Journal of International Conference Proceedings (JICP) Vol. 6 No. 5, pp. 179-193, November, 2023
P-ISSN: 2622-0989/E-ISSN: 2621-993X
https://www.ejournal.aibpmjournals.com/index.php/JICP

Performance of K-Means and DBSCAN Algorithm in Clustering Gross Regional Domestic Product

Jonathan K. Wororomi[1], Caecilia Bintang Girik Allo[2], Nicea Roona Paranoan[3], Wickly Gusthvi[4]
Universitas Cenderawasih[1, 2, 3, 4]
Jayapura, Papua, Indonesia
Correspondence Email: jkwororomi@gmail.com

ARTICLE INFORMATION

Publication information

Research article

HOW TO CITE

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. *Journal of International Conference Proceedings*, *6*(5), 179-193.

DOI:
https://doi.org/10.32535/jicp.v6i5.2710

Received: 11 September 2023
Accepted: 14 October 2023
Published: 13 November 2023

ABSTRACT

Gross Regional Domestic Product (GRDP) is one of important indicator to determine the economic conditions of a region. GRDP are obtained from sum of value added produced by all unit of production in a region. This study uses GRDP by production approach that grouped into seventeen categories of industry. The government always put the big efforts to increase the economic growth after Covid-19 pandemic. The aim of this study is determined the cluster GDRB based on province in Indonesia at current prices and analyses the performance of the cluster method. The results showed that by using the DBSCAN, two clusters were formed and one province can be detected as an outlier. On the other hands, performance of the method by K-Means showed two clusters. The silhouette value using K-Means is higher than the DBSCAN. For this case, the performance of K-Means is more appropriate than DBSCAN to use in clustering province in Indonesia based on GRDP at Current Market Prices. Moreover, performance of DBSCAN shows more sensitive on outliers detection.

Keywords: DBSCAN, Gross Regional Domestic Product, K-Nearest Neighbor, K-Means

INTRODUCTION

Indonesia is country that continues to develop in various sectors. One sector that the government keep paying attention to is the economic sector. The growth of technology that facilitate the spread of information can be used as one of tools to encourage the business development (Mgunda, 2019). One of the indicators in assessing the success of development on a macro basis is economic growth. Economic growth is characterized by a situation where the total output of actual services resulting from the utilization of production factors in a specific year surpasses the actual income of the population in the preceding year (Runtunuwu & Kotib, 2021). Economic growth is observed through the sustained increase in income over an extended period in each geographical area (Kolinug & Winerungan, 2022). Increase in Gross Domestic Product (GDP) or Gross National Product (GNP) can be used to interpreting economic growth regardless of whether the increase is greater or smaller than the population growth rate, and whether changes in the economic structure occur or not (Lincolin, 1997). GDP is another term for Gross Domestic Product (GDP). During the COVID-19 pandemic in 2020, GDP growth in Indonesia was -2.07% so the government create constructive policies and strategies in an effort to increase GDP growth. One of the efforts taken by the government is the program for handling COVID-19 and National Economic Recovery. In 2021, GDP growth in Indonesia is 3.70% and 5.31% in 2022 (Badan Pusat Statistik [BPS], 2023a).

Determining GDP in Indonesia can not be separated from the GDP in various regions or known as Gross Regional Domestic Product (GRDP) in Indonesia. GRDP is the total value produced by all business units in a certain area, or is the total value of final goods and services produced by all economic units in a region (BPS, 2023b). The GRDP calculation uses three approaches, namely the production, expenditure, and income approaches. The government keep paying attention to these factors so GDP growth in Indonesia continues to increase. In the midst of global economic issues that tend to be negative in 2023, the government needs policies that are right on target. Clustering analysis of GRDP in Indonesia can help the government to find areas that have similarities in GRDP factors. The government can provide different policies between regions based on the grouping results.

Clustering is an analysis can be used to help government knows the region that has same characteristics. There are many methods in cluster analysis. Some of them are hierarchical, partition based, are density-based. Hierarchical and partition-based are popular, but the problem is when they must work on data that has outlier. The principle of clustering is maximizing the intraclass similarity and minimize the interclass similarity so that objects in one cluster have high similarity but are very different from objects in other clusters (Han, Kamber, & Pei, 2012).

K-Means is one of popular method in clustering. However, the K-means algorithm has many challenges that negatively affect its clustering performance (Ikotun, Ezugwu, Abualigah, Abuhaija, & Heming, 2023). Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a new method in clustering. DBSCAN is here to overcome the shortcomings in K-Means. The advantage of the DBSCAN method is that it does not need to determine the number of clusters, but requires a function to calculate the distance between values and some guidelines for the number of centroids that are considered to be close or have the same characteristics. Another advantage of DBSCAN is able to detect noise (Furqon & Muflikhah, 2016).

