Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14052
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Performance Single Linkage and K-Medoids on Data with Outliers

Caecilia Bintang Girik Allo[1)*], Winda Ade Fitriya B[2)], Nicea Roona Paranoan[3)]
[1,2,3)]Universitas Cenderawasih, Indonesia
[1)]caecilia.bintang@fmipa.uncen.ac.id, [2)]windaafb97@gmail.com, [3)]nicearoona12@gmail.com

Submitted :Aug 29, 2024 | **Accepted** : Sept 12, 2024 | **Published** : Oct 2, 2024

Abstract: One way to assess the economic growth of a province is by examining its Gross Regional Domestic Product (GRDP). GRDP calculated through the production approach reflects the total value added by goods and services from various sectors within a particular region over a specified period. To determine the GRDP, 17 business sectors are considered. In 2023, the GRDP growth rate in Papua has decreased to 3.44%, down from 4.11% the previous year. To help the government improve Papua's GRDP, an analysis is required. Clustering methods can group regencies and cities with similar characteristics. Boxplots are used to identify outliers in the data. The data contains outliers, so one method that can be used is K-Medoids. Euclidean Distance is used to calculate the distance matrix. Before calculating the distances, standardization using z-score normalization is performed to ensure that the data ranges are the same. This article aims to identify the most effective method for clustering regencies and cities in Papua using GRDP at constant price data. Both Single Linkage and K-Medoids methods are applied in this study. The DBI is used for evaluation, with lower DBI values indicating better methods. According to the DBI results, Single Linkage outperforms K-Medoids for clustering regencies and cities in Papua, with the optimal number of clusters being three.

Keywords: Euclidean Distance; Davies Bouldin Index (DBI); Gross Regional Domestic Bruto; K-Medoids; Single Linkage; z-score Normalization

INTRODUCTION

Several efforts have been made by the government to improve the economic growth in Indonesia. The economy in Indonesia is closely tied to the economic growth of its provinces. One of the indicators to measure the economic growth of provinces in Indonesia is by observing the Gross Regional Domestic Product (GRDP).

There are three approaches to calculate GRDP, namely production approach, expenditure approaches, and income approach. Gross Regional Domestic Product (GRDP) using the production approach represents the total value added from goods and services produced by different sectors within a specific region over a defined timeframe. GRDP using the production approach is produced by 17 business sectors.

GRDP is presented in two forms: GRDP at current prices and GRDP at constant prices. GRDP at current (nominal) prices reflects the economic output capacity generated by a region. GRDP at constant (real) prices can be used to indicate the overall economic growth rate or the growth of individual sectors from year to year.

As a newly expanded province, the Papua Provincial Government certainly requires an analysis that can enhance its GRDP. In 2022, the GRDP growth rate at constant prices in Papua Province reached 4.11%. However, in 2023, there was a decline in the GRDP growth rate at constant prices, dropping to only 3.44% (BPS Provinsi Papua, 2024). When viewed from the GRDP value at constant prices, the increase from 2022 to 2023 was not as high as the rise from 2021 to 2022.

In this case, clustering analysis can be used to group regencies/cities with similar characteristics, making it easier for the government to formulate policies. Broadly speaking, the concept of clustering is that each member within a cluster shares similar characteristics, while members across different clusters have distinct characteristics. Various clustering methods have been applied to different types of data.

In general, there are two types of clustering methods: hierarchical and non-hierarchical methods. Fathia *et al*. (2016) compare Ward's and Single Linkage on variables that represent village potential in Semarang Regency. The result show that Single Linkage is better than Ward's method. Syafiyah *et al*. (2020) conducted a comparison of Hierarchical and Non-Hierarchical methods using employment indicator data from West Java. The authors use Single Linkage, Average Linkage, and Complete Linkage for the hierarchical method and K-Means for the non-hierarchical method. The study found that Hierarchical Clustering performed better.

