Journal of Data Analysis
Vol.7, No.2, December 2024, p. 51-64

People Living with HIV/AIDS in Areas of Poverty and Low Education Level: A K-Medoids Clustering Analysis

Andi Rosilala[1*] and Dewi Juliah Ratnaningsih[2]

[1,2] Department of Statistics, Universitas Terbuka, Tangerang Selatan, Indonesia
E-mail: arosilala@gmail.com*; djuli@ecampus.ut.ac.id

* = corresponding author

Abstrak

Rendahnya tingkat pendidikan dan kondisi ekonomi yang buruk sering dikaitkan dengan penyebaran HIV/AIDS. Penelitian ini bertujuan untuk melakukan analisis klaster terhadap jumlah ODHA, tingkat kemiskinan, dan pendidikan di 34 provinsi di Indonesia dengan menggunakan algoritma K-Medoids. Metode klaster K-Medoids digunakan untuk membagi data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik masing-masing provinsi terkait jumlah ODHA, tingkat kemiskinan, dan pendidikan. Hasil studi menunjukkan bahwa tiga klaster dihasilkan dari analisis K-Medoids. Klaster pertama terdiri dari 13 provinsi yang ditandai dengan persentase penduduk miskin yang tinggi, tetapi dengan tingkat pendidikan dan kasus ODHA yang rendah. Kluster kedua terdiri dari 17 provinsi dengan persentase penduduk miskin yang rendah, tingkat pendidikan yang sedang, dan kasus ODHA yang sedang. Klaster ketiga terdiri dari 4 provinsi dengan tingkat kemiskinan sedang, namun memiliki tingkat pendidikan dan kasus ODHA yang tinggi. Uji MANOVA satu arah (*p-value* < α (0,05)) menunjukkan bahwa terdapat perbedaan karakteristik yang nyata di antara klaster, yang mengindikasikan bahwa ketiga klaster tersebut berbeda secara signifikan satu sama lain. Uji ANOVA satu arah lebih lanjut menunjukkan bahwa ketiga variabel tersebut secara signifikan mempengaruhi pengelompokan provinsi-provinsi di Indonesia.

Informasi Artikel

Sejarah Artikel:
Diajukan Nov 5, 2024
Diterima Dec 26, 2024

Kata Kunci:
HIV/AIDS
K-Medoids
Klaster
Kemiskinan
Pendidikan

Abstract

Low level of education and poor economic conditions are often associated with the spread of HIV/AIDS. This study aims to conduct a cluster analysis on number of the PLWHA, poverty levels, and education across 34 provinces in Indonesia using the K-Medoids algorithm. The K-Medoids clustering method is used to divide the data into several groups based on similar characteristics of each province regarding the PLWHA, poverty levels, and education. The study's result indicate that three clusters were generated from the K-Medoids cluster analysis. The first cluster consists of 13 provinces characterized by a high percentage of poor populations, but with low levels of education and the PLWHA cases. The second cluster consists of 17 provinces with a low percentage of poor populations, moderate levels of education, and moderate the PLWHA cases. The third cluster includes 4 provinces with moderate poverty levels but high levels of education and the PLWHA cases. A one-way MANOVA test (p-value < α (0.05)) showed that there are distinct characteristic differences among the clusters, indicating that the three clusters are significantly different from each other. A one-way ANOVA test further indicated that all three variables significantly influenced the clustering of provinces in Indonesia.

Keyword:
Clustering
Education
HIV/AIDS
K-Medoids
Poverty

Journal of Data Analysis
Vol.7, No.2, December 2024, p. 51-64

1. Introduction

Human Immunodeficiency Virus (HIV) remains a serious global health challenge due to the rising number of cases and the lack of a cure. According to a World Health Organization (WHO) report, by the end of 2022, there were 38.4 million people living with HIV/AIDS (PLWHA) worldwide, with 1.5 million new HIV infections and 680,000 AIDS-related deaths [1].

In Indonesia, HIV/AIDS ranks as the 16th leading cause of fatality [1]. Since it was first detected in 1987 up until March 2023, HIV/AIDS has been reported in 507 out of 514 districts/cities, covering 99 percent of the country. The number of reported HIV cases consistently increased from 2004 to 2019, then declined until 2021, only to rise again in 2022. As of March 2023, the cumulative number of reported HIV cases reached 377,650, while AIDS cases totaled 145,037 [2].

