J-INTECH (Journal of Information and Technology)
Accredited Sinta 4 Ministry of Higher Education, Science and Technology Republic of Indonesia SK No. 10/C/C3/DT.05.00/2025
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Clustering HIV Screening Data in Teluk Bintuni Using K-Means

Yuliana Regina Nelce Manobi[1*], Alex De Kweldju[2], Julius Panda Putra Naibaho[3]

[1,2,3]*University of Papua, Faculty of Engineering, Informatics Engineering, Jl. Gunung Salju, Amban, Manokwari, Papua Barat 98314, Indonesia*

Keywords
Clustering; Health Analytics; HIV Screening; K-Means Algorithm;Teluk Bintuni

Corresponding Author:
manobimarion@gmail.com

Abstract
Human Immunodeficiency Virus (HIV) remains a major public health challenge in Papua, Indonesia, where geographical barriers and limited resources complicate service delivery. This study applies clustering methods to HIV screening data from 22 Health Service Units (UPK) in Teluk Bintuni Regency during 2024–2025. The final dataset consisted of 29 valid unit-year observations containing variables related to total tests, HIV-positive cases, and sex-disaggregated distributions. Data preprocessing followed the Knowledge Discovery in Database (KDD) framework, including data selection, cleaning, log transformation, normalization, and the construction of derived variables such as positivity rate and gender ratios. Four clustering algorithms were compared, namely K-Means, Hierarchical Clustering, Fuzzy C-Means, and DBSCAN, using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The results indicate that K-Means produced the most stable and interpretable clustering structure, forming three groups of UPKs: intermediate screening units with moderate coverage and low positivity, priority units with limited testing but high positivity rates, and active screening units with the highest testing volume and case detection. These findings reveal heterogeneity in HIV screening performance across UPKs and support differentiated intervention strategies. Units with high apparent positivity but low testing coverage require expanded outreach, field verification, and improved access to HIV screening services, while active screening units should be strengthened through counseling, referral, and follow-up services. This study demonstrates the usefulness of clustering analysis in identifying service gaps and supporting evidence-based HIV intervention planning at the local level.

1. Introduction

Human Immunodeficiency Virus (HIV) remains one of the public health issues requiring serious attention. This disease not only impacts individual health but also influences health service planning, transmission prevention, and policy-making at the local level. In Indonesia, HIV and AIDS cases are still found in various regions with uneven rates of spread. Papua and West Papua are regions with epidemiological characteristics that differ from many other areas in Indonesia, thus requiring an analytical approach more suited to local conditions [1] . A report from the Ministry of Health of the Republic of Indonesia indicates that strengthening

HIV services remains a crucial component of efforts to control infectious diseases [1] . In Teluk Bintuni Regency, HIV services also face unique challenges, such as geographical conditions, limited health resources, and the need to strengthen programs based on minimum service standards [2] .

HIV screening data collected by Health Service Units (UPKs) holds significant value as it can illustrate testing coverage, the number of positive cases, and service patterns across units. However, this data struggles to provide in-depth insights when analyzed manually. Differences in the number of tests, positivity rates, and service distribution across UPK units are not always clearly visible without the aid of data analysis methods. Therefore, a data mining approach is necessary to transform health data into more structured and useful information for decision-making [3],[4].

Clustering is one of the data mining methods that can group data based on similar characteristics [5] . In the context of HIV services, clustering can help identify groups of UPKs with different screening patterns, such as units with low coverage, units with active services, or units with positivity rates that require attention. Previous studies have shown that K-Means can be used in HIV data analysis, including for clustering AIDS cases by province [6] , in Banjar District [7] , and in West Java [8] , [9] . Additionally, K-Means has been applied to evaluate the performance of health facilities in reporting HIV indicators [10] , and for clustering districts in Papua [11] . Unsupervised machine learning approaches have been widely used in clinical population segmentation [12], and clustering has also proven helpful in identifying patterns in HIV services and financing across various countries [13].

Unlike previous HIV clustering studies in Indonesia, which generally used aggregate data at the provincial, district/city, or subdistrict levels, this study focuses on the Health Service Unit (UPK) level, thereby providing a more operational picture of the characteristics of HIV screening services. This study not only uses variables such as the absolute number of tests and the number of positive cases but also utilizes derived variables such as the positivity rate, and the distribution of tests and positive cases by gender to represent service characteristics more comprehensively.