One of famous algorithm in non-hierarchical method is K-Means. Sapriyanti & Rianto (2020) compare K-Means and Single Linkage to clustering agent contact center of a company. The result show that K-Means is better

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14052
e-ISSN : 2541-2019
p-ISSN : 2541-044X

than Single Linkage. Wororomi *et al*. (2023) compare K-Means and DBSCAN to clustering provinces in Indonesia. The authors use silhouette coefficient to show the performance of K-Means and DBSCAN. The result is K-Means better than DBSCAN. Nurjannah *et al*. (2024) use K-Means to clustering the child growth and development. The analysis result is 38 samples fall into three clusters. Wardani *et al*. (2024) also apply the silhouette coefficient to demonstrate the effectiveness of K-Means in clustering regency/city in North Sumatra based on poverty indicators.

In reality, there may be outliers in the data, which can make K-Means ineffective (Han & Kambers, 2006). K-Medoids is one of the clustering algorithms that can be used for data with outliers (Kauffman & Rousseeuw, 1990). Nahdliyah *et al*. (2019) applied the K-Medoids method to group regency/city in Central Java according to the number of criminal cases. The K-Medoids method also was used by Ibrahim *et al*. (2020) to clustering villages/sub-districts in Kutai Kartanegara Regency according to village potential data.

Based on previous research, this study aims to compare the Single Linkage and K-Medoids methods based on GRDP at constant price in 2023 in Papua. The effectiveness will be show in Davies Bouldin Index (DBI). At the end, this study will produce clustering of regency/city in Papua and hope this result will be useful for the government in making policy to ensure continuous GDP growth in Papua.

LITERATURE REVIEW

Several studies about clustering various dataset have been done. Reinaldi *et al*. (2021) compare three algorithms within the hierarchical method for community welfare data in East Java. The three algorithms are Single Linkage, Complete Linkage, and Average Linkage. There are five variables and 38 regencies/cities. As an initial step, the authors test the adequacy of the data. The evaluation method used by the authors is the silhouette coefficient. The highest silhouette value is found with three clusters using the average linkage method. Thamrin & Murni (2022) use Single Linkage to clustering regencies/cities in West Sumatera based on health indicator. The authors performs normalization using z-score. The distance matrix is calculated using the Manhattan distance formula. To determine the characteristics of each cluster, the authors computes the average of each variable. The number of clusters is determined based on the author's subjective judgment. Ulvi & Ikhsan (2024) compare K-Means and K-Medoids in Export and Import of Goods data in Indonesia. The authors use data from 2021 until 2023. There are 1176 data with categories of 98 goods. The authors use Elbow Method to determine number of clusters. The authors use DBI for evaluation. The result showed that K-Means was better than K-Medoids. Insani et al. (2024) use K-Medoids to clustering COVID-19 Patients. The authors used patient data from March 2020 until September 2021. Total data is 916 patients. The authors also show the DBI and silhouette coefficient for evaluation.

Several studies about clustering GRDP have been done. Febriyati *et al*. (2020) use K-means to clustering GRDP Growth Rate based on business field in Surabaya. The authors use GRDP Growth Rate in 2017 – 2019. The steps of the study are normalize data, determine number of cluster, modelling, and conclusion. The authors choose to make 3 clusters, namely high cluster, medium cluster, and low cluster. The authors initialized the centroids for the first iteration by setting the maximum value for the high cluster, the average value for the medium cluster, and the minimum value for the low cluster. Then did the next iteration with K-Means algorithm. Davies Bouldin Index (DBI) is used to evaluate the result. The DBI of 3 clusters is -0.712.

Another case of clustering has been done by Ningrum and Ahadi in 2022. Ningrum & Ahadi (2022) use K-Means to clustering regency/city in East Java based on GRDP Growth Rate. The authors also use 17 business field same with previous study mentioned. The steps of the study are descriptive analysis, modelling, evaluation. The authors used data of GRDP Growth Rate in 2021. The authors also use silhouette coefficient to determine the optimal cluster. There are 38 regencies/cities in East Java clustered. Based on silhouette coefficient, the optimal cluster was 2 clusters. One of the result is Cluster 1 was stronger than Cluster 2 during the COVID-19 period.

The gap with the studies mentioned earlier is that this study will consider the outliers in GRDP data with using K-Medoids, which has not been employed in previous studies in GRDP data. This study will evaluate Single Linkage and K-Medoids use Davies Bouldin Index (DBI). From the DBI, optimal number of clusters will be obtained.