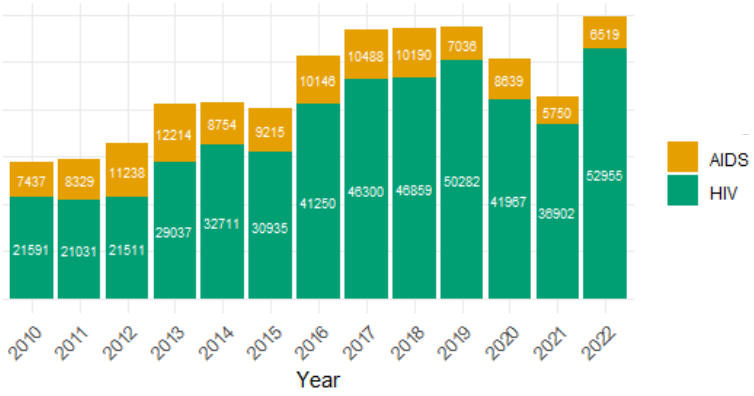

Figure 1 The number of HIV/AIDS cases in Indonesia 2010-2022

According Figure 1, there are four key takeaways from the data. First, the majority of PLWHA (68%) are in the productive age group of 25–49 years, followed by the 20–24 age group at 16.9%. This means that 85% of cases are within the productive age range, which poses a serious threat to population productivity and the spread of HIV. It's no surprise that Indonesia continues to rank first in the Asia-Pacific region in the number of HIV cases [3]. Second, the number of individuals diagnosed as HIV-positive depends on the number of tests conducted within a given period, as shown in the Figure 2 below.

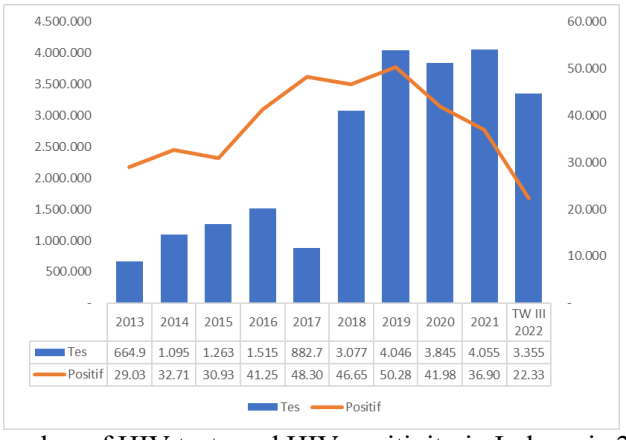

Figure 2 The number of HIV tests and HIV positivity in Indonesia 2013 - Q3 2022

Third, aligned with the previous point, experts suggest that the reported number of HIV/AIDS cases in Indonesia likely does not reflect the actual number. The "Iceberg Phenomenon" applies to this disease, where only a small portion of cases are detected. The WHO

estimates that for every identified HIV case, there may be around 100-200 additional undetected cases [4].

Fourth, HIV/AIDS is a complex social issue that cannot be addressed solely with a health-based approach. According to the Joint United Nations Programme on HIV/AIDS (UNAIDS), HIV/AIDS is closely linked to poverty and education levels, as well as issues like gender inequality [5]. Data from the U.S. Centers for Disease Control and Prevention (CDC) show that HIV infection rates among low-income individuals could be 10 to 20 times higher than the general population in the United States. Additionally, living below the poverty line, not completing secondary education, unemployment, and homelessness are significantly associated with higher HIV prevalence [6]. Khairunisa and Sihaloho also revealed that low education levels and limited knowledge about HIV transmission are closely related to the spread of this virus. Due to educational limitations, many face difficulties securing jobs, ultimately turning to commercial sex work to make ends meet [7].

Previous studies have explored the use of K-Means clustering to analyze public health data. For example, Lesmana, Purwanto, and Dinimaharawati [8], using the K-Means algorithm to develop a web-based application for data processing. Wiliani et al. [9] also used K-Means for clustering HIV/AIDS cases, identifying four high-risk provinces. They focused solely on AIDS transmission, using Rapidminer. However, K-Means is known to be sensitive to outliers, which can distort cluster centroids and result in misleading classifications [10]. Similarly, Noor, Dharmawati, and Qur'ana [11] applied K-Means clustering to categorize HIV/AIDS cases in Banjar, classifying cases into three clusters based on gender, marital status, occupation, and education. Unlike these K-Means studies, this research adopts K-Medoids for clustering, given its robustness to outliers, crucial for handling extreme HIV/AIDS cases in DKI Jakarta, which can skew results in other provinces.

To address these limitations, studies have increasingly turned to K-Medoids clustering, which is more robust against outliers due to its use of medoids instead of centroids. For instance, Rousseeuw [12] demonstrated the effectiveness of K-Medoids in handling noisy datasets, making it a preferred method in complex scenarios. A comparable methodological approach by Purba, Saifullah, and Dewi [13] used K-Medoids for clustering AIDS cases across 34 provinces but only mapped cases into two clusters without determining the optimal cluster count. This study, however, establishes three clusters based on calculated criteria and graphical displays in R Studio, aiming for more precise clustering and a comprehensive view of poverty and education factors alongside HIV/AIDS data.

Based on the description above, we are interested in analyzing the clustering of PLWHA cases, poverty rates, and education levels across 34 provinces in Indonesia. This study uses K-Medoids clustering analysis, known for its robustness against outliers, to categorize the severity level of each province concerning these three issues amid the uncertainty of actual data numbers. Data processing using K-Medoids Clustering is conducted with secondary data obtained from the websites of the Central Bureau of Statistics and the Ministry of Health.