Additionally, this study compares four clustering methods—K-Means, Hierarchical Clustering, Fuzzy C-Means, and DBSCAN—on a small-scale health service dataset to evaluate cluster stability, data separation quality, and the ease of interpreting clustering results. A combined approach involving UPK-level analysis, the use of feature engineering on HIV service data, and a comparative evaluation of several clustering algorithms remains relatively limited in data mining-based HIV research in Indonesia.

K-Means was selected as the primary method due to its simplicity, efficiency, and widespread use in various health contexts [5],[6],[7],[8]. As a comparison, Hierarchical Clustering was used, which does not require the number of clusters to be specified at the outset and produces an informative dendrogram structure [14] , Fuzzy C-Means (FCM), which allows for partial membership, making it more flexible in handling ambiguous data [15] , [16] and DBSCAN, which is capable of detecting clusters with arbitrary shapes and automatically identifying outliers [17] . This study contributes to the mapping of HIV services based on health service units, an area that has not been extensively studied in Indonesia.

To the best of our knowledge, comparative clustering studies on HIV screening data at the Health Service Unit (UPK) level in Papua remain limited. This study contributes by combining UPK-level analysis, feature engineering using positivity rate and gender-based indicators, and comparative evaluation of multiple clustering algorithms to support evidence-based HIV service planning in Teluk Bintuni Regency. In addition, this study emphasizes cluster stability and interpretability in a small-scale local health dataset through multi-metric internal evaluation.

Based on the above description, this study aims to cluster health service units in Teluk Bintuni Regency based on HIV screening performance using K-Means and compare it with other clustering methods. The results of this study are expected to identify differences in service patterns among health service units and provide supporting information for determining priorities for HIV interventions at the regional level.

2. Research Method

The research methodology adopted in this study follows a systematic workflow based on the Knowledge Discovery in Database (KDD) framework. The overall process, illustrated in Figure 1. Research Method, begins with data collection and proceeds through data selection, preprocessing, and transformation before applying clustering algorithms. Four clustering methods K-Means, Hierarchical Clustering, Fuzzy C-Means, and DBSCAN were implemented to analyze HIV screening data. The results were then evaluated, compared, and interpreted to provide meaningful insights for HIV service planning in Teluk Bintuni Regency.

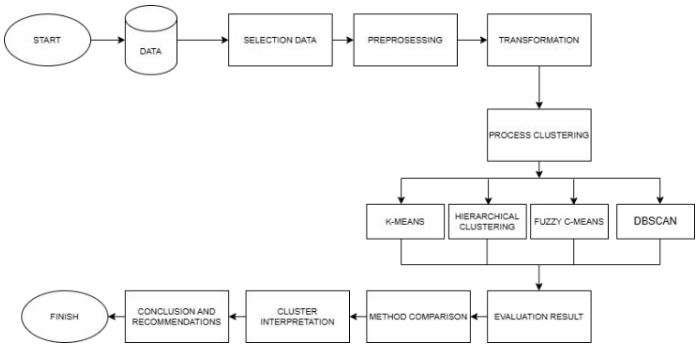

Figure 1. Research Method

2.1 Data

This study uses HIV screening data from Health Service Units (Unit Pelayanan Kesehatan/UPK) in Teluk Bintuni Regency. Data were obtained from HIV screening reports for 2024 and 2025. The dataset covers 22 UPKs with primary attributes including UPK name, reporting year, number of individuals tested for HIV, number of HIV-positive cases, number of males and females tested, and the number of positive cases by sex. A summary of the dataset characteristics and the number of observations used in this study is presented in Table 1.