METHOD

This research use data GRDP at constant price in 2023 based on 17 business field in Papua. There are 9 regencies/cities in Papua. The data is obtained from publications by BPS. The steps in this study is enhanced with the steps from Febriyati *et al*. (2020) and Ningrum & Ahadi (2022). In this study, we will add another descriptive to see the outliers of the data. Figure 1 shows the steps in this study.

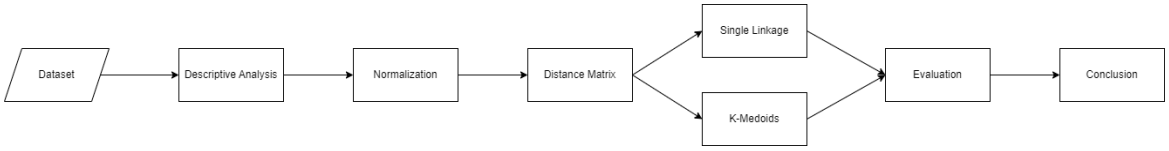

Fig. 1 The Study Stages

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14052
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In descriptive analysis, the data will shows in some graphics and boxplot to see outliers. After completing the descriptive analysis, data normalization will be performed using the z-score normalization method. The formula of z-score normalization can be seen in (1).

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \qquad (1)$$

where:
$i = 1, 2, \dots, n$; $j = 1, 2, \dots, p$
n : number of observation
p : number of variables
x_{ij} : observation i on jth variable
$\bar{x}_j$: mean of jth variable
σ_{ij} : standard deviation jth variable

The next step is to create a distance matrix. The distance matrix will be constructed using the Euclidean distance. The formula of Euclidean distance can be seen in (2).

$$d_{im} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{mj})^2} \qquad (2)$$

where:
d_{im}: Euclidean distance between observation ith observation and mth
x_{ij}: observation i on variable j
x_{mj}: observation m on variable j

After completing the distance matrix, the model will be applied. There are 2 models will be applied on the data. The first is Single Linkage Method. There are 4 steps in Single Linkage. **Step 1** is to create the distance matrix. **Step 2** is to determine the smallest distance and merge the two observations or objects into a single cluster. For instance, if the objects with the smallest distance are *U* and *V*, combine *U* and *V* into one cluster (*UV*). **Step 3** is to update the distance matrix by calculating the distances involving the newly merged cluster using (3). **Step 4** is repeat steps 2 and 3 until all objects are merged into a single cluster.

$$d_{UV} = \min\{d_{UW}, d_{VW}\} \qquad (3)$$

Another model is K-Medoids. K-Medoids is also known as Partitioning Around Medoids (PAM). There are 7 steps in this algorithm. **Step 1** is to determine number of cluster, namely k. **Step 2** is to randomly select medoids for each cluster. **Step 3** is to Calculate the distance of each object to each medoid in the clusters using (2). **Step 4** is to determine cluster membership by selecting the cluster with the minimum distance from each object. **Step 5** is to calculate the Total Cost. Total Cost is obtained by summing the minimum distances selected for all clusters. The initial Total Cost is denoted by TC_o. **Step 6** is to select new candidate medoids randomly, as in step 2. Then, do step 3 and step 4. The updated Total Cost is denoted by TC_n. **Step 7** is to calculate the difference in Total Cost, namely S where $S = TC_n - TC_o$. If $S > 0$, then the iteration stops, and the cluster memberships used are those from the calculation of TC_o. If $S > 0$, then do step 6 and it should be noted that TC_n in step 6 become TC_0 and TC_n is obtain from the new medoid determination.

The result from each algorithm will be evaluated by Davies Bouldin Index (DBI). The DBI approach aims to maximize the distance between different clusters while minimizing the distance between objects within the same cluster. The lowest DBI value indicates that the clustering in data mining has high quality, with the formed clusters being significantly distinct from each other (Warisa & Nurahman, 2023). After compare the DBI of both of the algorithm, then the conclusion will be obtained.