Research related to clustering analysis on GDP or GRDP calculation factors has been carried out widely. Onwukwe and Ezeorah (2009) used Single Linkage to classify the GDP growth rate in Nigeria using GDP data for 1994-2003. Oktaviana and Amalia (2018) used GRDP data from Bangka Belitung Province for 2010-2017 to predict GRDP for the next year using Trend Analysis. Ningrum and Ahadi (2022) used K-Means to group 38 districts/cities in East Java Province based on factors that support GRDP. Hidayah (2019) used K-Means to determine capital allocation for micro, small, and medium sized enterprises (MSMEs). The MSMEs determined based on its characteristics. They are total assets, sales, and industry.

LITERATURE REVIEW

Clustering is a popular method if people are faced with grouping cases. Cluster is divided into two methods. They are hierarchical and non-hierarchical. Popular algorithms in hierarchical clustering are agglomerative and divisive algorithm. Agglomerative algorithm starts with every object as an individual cluster and every step merge object with another object until become one cluster. There are some methods that popular in agglomerative algorithm. They are single linkage, complete linkage, average linkage, and ward's method. Divisive algorithm is reverse from agglomerative algorithm. Non-hierarchical clustering is also known as partition clustering. Non-hierarchical clustering forming a cluster depends on objects which maximize or minimize some evaluation criteria. Popular calculations in non-hierarchical clustering such as K-Means, K-Medoids, Density-Based Spatial Clustering with Noise (DBSCAN), etc.

Clustering is very useful applied in various applications. K-Means algorithm and DBSCAN algorithm are the two most commonly used clustering techniques that grouping the data based on different criteria. Actual K-Means suffers from several drawbacks, such as it needs a predefined number of clusters and most importantly it does not have the ability to handle noisy data or outliers. But DBSCAN algorithm is free from all these drawbacks and most importantly, it can handle noisy data or outliers data so efficiently. Thus, these two clustering techniques are also efficiently applied on incremental databases whose data is frequently updated. K-Means algorithm is renowned for its simplicity rather than DBSCAN algorithm.

Several previous studies about K-Means and DBSCAN have been done. Ahmar et al. (2018) have been done to cluster provinces in Indonesia based on population density, school participant, human development index, and open unemployment. Wei, Lao, Sato, and Han (2019) used multiple clustering (K-Means, Agglomerative, DBSCAN combine with agglomerative) to clustering product review. The result shows that the best algorithm is DBSCAN combine with agglomerative. Zhang (2019) clustered the member of galaxy using DBSCAN. Çataltaş, Doğramaci, Yumuşak, and Öztoprak (2020) used DBSCAN to find product defects from customer reviews. This study is text mining, so another result can be obtained words often mentioned related to product defects in customer reviews. To determine the epsilon (ε), they use K-Nearest Neighbors. Pamuji and Rongtao (2020) compare K-Means and DBSCAN algorithms on rainfall in Jakarta. The result shows different cluster produced by two methods. The conclusion is K-Means is more efficient and accurate than DBSCAN.

Muningsih, Maryani, and Handayani (2021) used K-Means to cluster province in Indonesia based on village potential and also using Davies Bouldin Index (DBI) to evaluate the cluster. The best cluster is cluster that have smallest DBI (Dista & Abdulloh, 2022). Dewi et al. (2021) compared DBSCAN and K-Means in grouping village status against COVID-19 in Central Java Province. The result of this research is that the silhouette value of DBSCAN is higher than K-Means, so DBSCAN is better than K-Means

for this case. Mulyo and Heikal (2022) used K-Means to cluster the customer in shopping mall in Indonesia. It is very useful to know the characteristics of customer. It is also can be used to know the product categories and target segmentation that attack customer attention. DBSCAN algorithm can be applied on big data. Zhang (2022) applied DBSCAN in information security detection and also combine BIRCH with DBSCAN. Li, Yang, Jiao, and Li (2022) used KMNN-DBSCAN and Partition KMNN-DBSCAN in Rail Damage Data and used silhouette coefficient to evaluate both methods. The result is Partition KMNN-DBSCAN better than KMNN-DBSCAN.

From some studies, the accuracy of K-Means and DBSCAN is depending on the data. In DBSCAN algorithm, determine MinPts and ε can affects the result too. In K-Means, determine the number of cluster (k) can affects the result. In this study, the researchers compared the K-Means algorithm with the DBSCAN algorithm to see the accuracy performance based on the highest silhouette value in Gross Regional Domestic Product (GRDP) case in Indonesia. On the other hand, the researchers want to see how DBSCAN and K-Means work on outlier data.