Table 1. Summary Of HIV Screening Dataset

Indicator	Value
Number of initial unit-year observations	44
Number of excluded observations	15
Final observations used in analysis	29
Year of analysis	2024, 2025
Total individuals tested	12,347
Total HIV-positive cases	255
Overall positivity rate	2.07%
Total male tested	4,914
Total female tested	7,433
Total male positive	121
Total female positive	134

2.2 Data Selection

Data selection was conducted to ensure that only complete and relevant observations were used in the clustering process. The initial dataset consisted of 44 unit-year observations from 22 Health Service Units (UPK) during 2024–2025. A total of 15 observations were excluded because several UPKs did not conduct HIV screening activities during the reporting period, resulting in missing values in key variables, including the total number of individuals tested, total HIV-positive cases, male and female testing data, and male and female positive cases. Most excluded records were therefore associated with the absence of screening activities rather than data entry errors. According to information from the District Health Office, incomplete reporting in several UPKs was also related to structural and operational constraints, including the absence of HIV program management personnel, limited internet access for reporting through the SIHA 2.1 application, and the lack of HIV screening implementation during the reporting period. After the data selection process, 29 valid observations were retained for further analysis. This complete-case approach was used to avoid biased clustering results caused by incomplete screening records and to preserve the integrity of the observed service conditions.

2.3 Data Preprocessing and Transformation

After data selection, several preprocessing and transformation steps were performed before clustering. First, derived variables were calculated, including the positivity rate and gender ratios. The positivity rate was obtained by dividing the total number of HIV-positive cases by the total number of individuals tested. Gender ratios were used to represent the proportion of male and female testing or positive cases within each observation. Second, log transformation was applied to the total number of individuals tested and total positive cases to reduce the influence of highly skewed values. Finally, all numerical variables used in clustering were standardized using StandardScaler so that each variable had a comparable scale and contributed proportionally to the clustering process.

2.3.1 Positivity Rate

$$\text{Positivity Rate} = \frac{\text{Total Positive Cases}}{\text{Total Individuals Tested}} \qquad (1)$$

Where *Total Positive Cases* represents the number of HIV-positive cases, while *Total Individuals Tested* represents the total number of individuals who underwent HIV testing. A higher positivity rate indicates a higher proportion of positive cases among the tested population.

2.3.2 Male Test Ratio

$$\text{Male Test Ratio} = \frac{\text{Male Tested}}{\text{Total Individuals Tested}} \qquad (2)$$

Where Male Tested represents the number of male individuals who underwent HIV testing, while Total Individuals Tested represents the total number of individuals tested. This ratio is used to describe the proportion of male participation in HIV screening services.

2.3.3 Female Test Ratio

$$\text{Female Test Ratio} = \frac{\text{Female Tested}}{\text{Total Individuals Tested}} \qquad (3)$$

Where Female Tested represents the number of female individuals who underwent HIV testing, while Total Individuals Tested represents the total number of individuals tested. This ratio is used to represent the proportion of female participation in HIV screening services.

2.3.4 Male Positive Ratio

$$\text{Male Positive Ratio} = \frac{\text{Male Positive Cases}}{\text{Total Positive Cases}} \qquad (4)$$

Where Male Positive Cases represents the number of HIV-positive cases among males, while Total Positive Cases represents the total number of HIV-positive cases. This ratio is used to identify the proportion of positive cases occurring in male individuals.

2.3.5 Female Positive Ratio

$$\text{Female Positive Ratio} = \frac{\text{Female Positive Cases}}{\text{Total Positive Cases}} \quad (5)$$

Where Female Positive Cases represents the number of HIV-positive cases among females, while Total Positive Cases represents the total number of HIV-positive cases. This ratio is used to identify the proportion of positive cases occurring in female individuals.

2.3.6 Log Transformation

$$x' = \log(1 + x) \quad (6)$$

Where x represents the original value, while x' represents the transformed value. The addition of 1 is used to allow zero values to be transformed. Log transformation reduces the influence of extreme values and helps make the data distribution more balanced before clustering.

2.3.7 StandardScaler / Z-score Normalization

$$z = \frac{x - \mu}{\sigma} \quad (7)$$

Where x represents the original value, μ represents the mean value of the variable, and σ represents the standard deviation. This normalization process ensures that each variable has a comparable scale before being used in distance-based clustering algorithms such as K-Means.

2.4 Process Clustering

This study applied four clustering algorithms: K-Means, Hierarchical Clustering, Fuzzy C-Means, and DBSCAN. K-Means was selected as the primary method due to its simplicity, computational efficiency, and widespread adoption in health data clustering studies, including HIV/AIDS-related analyses [10]. Hierarchical Clustering was used as a comparison method because it produces a layered clustering structure that can reveal relationships between objects based on similarity.