RESULT

For the first, the authors do some descriptive analysis. Fig. 1 shows top 3 regencies/cities with the highest values for each variable. From Fig. 1, we can see that either Kota Jayapura or Jayapura become the lead of each variable. Kota Jayapura and Jayapura dominate the top three. Kota Jayapura become the lead for every variable except in Agriculture, Forestry and Fishing; Mining and Quarrying; and Transportation and Storage. When Kota Jayapura is not the highest, Jayapura holds that position.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14052
e-ISSN : 2541-2019
p-ISSN : 2541-044X

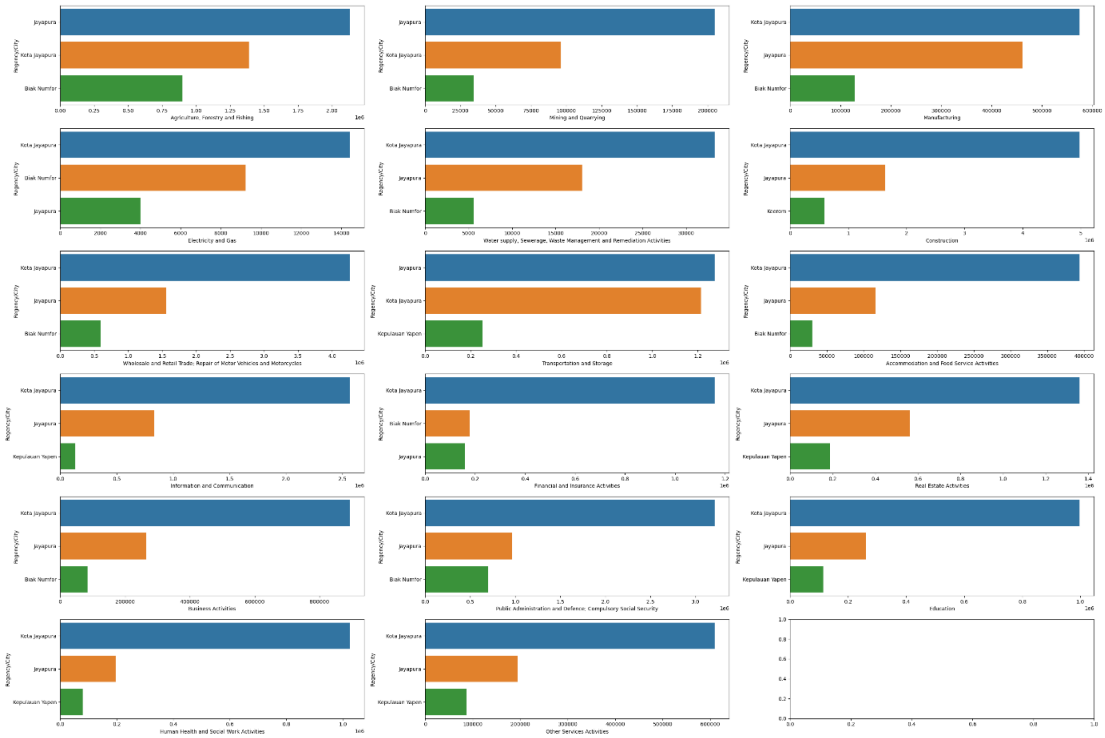

Fig. 1. Top 3 Regencies/Cities with The Highest Values for Each Variable

Fig. 2 shows that Mamberamo Raya does not contribute to Electricity and Gas; Water supply, Sewerage, Waste Management and Remediation Activities; Information and Communication. There are 3 regencies/cities do not contribute to Water supply, Sewerage, Waste Management and Remediation Activities.

Another descriptive analysis is in Fig. 3. Boxplot can be used to see outliers in data. From Fig. 3, we can see that in every variable there is outlier. Next, normalization is performed using z-score normalization, followed by distance calculation using Euclidean Distance.

Next, modeling is performed using Single Linkage and K-Medoids, and the DBI (Davis-Bouldin Index) is calculated for each possible number of clusters. Based on Table 1 shows that the lowest DBI value for Single Linkage is at 2 clusters, while for K-Medoids it is at 8 clusters. Fig. 4 is the mapping for the optimal cluster, that is 3 cluster from Single Linkage.