Elbow Method

Elbow method can be used to determine the number of clusters by comparing the difference SSE of each cluster. The most extreme difference forming the angle of elbow. Below is the formula of Sum of Square Error (SSE) (Bholowalia & Kumar, 2014).

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in K_k} \|x_i - c_k\|^2 \quad (1)$$

Where:
K: number of clusters
x_i: object i in cluster k
c_k: centroid cluster k

K-Means

K-Means method partitions data into groups so data that has the same characteristics is collected into one group and data that has different characteristics collaborates into different groups. The K-Means algorithm describes that each data collected has the closest centroid. One of the advantages of K-Means is its simplicity and computationally efficient, making it suitable for large datasets. However, it has se limitations. First, it assumes that clusters are spherical and similar in size, which may not always align with complex structures in real-world data. Second, it requires determining the number of clusters in advance, which may be difficult to do without prior knowledge of the data. The K-Means algorithm is as follows (Johnson & Wichern, 2007):

a. Determine the number of groups/clusters (k);
b. Calculate the centroid value using the formula in Equation 1;
c.

$$v_{ij} = \frac{\sum_{k=1}^{n_i} x_{kj}}{n_i} \quad (2)$$

where:
v_{ij} : centroid or average of the i^{th} cluster for the j^{th}variable
n_i : the number of the data in the i cluster
x_{kj}: kth data value in the j^{th} cluster for the j^{th} variable

d. Groups items based on the closest centroid based on the distance of the item on the centroid. The distance calculation used is the Euclidean distance as in Equation 2;

$$D_e = \sqrt{\sum_{i=1}^{p}(x_i - s_i)^2} \quad (3)$$

Where:
p : number of variables
x_i : i^{th} data
s_i : i^{th} center

e. Recalculate the group centroid when new items enter or leave the cluster;
f. Do iterations b to d until no items enter or leave the cluster.

The K-Means method is a simple and effective method. After the iteration is stable, it can be guaranteed that the distance between members to each centroid is minimum. But the minimum created is a local minimum. Therefore, K-Means is quite sensitive to the cluster centers selected during the initial step.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
DBSCAN is a clustering method that focuses on data density. One of the advantages of DBSCAN is that it does not require determining the number of clusters. However, DBSCAN requires two important parameters, called minimum points (MinPts) and epsilon (ε) (Safitri, Wuryandari, & Rahmawati, 2017). In DBSCAN algorithm, there are several terms that need to be understood. Core is the central point in a cluster based on density. The border is the point that becomes the boundary within the central point (core) area. Noise is a point that can't be reached by the core and is not a boundary. Direct reachable density is a point that is directly connected to the central point (core). Affordable density is a point that is connected indirectly to a central point (core). Connected density is points that are connected to each other by other points. DBSCAN algorithm is as follows (Devi, Putra, & Sukarsa, 2015): (a) Initialize MinPts and ε parameters; (b) Determine the initial point (p) randomly; (c) Calculate ε or all point distances that the density reaches to p; (d) If the point that satisfies ε is more than MinPts then point p is a core and a cluster is formed; (e) If p is a boundary and there are no points that fall within the reachable density of p, then the process continues to another point; and (f) Repeat steps c to e until all points are processed.

Silhouette Coefficient
Silhouette coefficient can be used to determine best clusters. The score of silhouette coefficient is between -1 and 1. The silhouette coefficient close to 1 indicates that the data is well apart from neighboring clusters and clearly distinguished. A value of 0 indicates that the data is at or very close to the decision boundary between two neighboring clusters and a negative value indicates that the data may be assigned to the wrong cluster. Table 1 shows the criteria of silhouette coefficient (Kaufman & Rousseuw, 1990). Below is the silhouette formula (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (4)$$

Where:
$a(i)$: average dissimilarity of i to all other objects in one cluster, suppose to be A
$b(i)$: minimum average dissimilarity of i to all objects in other cluster, suppose to be C.