2. Materials and Methods

The subjects of this study are number of people living with HIV/AIDS, number of individuals aged 19-21 who have completed high school education, and number of those living below the poverty line. The data for these subjects are secondary data collected from the HIV/AIDS Information System (SIHA) of the Ministry of Health, covering the number of cases, mortality rates, and distribution across 34 provinces. Additionally, poverty and education level data were obtained from the Statistics Indonesia in the 2022 publication [14] [15].

K-Medoids Clustering is a variation of the K-Means method. The difference between this method and K-Means lies in the use of medoids instead of the mean of each cluster's objects [16]. K-Medoids is used to reduce the sensitivity of partitioning to outliers within a dataset [17]. The

purpose of K-Medoids Clustering is to address K-Means Clustering's limitations due to its sensitivity to outliers, which can impact data distribution [18].

K-Medoids Clustering is a partitioning method that groups n objects into k clusters. An object that represents a cluster is called a medoid [19]. K-Medoids is an algorithm used to find medoids within a cluster, which serves as the central point of that cluster. The K-Medoids algorithm is considered better than K-Means because K-Medoids selects k representative objects to minimize the sum of dissimilarities among data points, while K-Means relies on the sum of Euclidean distances between data points [20].

The steps for the K-Medoids method are as follows:

a. Initialize cluster centers: Set k cluster centers (determine the number of clusters k based on theoretical and conceptual considerations).
b. Allocate each data point: Assign each data point (object) to the nearest cluster using the Euclidean distance metric (calculate the distance from each object to the closest cluster using Euclidean distance).

$$d_{ab} = \sqrt{\sum_{k=1}^{p} (x_{ak} - x_{bk})^2} \quad (1)$$

This formula represents the Euclidean distance between two data points a and b in a p-dimensional space. Here, x_{ak} and x_{bk} are the coordinates of points a and b in the k-th dimension. The formula calculates the straight-line distance by summing the squared differences of their coordinates across all dimensions (k=1 to p) and taking the square root. In K-Medoids, this metric is used to assign data points to the nearest medoid and measure the total distance within clusters for optimization.

c. Randomly select objects: Randomly choose an object in each cluster as a candidate for the new medoid.
d. Calculate distances: Compute the distance of each object in each cluster to the candidate medoid.
e. Calculate total deviation (S): Calculate the new total distance – previous total distance. If $S < 0$, replace the object with cluster data to form a new set of k medoids.

Repeat steps 3 to 5 until no further medoid changes occur, resulting in the final clusters and their respective members.

3. Results and Discussion

3.1. Data Exploration

Data exploration will involve generating descriptive statistical summaries and box plots.

Table 1 Descriptive statistic

Variable	Min	Q1	Median	Mean	Q3	Max	SD
Percentage of Poor Population	4.45	6.26	8.53	10.24	12.22	26.56	5.25
High School Completion Rate	38.47	59.31	66.34	64.68	68.19	87.92	10.86
Number of the PLWHA	18.00	137.75	250.00	636.03	568.25	3576.00	931.08

According to Table 1, for the variable Poverty, values range from 4.45% to 26.56%, with an average of 10.24%. Most data lie between 6.26% (Q1) and 12.22% (Q3), with a standard deviation of 5.25, indicating moderate variability. For School Completion, the data ranges from 38.47% to 87.92%, averaging 64.68%. The median is 66.34%, with most values between 59.31% (Q1) and 68.19% (Q3). The standard deviation is 10.86, showing moderate variation. Lastly for PLWHA,

the most varied variable, ranging from 18 to 3576 cases, with an average of 636. The median is 250, and the standard deviation is 931.08, indicating significant dispersion. Overall, PLWHA shows the widest variability, reflecting potential disparities in the HIV/AIDS burden across provinces. Poverty and school completion are more centrally distributed but still vary moderately.

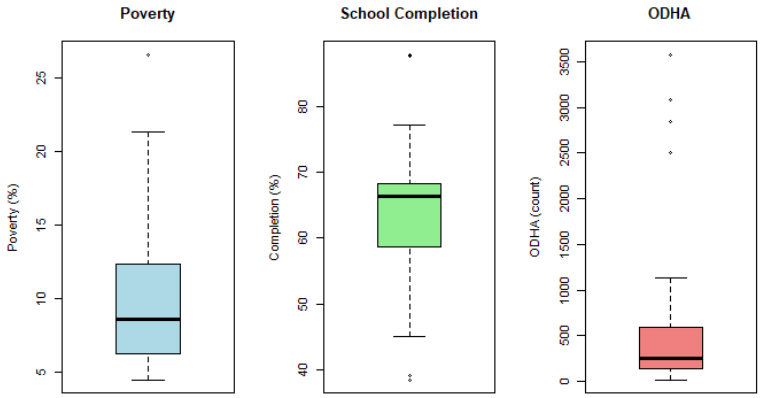

Figure 3 Box plot

Referring to Figure 3, the box plot for Poverty shows a relatively compact distribution, with most provinces having poverty rates between 6% and 12%. A few outliers are present above 20%, suggesting that some provinces experience significantly higher poverty levels compared to the majority. For School Completion, the data exhibits a slightly wider spread, with most values falling between 59% and 68%. There are outliers at both the lower and upper ends, indicating that some provinces have unusually low or high school completion rates compared to the general trend. The box plot for PLWHA reveals a highly skewed distribution, with most provinces reporting fewer than 1,000 cases. However, several significant outliers exceed 2,500 cases, reflecting substantial disparities in the number of HIV/AIDS cases across provinces.