Fig. 2. Top 3 Regencies/Cities with The Lowest Values for Each Variable

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14052
e-ISSN : 2541-2019
p-ISSN : 2541-044X

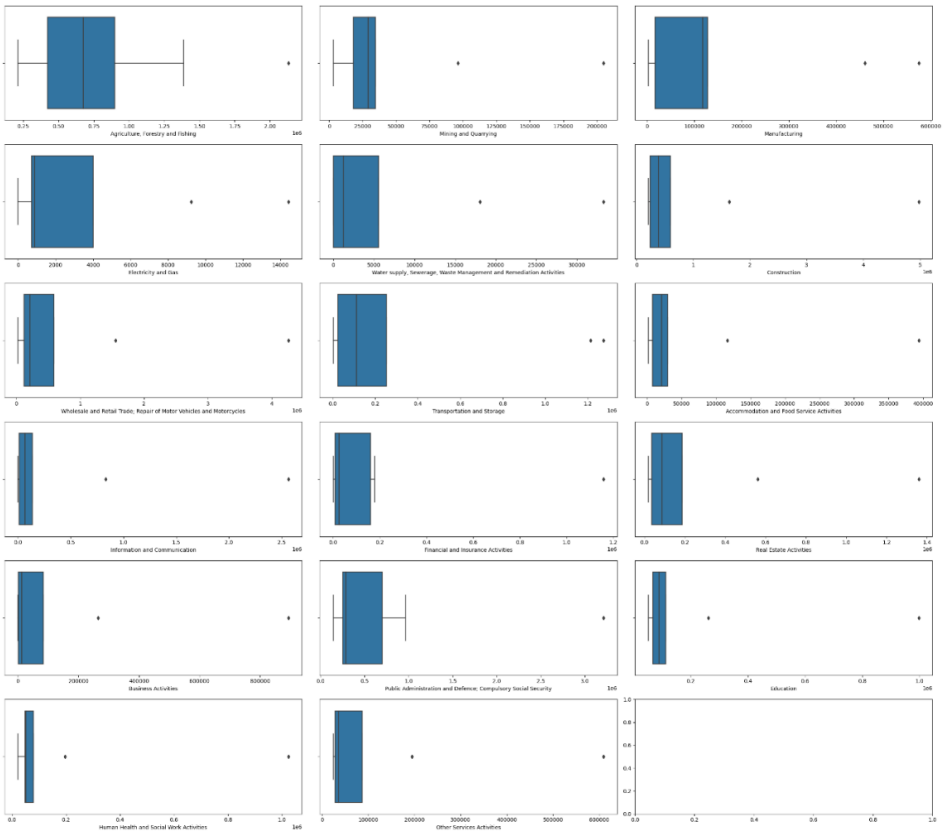

Fig. 3. Boxplot for Every Variable

Table 1. DBI for Single Linkage and K-Medoids

Number of Cluster	DBI Single Linkage	DBI K-Medoids
2	0.148578	0.614225
3	0.116125	0.725873
4	0.162401	0.728958
5	0.206816	0.668633
6	0.238390	0.592863
7	0.341246	0.561859
8	0.241716	0.522882