For the illustration to calculate $a(i)$ and $b(i)$ is if there are three clusters. They are Cluster X, Cluster Y, and Cluster Z. So, it can be calculated the $a(i)$ and $b(i)$ using this formula (Bariklana and Fauzan, 2023) as follows:

$$a(i) = \frac{1}{|X| - 1} \sum_{j \in X, j \neq i} d(i,j)$$

$$b(i) = \min_{Z \neq A} d(i,Z) \quad \text{where} \quad d(i,Z) = \frac{1}{|Z|} \sum_{j \in Z} d(i,j)$$

Table 1. Criteria of Silhouette Coefficient

Score	Interpretation Structure
0.71 – 1.00	Strong
0.51 – 0.70	Good
0.26 – 0.50	Weak
≤0.25	Bad

Z-Score Normalization

Z-score normalization is one method in data transformation. The formula for Z-score normalization can be shown in (5) (Allo, Putra, Paranoan, & Gunawan, 2023). Data transformation is used when the variables of data have big range or not same in scale of measurement between variables.

$$x_i' = \frac{x_i - \bar{x}}{\sigma_x} \tag{4}$$

Where $\bar{x}$ is average of x and σ_x is standard deviation of x.

RESEARCH METHOD

Data are collected from publication of Badan Statistik Indonesia (BPS) in 2023. The data is GRDP from 34 provinces in Indonesia based on 17 categories. Table 2 shows the categories.

Table 2. 17 Categories that Forming GRDP by Industry

Categories	Definition
X1	Agriculture, Forestry and Fishing
X2	Mining and Quarrying
X3	Manufacturing
X4	Electricity and Gas
X5	Water supply, Sewerage, Waste Management and Remediation Activities
X6	Construction
X7	Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
X8	Transportation and Storage
X9	Accommodation and Food Service Activities
X10	Information and Communication
X11	Financial and Insurance Activities
X12	Real Estate Activities
X13	Business Activities
X14	Public Administration and Defence; Compulsory Social Security
X15	Education
X16	Human Health and Social Work Activities
X17	Other Services Activities

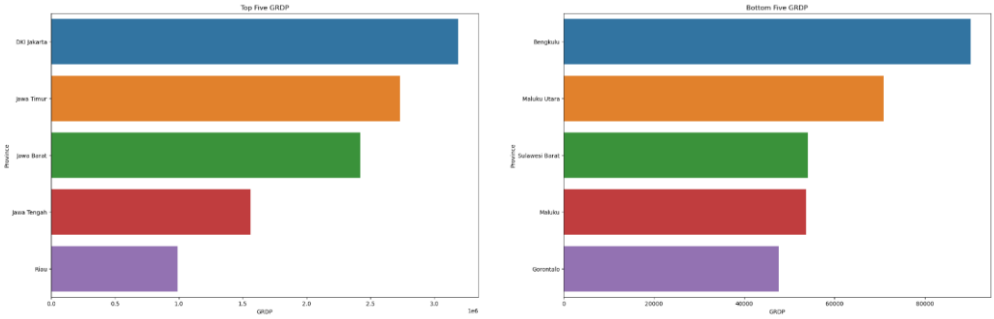
There are four steps in this research. First, using statistic descriptive to describe or provide an overview of the characteristics of categories that forming GRDP data. Second, do data transformation using z-normalization. Third, clustering methods (K-Means and DBSCAN algorithm) are applied to the data. Before K-Means is applied to data, Elbow Method can help to determine the number of clusters. K-Nearest Neighbors (K-NN) is used to determine the epsilon in DBSCAN algorithm. After all methods have been applied, the fourth step is evaluation using silhouette coefficients. The method with the largest silhouette coefficient is chosen to be the best method.

RESULTS

Gross Regional Domestic Bruto (GRDP) is sum of 17 categories used in this paper. There are two pictures in Figure 1. Top five provinces that has highest GRDP (left) and bottom five province that has lowest GRDP in 2022 (right). Figure 1 shows the big contributor to Gross Domestic Product (GDP) is four provinces in Java. They are DKI Jakarta, East Java, West Java, and Central Java. DKI Jakarta as the center of government, business, trade, and services with the largest GRDP contribution which is very far above other provinces. The GRDP data shows a very unequal condition between DKI Jakarta and other provinces in Indonesia in terms of economic capacity. Figure 1 also shows that Gorontalo has the lowest GRDP in 2022. Based on 17 categories, the category Agriculture, Forestry and Fishing has highest value in Gorontalo. So, Gorontalo has potential to increase the GRDP in Agriculture, Forestry and Fishing sector.

The value of each category is sorted from the highest to lowest. Figure 2 is the five provinces with highest value each category. Figure 2 shows the potential each category can be developed by government. Figure 2 shows that DKI Jakarta leading many categories formed GRDP in 2022. It is related that DKI Jakarta has high contribution to GDP in Indonesia. There are three provinces which are almost in every category shown in Figure 2. They are West Java, East Java and DKI Jakarta. Figure 2 also shows that the top five are dominated by provinces in Java and Sumatera. It is related with Figure 1.