Fuzzy C-Means was included because it allows a single UPK to have membership degrees in multiple clusters simultaneously. This characteristic is suitable for health data, given that a service unit may exhibit characteristics not exclusively possessed by a single group. The use of Fuzzy C-Means in clustering indicators of infectious illnesses has been demonstrated [11]. DBSCAN is used due to its ability to form clusters based on data density and detect potential noise or outliers. Comparative analyses involving K-Means, DBSCAN, and Hierarchical Clustering have been used in previous studies to evaluate differences in clustering quality among algorithms [12]. Comparisons between K-Means, Fuzzy C-Means, and Hierarchical Clustering have also been applied to identify the most suitable method based on data characteristics [13].

For the K-Means, Hierarchical Clustering, and Fuzzy C-Means methods, the number of clusters was tested within the range of K = 2 to K = 8. The lower bound of K = 2 was chosen because a single cluster does not produce meaningful group separation and cannot be evaluated using internal validation metrics such as the Silhouette Score. Meanwhile, the upper limit of K = 8 was chosen to provide a sufficiently wide range of cluster candidates while still avoiding excessive cluster fragmentation given the relatively small final dataset size of 29 observations. The optimal number of clusters was then determined empirically by comparing the results for each value of K using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Although the final sample size was relatively small, the clustering analysis in this study was intended as an exploratory approach to identify service performance patterns rather than to generate population-level inference. To improve clustering stability and reduce sensitivity to sample variation, preprocessing procedures including log transformation and Z-score normalization were applied, and robustness was further assessed through sensitivity analysis by comparing clustering patterns across multiple clustering algorithms

and different values of K. The relative consistency of the clustering results and validation metrics across these approaches supports the stability of the selected clustering structure. Nevertheless, the findings should still be interpreted cautiously, as clustering results derived from small datasets may remain sensitive to the inclusion or exclusion of individual observations.

For DBSCAN parameters were determined through exploratory tuning and nearest-neighbor distance inspection. A k-distance graph using the 2-nearest neighbor distance showed an elbow pattern around eps = 1.2 after data standardization. Therefore, eps = 1.2 was selected as the neighborhood radius, while min_samples = 2 was chosen considering the relatively small dataset size.

2.5 Evaluation Result

Clustering results were evaluated using three internal validity metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The Silhouette Score measures intra-cluster cohesion and inter-cluster separation. The Davies-Bouldin Index assesses cluster separation quality, with lower values indicating better results. The Calinski-Harabasz Index evaluates within-cluster compactness and between-cluster distance, with higher values reflecting superior clustering quality. The use of quantitative evaluation metrics ensures that cluster selection is not based solely on visual inspection, but is supported by comparable, algorithm-agnostic measures [18],[19],[20].

The quality of the clustering results was evaluated using three internal validation metrics: the Silhouette Score, the Davies-Bouldin Index, and the Calinski-Harabasz Index. These three metrics are commonly used to assess cluster quality based on data compactness within clusters and separation between clusters [14], [21]. The Silhouette Score is used to measure the extent to which a data point is closer to its own cluster compared to other clusters[14],[21]. The formula for the Silhouette Score is shown as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{8}$$

Where $a(i)$ is the average distance of the data to other members in the same cluster, while $b(i)$ is the average minimum distance of the data to the nearest other cluster. The Silhouette Score ranges from -1 to 1, where a higher value indicates better clustering results [14],[22].

The Davies-Bouldin Index is used to measure the ratio between intra-cluster cohesion and inter-cluster distance. The formula for the Davies-Bouldin Index is shown as follows:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \tag{9}$$

Where S_i and S_j represent the dispersion of data within clusters, while M_{ij} represent the distance between cluster center i and cluster j. A lower DBI value indicates more compact and more distinct cluster[14],[21],[22].

The Calinski-Harabasz Index is used to compare the dispersion between clusters with the dispersion within clusters. The formula for the Calinski-Harabasz Index is shown as follows:

$$CHI = \frac{BCSS/(k-1)}{WCSS/(n-k)} \tag{10}$$

Where BCSS is the between-cluster sum of squares, WCSSa is within-cluster sum of squares, k the number of clusters, and n is the number of observations. A higher CHI value indicates a better clustering structure [14],[21],[23].