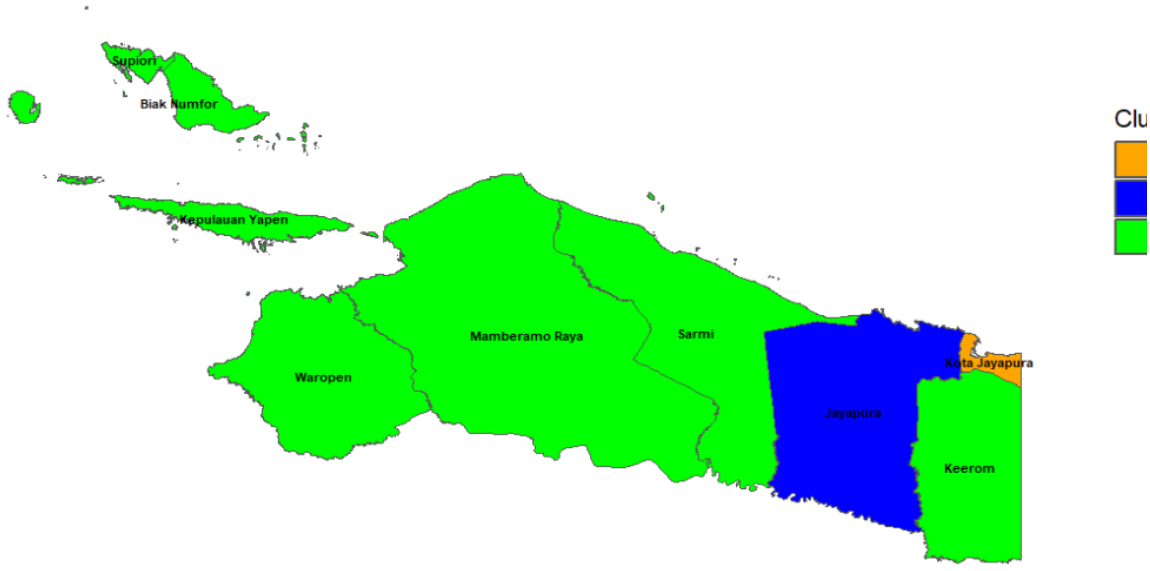

Fig. 4. Mapping of The Result Cluster

*name of corresponding author

DISCUSSIONS

The data consists of 9 Regencies/Cities and 17 variables, where the variables come from 17 business sectors. The data is sourced from publications by BPS. The results of the Boxplot show that there are outliers in the data, therefore K-Medoids is used in cluster modeling.

Suraya & Wijayanto (2022) compare hierarchical clustering, K-Means, K-Medoids, and Fuzzy C-Means to clustering provinces in Indonesia based on special index for handling stunting. The result show that Average Linkage is the best cluster method. Alamtaha *et al*. (2023) compare Single Linkage, Complete Linkage, K-Means, and K-Medoids use DBI to clustering the users of media social. The result show complete linkage has the lowest DBI. Based on the two previous studies, it is shown that the hierarchical method is better than the others. Therefore, this study chooses one algorithm from the hierarchical methods, namely Single Linkage.

Based on the DBI values for all possible clusters in Single Linkage and K-Medoids, it was found that the smallest DBI value is obtained with the Single Linkage method with 3 clusters with DBI value 0.116125. Cluster 1 consists of Kota Jayapura. Cluster 2 consists of Jayapura. Cluster 3 consists of Kepulauan Yapen, Biak Numfor, Sarmi, Keerom, Waropen, Supriori, and Memberamo Raya.

CONCLUSION

The decline in GRDP growth rates and the provincial expansion in Papua have led the government to require an analysis that can group regencies/cities in Papua Province with similar characteristics based on 17 business sectors. One analysis that can be used is cluster analysis. The cluster methods used are Single Linkage and K-Medoids. DBI is used as a method to evaluate both algorithms. The result shows Single Linkage is better than K-Medoids to clustering GRDP at constant price in Papua. The optimal number of clusters is 3. Cluster 1 has one regency/city. Cluster 2 has 1 regency/city. Cluster 3 have 7 regencies/cities.

ACKNOWLEDGMENT

This journal article is written based on research funded by the Research and Community Service Institution of Cenderawasih University in 2024. The content of the journal is the sole responsibility of the author.

REFERENCES

Alamtaha, Z., Djakaria, I., & Yahya, N. I. (2023). Implementasi Algoritma Hierarchical Clustering dan Non-Hierarchical Clustering untuk Pengelompokkan Pengguna Media Sosial. *Estimasi: Journal of Statistics and Its Application*, 4(1), 33-43.

Badan Pusat Statistik (BPS) Provinsi Papua. (2024). *Produk Domestik Regional Bruto Provinsi Papua Menurut Lapangan Usaha 2019 – 2023*. Jayapura: Badan Pusat Statistik (BPS) Provinsi Papua.

Fathia, A. N., Rahmawati, R. & Tarno. (2016). Analisis Klaster Kecamatan di Kabuapten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward dan Single Linkage. *Jurnal Gaussian*, 5(4), 801-810.