3.2. Determining the Optimal Number of Clusters

The optimal number of clusters using the K-Medoids method was determined through three approaches:

a. Elbow Method: This method calculates the total Within Sum of Squares (WSS) to identify the optimal value of *K*.

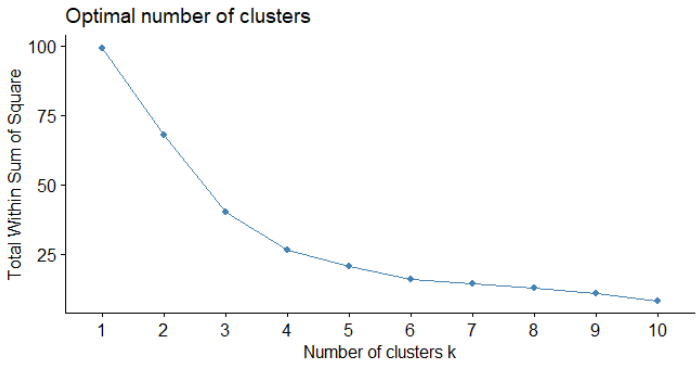

Figure 4 Elbow method

b. Silhouette Coefficient: This combines cohesion and separation, with higher average values indicating better clustering results.

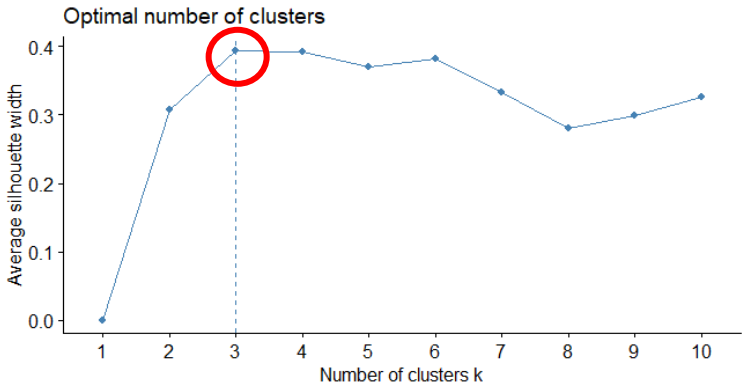

Figure 5 Silhouette coefficient

c. Gap Statistics: This compares total intra-cluster variation for different *K* values against expected values from a distribution without clustering.

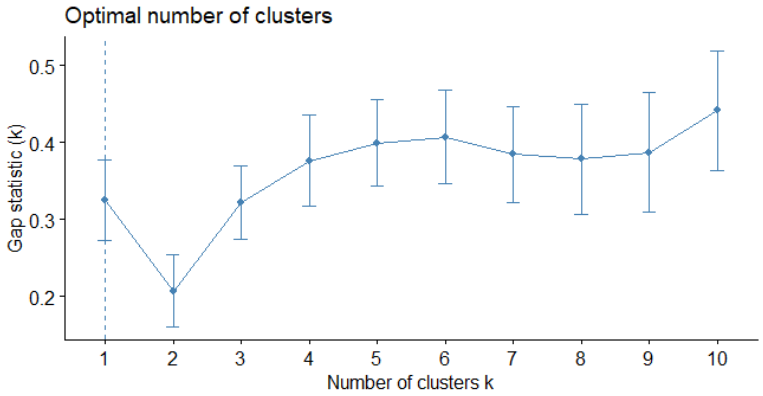

Figure 6 Gap statistic

Using these methods, the optimal *K* is 3.

3.3. K-Medoids Cluster Analysis

The analysis confirmed that the optimal number of clusters is 3, as shown in the comparison of Elbow, Silhouette, and Gap Statistic graphs. The results of the K-Medoids clustering with 3 clusters are shown below.