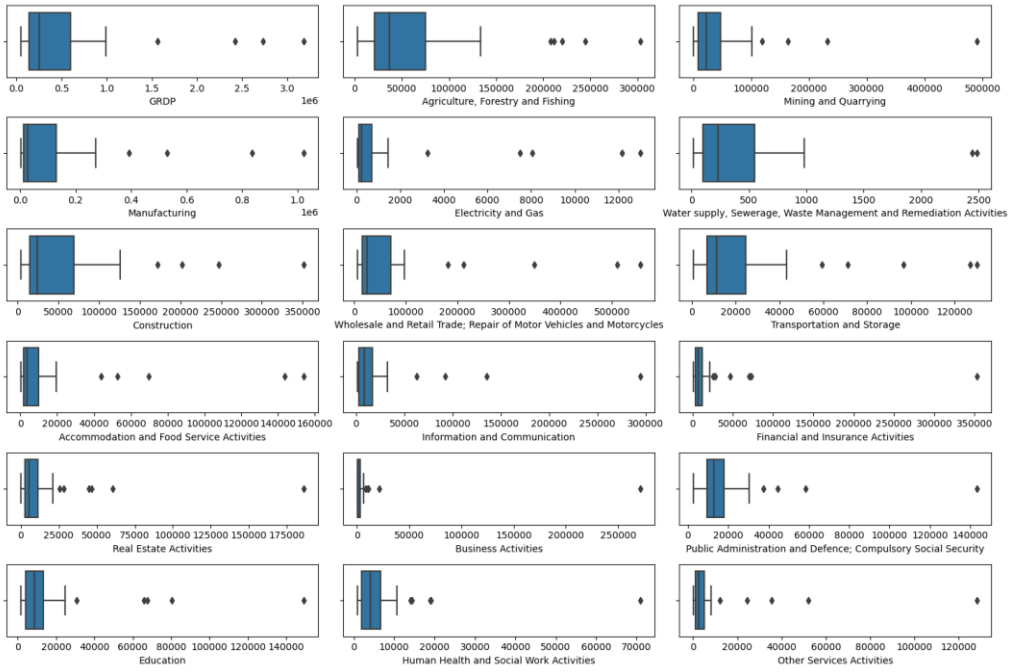
Figure 1. Top and Bottom Gross Regional Domestic Bruto



Before clustering, we should do outlier checking in all variables using boxplot. Figure 3 shows that there are outliers in all variables. The boxplot results also provide information that for each variable there are provinces whose values are very different from other provinces. Figure 3 shows that there are four provinces that become outlier. It means that the four provinces have high GRDP in 2022. The four provinces are DKI Jakarta, East Java, West Java and Central Java. After that, the researcher needs to standardized the data to handle the large range between variables. The standardization method used is Z-Score Normalization.

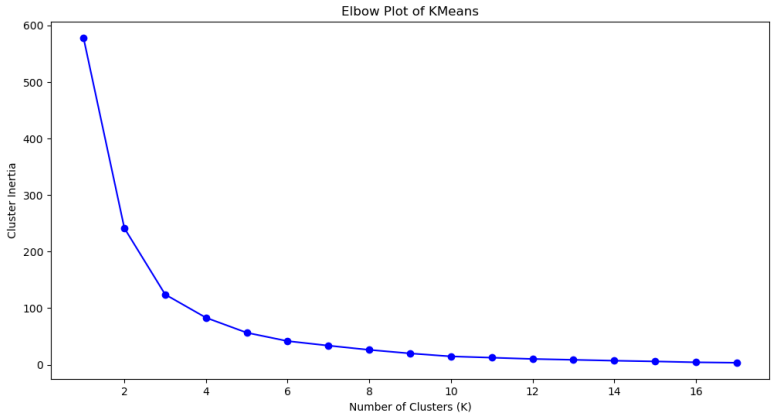
Figure 2. Top Five Provinces Each Category

Figure 3. Boxplot All Variables



Before start with K-Means algorithm. The problem is how many clusters needed for the data because K-Means algorithm requires information regarding the number of clusters. There are some methods to determine number of clusters. The Elbow method is a method that can be used to help determine the number of clusters. Based on the Elbow plot results (Figure 4), it can be concluded that the number of clusters is 2 or 3 clusters (see extreme difference forming the angle of elbow). Another method to determine the number of clusters is running some clusters by the K-Means algorithm and compare the silhouette coefficients, but needs more time to run the algorithm.

Figure 4. Elbow Plot for K-Means Algorithm



Next step, do modeling using K-Means. However, from the Elbow method shows that the number of clusters is 2 or 3 cluster, the researchers try to compare the silhouette coefficients in 2 until 17 clusters. The following table is a summary of the silhouette coefficients for each cluster. In table 3, the silhouette value is highest when two clusters are formed. So, the result is related to the result from elbow method.