2.6 Tools

Data analysis was conducted using the Python programming language on the Google Colab platform. Libraries used include Pandas for data manipulation, NumPy for numerical operations, Scikit-learn for preprocessing, K-Means, Hierarchical Clustering, DBSCAN, and clustering evaluation, and Scikit-fuzzy for Fuzzy C-Means

implementation. Clustering results were visualized using Matplotlib and Seaborn, while Principal Component Analysis (PCA) was applied to display cluster separation patterns in two-dimensional space.

3. Result and Discussions

3.1 Data Preparation Results

The values illustrate variation in screening coverage and positivity rates across UPKs before the clustering analysis was conducted.

Table 2. Selected Sample of HIV Screening Dataset

Year	UPK	Tested	Pstv	Pos (%)	M_Tested	F_Tested
2025	UPK1	340	4	0.0117	150	**190**
2025	UPK 3	2,040	64	0.0313	899	**1141**
2025	UPK 5	27	0.0	0.0	12	**15**
2025	UPK 8	43	0.0	0.0	15.0	**28**
2025	UPK 11	1	0.0	0.0	1.0	**0.0**
2024	UPK 14	359	3	0.0083	99	**260**
2025	UPK 17	1948	32	0.0164	718	**1230**
2025	UPK 21	1	1	1	1	**0.0**
2025	UPK 25	117	0	0	59	**58**
2025	UPK 29	35	2	0.0571	22	**13**

Table 2, presents selected observations from the final HIV screening dataset used in this study after the data selection and preprocessing stages. The table shows several key variables, including the total number of individuals tested, the number of HIV-positive cases, positivity rate, gender-based testing distribution, and gender-based positive cases across different Health Service Units (UPKs). These selected samples illustrate the variability of HIV screening characteristics among UPKs in Teluk Bintuni Regency before the clustering process was performed

3.2 Cluster Evaluation and Method Comparison

After the clustering process is performed using four algorithms, the next step is to determine the optimal number of clusters and the best method through internal evaluation. Table 3 presents the sensitivity analysis on the number of clusters using internal evaluation metrics, including Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

Table 3. Sensitivity Analysis on Number of Clusters

k	Inertia	Silhouette	DBI	CHI	Min Cluster Size	Singleton
2	87.090	0.629	0.755	17.954	2	FALSE
3	55.167	0.372	0.838	21.169	2	FALSE
4	39.816	0.365	0.554	22.014	1	TRUE
5	28.058	0.391	0.443	25.007	1	TRUE
6	20.122	0.322	0.642	28.548	1	TRUE
7	14.404	0.304	0.560	33.243	1	TRUE
8	9.934	0.337	0.484	40.789	1	TRUE

The range of K=2 to K=8 was selected because K=1 cannot yet be considered clustering, since all data are in a single group. The value K=2 is used as the minimum number of clusters to observe the initial separation of the data, while K=8 is chosen as the upper limit so that the number of clusters is not too large compared to

the relatively small dataset size. With a limited number of observations, using a value of K that is too large can result in clusters with very few members or singleton clusters, making the interpretation less stable.

Although K = 2 produced the highest Silhouette Score, the resulting clustering structure mainly separated the dataset into two broad groups and did not adequately capture the heterogeneity of HIV screening service characteristics among UPKs. In practice, the K = 2 solution provided limited interpretative value because it merged observations with substantially different screening profiles into larger generalized clusters.

In contrast, K = 3 produced a more informative and operationally meaningful segmentation without generating singleton clusters. The three-cluster solution allowed differentiation between UPKs with intermediate screening performance, highly active screening services, and units with extremely low testing coverage accompanied by high apparent positivity rates. This structure was considered more relevant for supporting targeted HIV intervention planning and resource allocation at the local level.

Therefore, although the K = 3 solution showed a lower Silhouette Score than K = 2, it provided a better balance between statistical quality, cluster stability, and substantive interpretability. The selection of K = 3 was thus based not solely on maximizing internal validation metrics, but also on the practical usefulness and interpretability of the resulting cluster profiles.