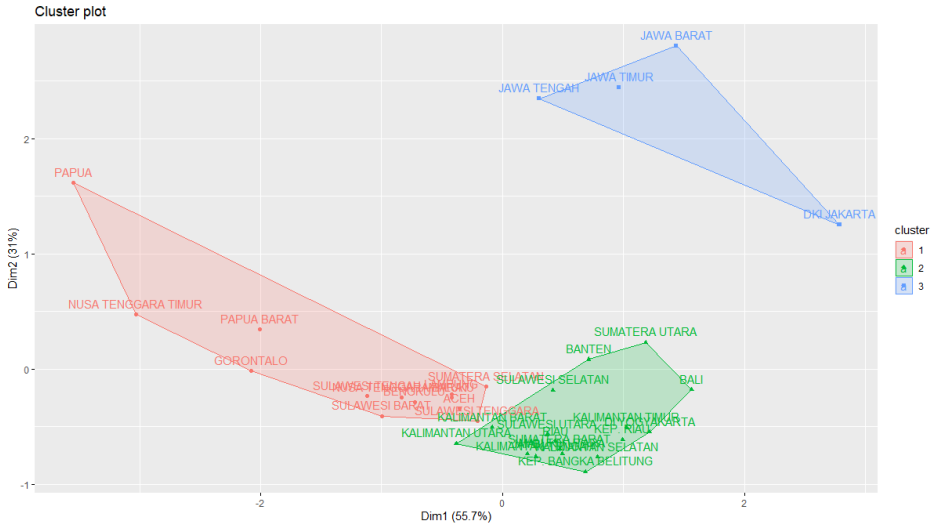

Figure 7 Cluster plot result

From the Figure 7, each cluster contains the following number of members: Cluster 1 has 13 provinces, Cluster 2 has 17 provinces, and Cluster 3 has 4 provinces, with each province within a cluster exhibiting high similarity in characteristics. The results are presented in table below.

Table 2 Cluster Result by Provinces

Cluster	Provinces
Cluster 1	Aceh, Sumatera Selatan, Bengkulu, Lampung, Nusa Tenggara Barat, Nusa Tenggara Timur, Sulawesi Tengah, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Papua Barat, Papua.
Cluster 2	Sumatera Utara, Sumatera Barat, Riau, Jambi, Kep. Bangka Belitung, Kep. Riau, Di Yogyakarta, Banten, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Selatan, Maluku Utara.
Cluster 3	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur

Subsequently, segmentation was performed based on the average values of each cluster, with the results shown in the following table.

Table 3 Cluster segmentation of HIV/AIDS cases, poverty levels, and education

Variable	Cluster 1	Cluster 2	Cluster 3	Pop. Avg
Percentage of Poor Population	15.46	6.66	8.51	10.24
High School Completion Rate	57.89	68.59	70.09	64.68
Number of the PLWHA	226.92	392.35	3001.25	636.03

Based on the Table 3, the green color indicates high indicators, signifying a high average. The red color represents medium indicators, indicating an average level. The yellow color denotes low indicators, reflecting a low average.

From the cluster positions, we can conclude that three groups were formed with the following characteristics: First. Cluster 1, consisting of 13 provinces, has high poverty percentages but low high school completion rates and low numbers of the PLWHA. Second. Cluster 2, made up of 17 provinces, has low poverty percentages with medium high school completion rates and medium numbers of the PLWHA. Third. Cluster 3, consisting of 4 provinces, has medium poverty percentages with high high school completion rates and high numbers of the PLWHA.

3.4. One-Way MANOVA and One-Way ANOVA Testing

Considering that the clustering results using the K-Medoids method produced 3 clusters (levels of factors) and there are 3 response variable indicators, testing was conducted using one-way MANOVA. A large Roy's Largest Root value (10.89) with $p < 0.05$ for Cluster indicates a significant difference between clusters on the variables tested in this multivariate analysis. The Observed Power value is 1.000 for all statistics, which indicates that the power of the test is very high, meaning that the test has an excellent ability to detect differences.

Prior to the one-way MANOVA and one-way ANOVA tests, multivariate normality and homogeneity of variance-covariance matrix assumptions were checked. Based on the Kolmogorov-Smirnov test, the p-value was found to be 0.472, which is greater than 0.05, leading to the decision to fail to reject H0. This indicates that the data meet the multivariate normal distribution assumption. According to Tabel 3, the one-way MANOVA test revealed a p-value $< \alpha$ (0.05), leading to the decision to reject H0. This indicates that there are significant differences in characteristics across the formed clusters.

Table 4 One-way MANOVA test

Effect	Test	Value	F	Hypothesis df	Sig.	Observed Power
Intercept	Pillai's Trace	.98	646.01	3.00	.00	1.00
	Wilks' Lambda	.01	646.01	3.00	.00	1.00
	Hotelling's Trace	66.82	646.01	3.00	.00	1.00
	Roy's Largest Root	66.82	646.01	3.00	.00	1.00
Cluster	Pillai's Trace	1.51	31.24	6.00	.00	1.00
	Wilks' Lambda	.03	43.00	6.00	.00	1.00
	Hotelling's Trace	12.39	57.83	6.00	.00	1.00
	Roy's Largest Root	10.89	108.99	3.00	.00	1.00

Subsequently, a one-way ANOVA analysis was conducted in Tabel 5, followed by data visualization using boxplots in Figure 8. Boxplots are a method in descriptive statistics used to graphically illustrate the distribution of numerical data. Below are the boxplots for each variable along with a review of the results from the one-way ANOVA tests conducted.

