sulawesi selatan)," *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, vol. 1, no. 2, pp. 22–31, 2019. DOI: 10.35580/variansiunm9357. Available online.

[21] H. Malikhatin, A. Rusgiyono, and D. A. I. Maruddani, "Clustering of prospective tki workers using k-modes algorithm and dunn index for categorical data," *Jurnal Gaussian*, vol. 10, no. 3, pp. 359–366, 2021, Published: 2021-12-30, Accepted: 2021-08-31. DOI: 10.14710/j.gauss.10.3.359-366. Available online.

[22] C. Ais, A. Hamid, and D. C. R. Novitasari, "Analysis of livestock meat production in indonesia using fuzzy c-means clustering," *Jurnal Ilmu Komputer dan Informasi*, vol. 15, no. 1, pp. 1–8, 2022. DOI: 10.21609/jiki.v15i1.993. Available online.