Febriyati, N. A., DS, A. D., & Wanto, A. (2020). GRDP Growth Rate Clustering in Surabaya City uses the K-Means Algorithm. *International Journal of Information System & Technology*, 3(2), 276-283.

Han, J., & Kamber, M. (2006). *Data Mining: Concept and Techniques*. San Fransisco. Morgan Kauffman Publisher.

Ibrahim, R. N., Hayati, M. N., & Amijaya, F. D. T. (2020). Penerapan Algoritma K-Medoids Pada Pengelompokkan Wilayah Desa atau Kelurahan di Kabupaten Kutai Kartanegara (Studi Kasus: Data Hasil Pendataan Potensi Desa (PODES) Tahun 2018). *Jurnal Eksponensial*, 11(2), 153-158.

Insani, P. N., Darmawan, E., & Sugiyarto. (2024). K-Medoids Algorithm to Clustering COVID-19 patients with Various Age Levels at Hospital in Yogyakarta. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(2), 1014-1018.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York. John Willey & Sons.

Nahdliyah, M. A., Widiharih, T. & Prahutama, A. (2019). Metode K-Medoids Clustering Dengan Validasi Silhouette Index dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018). *Jurnal Gaussian*, 8(2), 161-170.

Ningrum, A. F., & Ahadi, G. D. (2022). Analisis Cluster Kabupaten/Kota di Provinsi Jawa Timur Berdasarkan Laju Produk Domestik Regional Bruto dengan Pendekatan K-Means. *Jurnal Kompetitif: Media Informasi Ekonomi Pembangunan, Manajemen dan Akuntansi*, 8(2), 60-76.

Nurjannah, E., Nasution, M., & Muti'ah, R. (2024). Data Mining Clustering Analysis of Child Growth and Development Using K-Means Method. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(3), 1909-1919.

Reinaldi Y., Ulinnuha, N., Hartono, T. & Hafiyusholeh, M. (2021). Comparison of Single Linkage, Complete Linkage, and Average Linkage Methods on Community Welfare Analysis in Cities and Regencies in East Java. *Jurnal Matematika, Statistika & Komputasi*, 18(1), 130-140.

*name of corresponding author

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

Sapriyanti, S., & Rianto, Y. (2020). Komparasi Metode Clustering K-Means dan Single Linkage untuk Penentuan Kelompok Agent Pada Call Center. *JISAMAR: Journal of Information System, Applied, Management, Accounting and Research*, 4(3), 1-7.

Suraya, G. R., & Wijayanto, A. W. (2022). Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting. *Indonesian Journal of Statistics and Its Applications*, 6(2), 180-201.

Syafiyah, U., Asrafi, I., Wicaksono, B., Puspitasari, D. P., & Sirait, F. M. (2022). Analisis Perbandingan Hierarchical dan Non-Hierarchical Clustering Pada Data Indikator Ketenagakerjaan di Jawa Barat Tahun 2020. *Seminar Nasional Official Statistics*, 1, 803-812.

Thamrin, D. R., & Murni, D. (2022). Analisis ClusterHierarki Metode Single LinkagePada Kabupaten/Kota di Provinsi Sumatera Barat Berdasarkan Indikator Kesehatan. *Journal of Mathematics UNP*, 7(3), 45-51.

Ulvi, H. A., & Ikhsan, M. (2024). Comparison of K-Means and K-Medoids Clustering Algorithms for Export and Import Grouping of Goods in Indonesia. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(3), 1671-1685.

Wardani, S. E., Harahap, S. Z., & Muti'ah, R. (2024). Implementation of the K-Means Method for Clustering Regency/City in North Sumatra based on Poverty Indicators. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(3), 1429-1442.

Warisa, & Nurahman. (2023). Perbandingan Performa Cluster Model Algoritma K-Means Dalam Mengelompokkan Penerima Bantuan Program Keluarga Harapan. *Jurnal Sistem Informasi Bisnis*, 1, 20-28.

Wororomi, J. K., Allo, C. B. G., Paranoan, N. R., Gusthvi, W. (2023). Performance of K-Means and DBSCAN Algorithm in Clustering Gross Regional Domestic Product. *JICP: Journal of International Conference Proceedings*, 6(5), 179-193.

*name of corresponding author