Table 5 One-way ANOVA test

		Sum of Squares	df	Mean Square	F	Sig.
Percentage of Poor Population	Between Groups	584.115	2	292.057	27.873	.000
	Within Groups	324.827	31	10.478		
	Total	908.942	33			
High School Completion Rate	Between Groups	976.357	2	488.178	5.191	.011
	Within Groups	2915.288	31	94.042		
	Total	3891.645	33			
Number of the PLWHA	Between Groups	25562287.42	2	12781143.71	130.093	.000
	Within Groups	3045637.555	31	98246.373		
	Total	28607924.97	33			

Based on Tabel 4, all variables had a p-value < 0.05. Therefore, the decision is to reject H0, indicating that all three variables significantly influence the formation of provincial clusters in Indonesia.

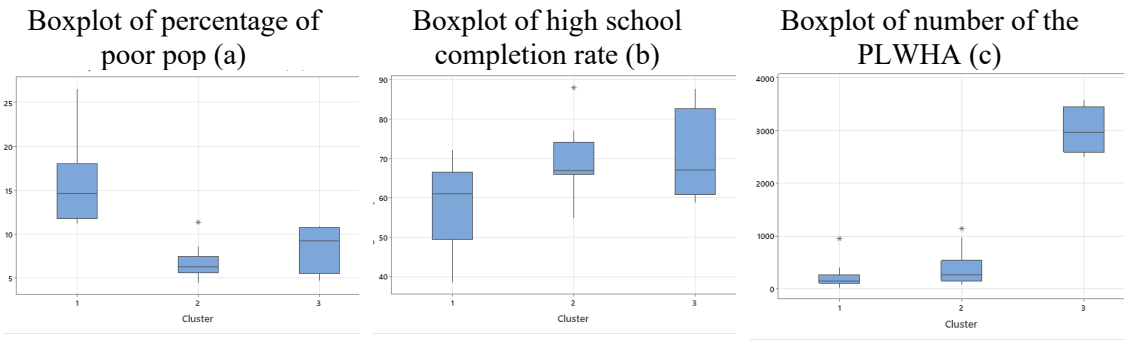

Figure 8 Box plot result

From Figure 8, we can interpret that: First, Cluster 1 represents areas with higher and more varied poverty rates, Cluster 2 has consistently low poverty rates with minimal variation, and Cluster 3 shows moderate poverty levels with some variability. The differences in medians and IQRs between clusters suggest distinct characteristics related to poverty distribution among them.

Second, Cluster 1 represents areas with generally lower and more varied education completion rates, Cluster 2 has moderate and consistent education completion rates with an outlier, and Cluster 3 has the highest education completion rates with considerable variability. This distribution suggests that each cluster has distinct characteristics regarding educational attainment. Third, Cluster 1 represents areas with low counts, Cluster 2 represents areas with moderate counts, and Cluster 3 represents areas with high counts of PLWHA. The boxplot suggests a clear separation between the clusters in terms of the number of PLWHA cases, with Cluster 3 having the highest concentration.

Due to social and economic issues, individuals with HIV/AIDS tend to isolate themselves and do not report their condition, suggesting that the actual number of PLWHA is much higher than recorded by the government. Therefore, cluster analysis of HIV/AIDS cases is essential for mapping severity across the 34 provinces of Indonesia, allowing for the identification of case clusters. Another finding from this analysis emerges when we include poverty and education levels alongside HIV/AIDS data. In Cluster 3, four provinces—West Java, DKI Jakarta, East Java, and Central Java—exhibit high levels of HIV/AIDS and education, with moderate poverty rates. It was previously noted that higher HIV/AIDS prevalence correlates with lower education levels. However, in this cluster, residents in Java generally have higher education, leading to greater awareness and proactive reporting regarding HIV/AIDS. This contributes to the higher recorded cases in these provinces.

Conversely, in Cluster 1, there are low numbers of the PLWHA, low education levels, and high poverty rates. Previous findings indicated that poor populations are 10 to 20 times more at risk for HIV/AIDS than the general population. This suggests that Indonesia's policy of not localizing areas where prostitution is legalized has caused sex workers—who are most vulnerable to HIV/AIDS—to spread randomly. Those from impoverished areas migrate to economic centers to work, resulting in unrecorded PLWHA in their home areas and artificially low data. In contrast, their workplaces record these individuals, leading to higher case numbers.