Table 3. Silhouette Coefficient for 17 Cluster

Number of Clusters	Silhouette Coefficient
2	0.75
3	0.69
4	0.54
5	0.58
6	0.57
7	0.44
8	0.42
9	0.30
10	0.23
11	0.20
12	0.17
13	0.15
14	0.15
15	0.20
16	0.17
17	0.17

Based on the result from Table 3, the number of members formed in each cluster is as follows.

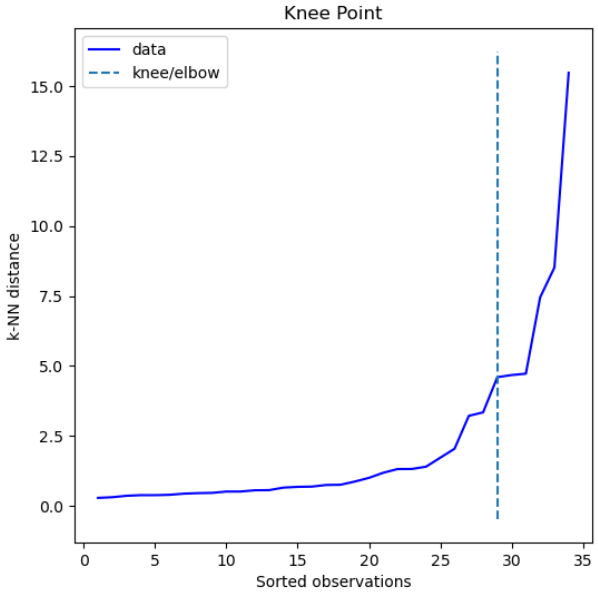
Table 4. The Number of Members Formed in Each Cluster

Cluster	Number of Members	Name of Members
1	31	Aceh, North Sumatera, West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Central Java, DI Yogyakarta, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua
2	3	DKI Jakarta, West Java, and East Java

The K-Means results show that there are 31 provinces formed in the first cluster. Meanwhile, there are 3 provinces in the second cluster that have the same characteristics of 17 categories of Industry in GRDP. The three provinces in second cluster are DKI Jakarta, West Java, and East Java have highest GRDP in 2022.

In the DBSCAN algorithm, two values are required, called Epsilon (ε) and Minimum Points. Minimum Points will be used to represent a cluster. Using KKN (K-Nearest Neighbors) is a method to determine epsilon from the K-NN results (Figure 5). The K-NN Distance value is 4.59, so the epsilon used is 4.6.

Figure 5. K-NN Distance



There are 2 clusters and 1 outlier formed by DBSCAN algorithm with a silhouette value of 0.72. Members of the cluster use DBSCAN algorithm as follows.

Table 5. Members of the Cluster Use DBSCAN

Cluster	Number of Members	Name of Members
1	31	Aceh, North Sumatera, West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, Central Java, DI Yogyakarta, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua
2	2	West Java and East Java
-1 (outlier)	1	DKI Jakarta

From the clusters formed, it shows that there are 31 provinces with the same characteristics in first cluster. Meanwhile, in second cluster are formed 2 provinces namely East Java and West Java.

DISCUSSION

Based on the cluster results, shows that silhouette coefficient is not very far apart using K-Means and DBSCAN algorithms but form different clusters. In DBSCAN, province of DKI Jakarta was detected as an outlier. The results of both algorithms also show the same number of members in the first cluster. In the K-Means algorithm, province of DKI Jakarta, East Java and West Java is form in the same cluster (cluster 2). Based on GRDP

value, three provinces in the same cluster have the highest GRDP in 2022. DBSCAN result can detect an outlier from the GDRP in 2022. It can be a positive and negative perspective. The positive perspective is the government should always pay attention to this province and still develop potential categories in this province. The negative perspective is if something bad happen to this province, then the Indonesian GDP's threatened decline. If we rank the GRDP of each province, the top three provinces are DKI Jakarta, East Java, and West Java. The difference between DKI Jakarta and East Java is 455563 billion rupiahs. The difference is quite big. It can be caused by the recovery of Covid-19 in DKI Jakarta is faster than others provinces. The member of Cluster 1 in both K-Means and DBSCAN algorithm are same. It means the characteristics based on 17 categories between the province is same. If it is compared the value between cluster, the highest value in each category except two categories (Mining and Quarrying & Electricity and Gas) are all in Cluster 2.