4. Conclusion and Recommendation

The application of cluster analysis using the K-Medoids algorithm to group HIV/AIDS cases across 34 provinces in Indonesia revealed several important insights. The optimal number of clusters, determined using the Elbow, Silhouette, and Gap Statistic methods, is three. This indicates that the provinces can be distinctly grouped into three clusters based on the characteristics of poverty rates, high school completion rates, and the number of PLWHA. The three clusters exhibit clear distinctions. The first cluster includes 13 provinces characterized by high poverty rates, low high school completion rates, and relatively low numbers of the PLWHA. The second cluster comprises 17 provinces with lower poverty rates, moderate high school completion rates, and moderate numbers of the PLWHA. Lastly, the third cluster consists of 4 provinces with moderate poverty rates, high high school completion rates, and a high number of PLWHA. These groupings highlight regional disparities in socioeconomic and health indicators, reflecting potential targets for policy intervention. Further statistical analysis strengthens these findings. The one-way MANOVA test yielded a p-value below 0.05, confirming that the differences between clusters are statistically significant across all three variables. Additionally, the one-way ANOVA test also resulted in a p-value below 0.05, demonstrating that poverty, school completion, and HIV/AIDS case numbers significantly influence the clustering of provinces. This analysis underscores the importance of considering these variables in developing region-specific strategies to address disparities in health and socioeconomic conditions, especially in a context like Indonesia, where clustering approaches have proven useful in other spatial and health-related studies.

Based on the cluster analysis, targeted recommendations address the specific needs of each group of provinces. Cluster 3, with moderate poverty, high education levels, and high PLWHA cases, requires focused HIV prevention programs, awareness campaigns, and improved healthcare access. Cluster 1, marked by high poverty, low education levels, and low PLWHA

cases, would benefit from education and economic development initiatives to reduce future HIV risks. Meanwhile, Cluster 2, characterized by low poverty, moderate education, and moderate PLWHA cases, requires ongoing monitoring and strengthened HIV prevention and treatment services to prevent escalation. These tailored strategies aim to improve public health outcomes across all clusters.

References

[1] World Health Organization. HIV and AIDS: Key Facts. https://www.who.int/news-room/fact-sheets/detail/hiv-aids. Accessed August, 5th 2024.

[2] Ministry of Health. Laporan Eksekutif Perkembangan HIV/AIDS dan Penyakit Infeksi Menular Seksual (PIMS) Triwulan III Tahun 2022. https://siha.kemkes.go.id/portal/files_upload/Laporan_TW_3_2022.pdf. Accessed August, 1th 2024.

[3] Ahdiat, A. 2022. Indonesia Punya Pengidap HIV Terbanyak di Asia Tenggara. https://databoks.katadata.co.id/datapublish/2022/09/22/indonesia-punya-pengidap-hiv-terbanyak-di-asia-tenggara. Accessed April, 1st 2024.

[4] Siregar, F. 2004. Pengenalan dan Pencegahan AIDS. Medan: USU Digital Library.

[5] United Nation. UNAIDS Input SGs Report Eradicating Poverty https://www.un.org/development/desa/socialperspectiveondevelopment/wp-content/uploads/sites/27/2020/07/UNAIDS-Inputs-to-the-Report-of-the-Secretary-General-2020.pdf. Accessed May, 2nd 2024.

[6] United Nation. Mainstreaming AIDS in Development Instruments and Processes at the National Level: A Review of Experiences. https://www.unaids.org/sites/default/files/media_asset/mainstreaming_aids-in_dev_instr_rep_28nov05_en_0.pdf. Accessed May, 3rd 2024.

[7] Khoirunisa, N and Sihaloho, E. 2019. Determinan Pembangunan Daerah dan Angka HIV/AIDS di Indonesia. Jurnal Ilmu Ekonomi dan Studi Pembangunan Vol. 19 No. 1.

[8] Lesmana, A, Purwanto, Y, and Dinimaharawati, A. 2019. Implementasi Algoritma K-Means untuk Clustering Penyakit HIV/AIDS di Indonesia. e-Proceeding of Engineering Vol.6, No.2.

[9] Wiliani, et.al. 2021. Penerapan Rapidminer Dengan K-Means Pada Clustering Daerah Terjangkit AIDS. Journal of Informatics and Advanced Computing. Vol. 2, No.2.

[10] Hartigan, J. A., and Wong, M. A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108.

[11] Noor, H, Dharmawati, A, and Quranan, T. 2021. Penerapan Algoritma K-Means Clustering Analysis pada Kasus Penderita HIV/AIDS (Studi Kasus Kabupaten Banjar). Technologia Vol 12, No. 2.

[12] 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.

[13] Purba, L, Saifullah, and Dewi R. 2019. Pengelompokan Kasus Penyakit AIDS Berdasarkan Provinsi dengan Data Mining K-Medoids Clustering. Konferensi Nasional Teknologi Informasi dan Komputer. Vol. 3, No. 1.

[14] Statistics Indonesia. Population and Migration. https://www.bps.go.id/subject/28/pendidikan.html#subjekViewTab3.html. Accessed March, 3rd 2024.

[15] Statistics Indonesia. Percentage of Poor Population https://www.bps.go.id/indicator/23/192/1/persentase-penduduk-miskin-p0-menurut-provinsi-dan-daerah.html. Accessed March, 3rd 2024.