CONCLUSION

The boxplot of 17 categories show that every category has outlier. To handle the large range between variables, Standardization using Z-Score Normalization is applied. Clustering using the K-Means algorithm is form two clusters with a silhouette coefficient of 0.75. Meanwhile, clustering using the DBSCAN algorithm is form two clusters and one outlier with a silhouette coefficient of 0.72. The difference in silhouette coefficient between the two algorithms in this study is 0.03. The value is quite small, but form quite different clusters. From the silhouette coefficient, the K-Means algorithm is better than the DBSCAN algorithm. Elbow methods in K-Means algorithm remain a valuable starting point for selecting the number of clusters, and often provide useful insights into the underlying structure of the data. However, it is important to complement it with other validation techniques when working with more complex data sets or different clustering algorithms. Based on this research results, using DBSCAN algorithm automatically adapts to the data, determining the number of clusters without explicit prior knowledge. Also, DBSCAN algorithm shines in handling data with complex structures and outliers. This proves that DBSCAN algorithm can overcome outliers. The outlier showed in DBSCAN can help stakeholders to keep attention on the outlier. On the other hand, K-Means requires determining the number of clusters first. This can be limiting when dealing with diverse and irregular data. While K-Means shines in computational efficiency, simplicity, and performs well when dealing with well-defined clusters.

ACKNOWLEDGMENT
This work has been funded by Faculty of Mathematics and Natural Sciences, Universitas Cenderawasih (FMIPA UNCEN) under PNPB 2023. The opinions expressed here are the authors and do not necessarily reflect the views of funding agency.

DECLARATION OF CONFLICTING INTERESTS
The authors have no affiliation with or involvement in any organization or entity that has a financial or non-financial interest in the subject matter or materials discussed in this paper.

REFERENCES

Ahmar, A. S., Napitupulu, D., Raham, R., Hidayat, R., Sonatha, Y., & Azmi, M. (2018). Using K-Means clustering to cluster provinces in Indonesia. *Journal of Physics: Conference Series*. 1028(012006). doi:10.1088/1742-6596/1028/1/012006

Allo, C. B.G., Putra, L. S.A., Paranoan, N. R., & Gunawan, V.A. (2023). Comparing Logistic Regression and Support Vector Machine in Breast Cancer Problem. *Jambura: Journal of Probability and Statistics, 4*(1), 1-8. doi:10.34312/jjps.v4i1.19246

Badan Pusat Statistik (BPS). (2023, February 6). Ekonomi Indonesia Tahun 2022 Tumbuh 5,31 Persen. *Badan Pusat Statistik*. Retrieved from https://www.bps.go.id/pressrelease/2023/02/06/1997/ekonomi-indonesia-tahun-2022-tumbuh-5-31-persen.html#:~:text=Ekonomi%20Indonesia%20tahun%202022%20tumbuh%20sebesar%205%2C31%20persen%2C%20lebih,Pergudangan%20sebesar%2019%2C87%20persen.

Badan Pusat Statistik (BPS). (2023, April 28). Produk Domestik Regional Bruto Provinsi-Provinsi di Indonesia Menurut Lapangan Usaha 2018 - 2022. *Badan Pusat Statistik*. Retrieved from https://www.bps.go.id/publication/2023/04/28/adbf2e4673599f3dbddca295/produk-domestik-regional-bruto-provinsi-provinsi-di-indonesia-menurut-pengeluaran--2018-2022.html

Bariklana, M., & Fauzan, A. (2023). Implementation of the DBSCAN method for cluster mapping of earthquake spread location. *Barekeng: Journal of Mathematics and Its Applications, 17*(2), 867-878. doi:10.30598/barekengvol17iss2pp0867-0878

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications, 105*(9), 17-24.

Çataltaş, M., Doğramaci, S., Yumuşak, S., & Öztoprak, K. (2020). Extraction of product defects and opinions from customer reviews by using text clustering and sentiment analysis. *2020 IEEE International Conference on Big Data (Big Data),* 4529-4534. doi:10.1109/BigData50022.2020.9377851

Devi, N. M. A. S., Putra, I K. G. D., & Sukarsa, I. M. (2015). Implementasi metode clustering DBSCAN pada proses pengambilan keputusan. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 6*(3), 185-191. doi:10.24843/LKJITI.2015.v06.i03.p05

Dewi, C., Siam, E. P., Wijayanti, G. A., Putri, M., Aulia, N., & Nooraeni, R. (2021). Comparison of DBSCAN and K-Means Clustering for grouping the village status in Central Java 2020. *Jurnal Matematika, Statistika, dan Komputasi*, *17*(3), 394-404. doi:10.20956/j.v17i3.11704

Dista, T. M., & Abdulloh, F. F. (2022). Clustering pengunjung mall menggunakan metode K-Means dan Particle Swarm Optimization. *Jurnal Media Informatika Budidarma*, *6*(3), 1339 – 1348. doi:10.30865/mib.v6i3.4172

Furqon, M. T., & Muflikhah, L. (2016). Clustering the potential risk of tsunami using Density-Based Spatial Clustering of Application with Noise (DBSCAN). *Journal of Environmental Engineering Sustain Technology*, *3*(1), 1-8. doi:10.21776/ub.jeest.2016.003.01.1

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd ed.). Cambridge: Morgan Kaufmann.

Hidayah, A. (2019). Implementing data clustering to identify capital allocation for Small and Medium Sized Enterprises (SMEs). *ASEAN Marketing Journal, X*(1), 66-74. doi:10.21002/amj.v10i1.10627

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. doi:10.1016/j.ins.2022.11.139

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River: Prentice Hall International Inc.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley & Sons Inc.

Kolinug, F. C., & Winerungan, R. (2022). The effect of exports and labor on economic growth in North Sulawesi. *Journal of International Conference Proceedings, 5*(2), 203-211. doi:10.32535/jicp.v5i2.1685

Li, Y., Yang, Z., Jiao, Z., & Li, Y. (2022). Partition KMNN-DBSCAN algorithm and its application in extraction of rail damage data. *Hindawi: Mathematical Problems in Engineering, 2022*, 1 – 10. doi:10.1155/2022/4699573

Lincolin, A. (1997). *Ekonomi Pembangunan* (3rd ed.). Yogyakarta: BP STIE YKPN.

Mgunda, M. I. (2019). The impacts information technology on business. *Journal of International Conference Proceedings, 2*(3), 149-156. doi:10.32535/jicp.v2i3.656

Mulyo, I, A., & Heikal, J. (2022). Customer Clustering Using The K-Means clustering algorithm in shopping mall in Indonesia. *Management Analysis Journal*, *12*(4), 365 – 371. doi:10.15294/MAJ.V11I4.61894

Muningsih, E., Maryani, I., & Handayani, V. R. (2021). Penerapan metode K-Means da optimasi jumlah cluster dengan Index Davies Bouldin untuk clustering propinsi berdasarkan potensi desa. *Evolusi: Jurnal Sains dan Manajemen*, *9*(1), 95 - 100. doi:evolusi.v9i1.10428

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. doi:10.36679/kompetitif.v8i2.5

Oktaviana, N., & Amalia, N. (2018). Gross regional domestic product forecasts using trend analysis: Case study of bangka belitung province. *Jurnal Ekonomi & Studi Pembangunan, 19*(2), 142-151. doi:10.18196/jesp.19.2.5005

Onwukwe, C. E., & Ezeorah, J. N. (2009). Application of single linkage clustering method in the analysis of growth rate of Gross Domestic Product (GDP) at 1990 constant basic prices (Million Naira). *Global Journal of Mathematical Sciences*, *8*(2), 83-93. doi:10.4314/gjmas.v8i2.53751

Pamuji, G. C., & Rongtao, H. (2020). A comparison study of DBScan and K-Means clustering in Jakarta rainfall based on the Tropical Rainfall Measuring Mission (TRMM) 1998-2007. *IOP Conference Series: Materials Science and Engineering*, *879*(1), 012057. doi:10.1088/1757-899X/879/1/012057

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

Runtunuwu, P. C. H., & Kotib, M. (2021). Analysis of the effect construction costs, human development index and investment: Does it have an impact on economic development?. *International Journal of Accounting & Finance in Asia Pasific, 4*(3), 100-113. doi:10.32535/ijafap.v4i3.1210

Safitri, D., Wuryandari, T., & Rahmawati, R. (2017). Metode DBSCAN untuk pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah berdasarkan produksi padi sawah dan padi ladang. *Jurnal Statistika Universitas Muhammadiyah Semarang, 5*(1), 8-13. doi:10.26714/jsunimus.5.1.2017.%25p

Wei, Y., Lao, Y., Sato, Y., & Han, D. (2019). Product-review classification combining multiple clustering algorithms. *ACM International Conference Proceeding Series*, 133–136. doi:10.1145/3338188.3338211

Zhang, M. (2019). Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to identify galaxy cluster members. *IOP Conference Series: Earth and Environmental Science, 252*(042033), 1 – 5. doi:10.1088/1755-1315/252/4/042033

Zhang, Y. (2022). DBSCAN clustering algorithm based on big data is applied in network information security detection. *Security and Communication Networks, 2022*, 1 – 8. doi:10.1155/2022/9951609