[16] Farissa, R, Mayasari, R, and Umaidah, Y. 2021. Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di

Puskesmas Karangsambung. Journal of Applied Informatics and Computing, 5(2), 109-116.

[17] Vercillis, C. 2009. Business Intelligence, Data Mining and Optimization for Desicion Making. Milan: Wiley.

[18] Han, J, and Kamber, M. 2017. Data Mining, Concept and Techniques. Waltham: Morgan.

[19] Setyawati, A. 2017. Implementasi Algoritma Partitioning Around Medoid (PAM) untuk Pengelompokkan Sekolah Menengah Atas di DIY Berdasarkan Nilai Daya Serap Ujian Nasional. Yogyakarta: Universitas Danata Darma, Fakultas Sains dan Teknologi.

[20] Marlina, D, Lina, N, Fernando, A, and Ramadhan, A. 2018. Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi 4(2), 64.

Appendix 1:

```
> kmed
Medoids:
                     ID Persentase.Penduduk.Miskin..P0.
NUSA TENGGARA BARAT 18                       0.65484535
RIAU                 4                      -0.65988908
JAWA TIMUR          15                       0.02605931
                    Tingkat.Penyelesaian.Pendidikan.SMA.sederajat
Jumlah.ODHA
NUSA TENGGARA BARAT                                    -0.3391719  -
0.5327473
RIAU                                                    0.2050523  -
0.4210488
JAWA TIMUR                                              0.2013689
2.6248819
Clustering vector:
                ACEH        SUMATERA UTARA        SUMATERA BARAT
                   1                     2                     2
               RIAU                 JAMBI      SUMATERA SELATAN
                   2                     2                     1
           BENGKULU               LAMPUNG KEP. BANGKA BELITUNG
                   1                     1                     2
         KEP. RIAU           DKI JAKARTA             JAWA BARAT
                   2                     3                     3
        JAWA TENGAH         DI YOGYAKARTA             JAWA TIMUR
                   3                     2                     3
             BANTEN                   BALI  NUSA TENGGARA BARAT
                   2                     2                     1
 NUSA TENGGARA TIMUR     KALIMANTAN BARAT    KALIMANTAN TENGAH
                   1                     2                     2
  KALIMANTAN SELATAN     KALIMANTAN TIMUR     KALIMANTAN UTARA
                   2                     2                     2
      SULAWESI UTARA      SULAWESI TENGAH     SULAWESI SELATAN
                   2                     1                     2
   SULAWESI TENGGARA            GORONTALO       SULAWESI BARAT
                   1                     1                     1
              MALUKU         MALUKU UTARA           PAPUA BARAT
                   1                     2                     1
               PAPUA
                   1
Objective function:
    build      swap
0.8821568 0.8719135

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
 [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

Normality dan Homogenity Test

```
mnorm.test<-function(x)
  {
  rata2 <- apply(x, 2, mean)
  mcov <- var(x)
  ds <- sort(mahalanobis(x, center = rata2,cov = mcov))
  n <- length(ds)
  p <- (1:n-0.5)/n
  chi <- qchisq(p, df = ncol(x))
  win.graph()
  plot(ds, chi, type = "p")
  return(ks.test(ds,chi,df = ncol(x)))
  }
mnorm.test(dataclus1[1:3])
```

```
> mnorm.test(dataclus1[1:3])
```

```
        Two-sample Kolmogorov-Smirnov test

data:  ds and chi
D = 0.20588, p-value = 0.4728
alternative hypothesis: two-sided
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Persentase_Penduduk_Miskin	Between Groups	584.115	2	292.057	27.873	.000
	Within Groups	324.827	31	10.478		
	Total	908.942	33			
Tingkat_Penyelesaian_Pendidikan_SMA	Between Groups	976.357	2	488.178	5.191	.011
	Within Groups	2915.288	31	94.042		
	Total	3891.645	33			
Jumlah_ODHA	Between Groups	25562287.42	2	12781143.71	130.093	.000
	Within Groups	3045637.555	31	98246.373		
	Total	28607924.97	33			

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Noncent. Parameter	Observed Power[d]
Intercept	Pillai's Trace	.985	646.005[b]	3.000	29.000	.000	1938.016	1.000
	Wilks' Lambda	.015	646.005[b]	3.000	29.000	.000	1938.016	1.000
	Hotelling's Trace	66.828	646.005[b]	3.000	29.000	.000	1938.016	1.000
	Roy's Largest Root	66.828	646.005[b]	3.000	29.000	.000	1938.016	1.000
Klaster	Pillai's Trace	1.515	31.246	6.000	60.000	.000	187.474	1.000
	Wilks' Lambda	.034	43.000[b]	6.000	58.000	.000	258.000	1.000
	Hotelling's Trace	12.394	57.837	6.000	56.000	.000	347.022	1.000
	Roy's Largest Root	10.899	108.990[c]	3.000	30.000	.000	326.970	1.000

a. Design: Intercept + Klaster

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

d. Computed using alpha = ,05