Table 4. Cluster Evaluation Result

Algorithm	C	Noise	Sil.	DBI	CHI	Singleton
K-Means	3	0	0.3718	0.8380	21.1688	FALSE
Hierarchical	3	0	0.3626	0.8693	20.9850	FALSE
Fuzzy C-Means	3	0	0.3418	0.7166	18.2187	TRUE
DBSCAN	1	4	-	-	-	FALSE

Based on Table 4, K-Means produced three clusters with the highest Silhouette Score, namely 0.3718, and did not form any singleton clusters. K-Means performed better in this study because the data had been log-transformed and standardized using z-scores, thereby making the distances between observations more balanced. Additionally, the relatively small dataset and the use of numerical features align with the characteristics of K-Means, which is based on distance to the centroid. Hierarchical Clustering also produced three clusters without any singleton clusters, but it had a slightly lower Silhouette Score, namely 0.3626. Fuzzy C-Means produced the lowest Davies-Bouldin Index value, namely 0.7166, but formed a singleton cluster, indicating that there was one cluster containing only a single observation. This condition can reduce the reliability of the clustering results because such a cluster likely represents isolated or extreme observations rather than stable group patterns. Therefore, although Fuzzy C-Means yields the lowest Davies-Bouldin Index value, the resulting cluster structure is considered less reliable for interpretation compared to K-Means and Hierarchical Clustering, which do not form singleton clusters. Meanwhile, DBSCAN formed only one cluster and identified four data points as noise, so internal evaluation metrics could not be calculated because there was only one cluster. DBSCAN is less suitable for use on this research dataset because the number of observations is relatively small and the features used are low-dimensional. Additionally, the data distribution does not show clear density separation, so DBSCAN was unable to form meaningful clusters. This condition causes most of the data to be grouped into a single main cluster, while several other observations are identified as noise. Based on considerations of evaluation values, cluster structure stability, and ease of substantive interpretation, K-Means was selected as the primary method for further interpretation, while the other three methods were used as comparators.

Theoretically, the results of this study indicate that K-Means remains relevant for use on small-scale health data provided that numerical features have undergone adequate transformation and standardization. The use of positivity rate, log-transformed screening volume, and gender-based ratios helps create a more balanced data representation, ensuring that differences between UPKs are not determined solely by the absolute number of tests or positive cases. These findings contribute to the literature on data mining and health informatics, particularly regarding the application of unsupervised learning to support the mapping of local health services in areas with limited data. Since K-Means demonstrated the most stable and interpretable performance compared to other methods, further analysis focused on the cluster profiles resulting from K-Means.

Table 5. Distribution of Observation by Clustering Method

Algorithm	Clusters	0	1	2	Noise
K-Means	3	18	2	9	**0**
Hierarchical Clustering	3	2	16	11	**0**
Fuzzy C-Means	3	1	18	10	**0**
DBSCAN	1	25	–	–	**4**

Table 5 shows that K-Means, Hierarchical Clustering, and Fuzzy C-Means each form three clusters, while DBSCAN produces only one main cluster and identifies four observations as noise. K-Means forms clusters with 18, 2, and 9 observations respectively, without forming any singleton clusters. Hierarchical Clustering also produces three clusters without any singleton clusters, although the distribution of members differs from that of K-Means. Fuzzy C-Means forms three clusters, but one of the clusters contains only a single observation, indicating the presence of a singleton cluster. Meanwhile, DBSCAN is considered less suitable for further interpretation because it is unable to separate the data into meaningful clusters.

3.3 Cluster Profile and Visualization

Table 6. Profile of K-Means Clusters Based on HIV Screening Indicators

C	N	Tested	Positive	Pos. (%)	Male (%)	Female (%)
0	18	116.61	0.61	0.98	38.6	**61.36**
1	2	1.00	0.50	50.00	100	**0.0**
2	9	1138.44	27.00	1.99	41.93	**58.07**

Based on Table 6, the K-Means results formed three clusters with different HIV screening service characteristics. Cluster 0 is the largest group with 18 observations, having an average number of tests of 116.61 and an average number of positive cases of 0.61, with an average positivity rate of 0.0098. These characteristics indicate that Cluster 0 represents a group with an intermediate screening profile, meaning that screening services are in place but the number of positive cases detected remains relatively low. From a programmatic perspective, this cluster may benefit from strengthening targeted screening strategies, improving community outreach, and increasing early detection efforts among key populations to improve case finding performance.

Cluster 1 must be interpreted with caution. This cluster consists of only two observations with a very low number of tests, so the high positivity rate reflects the influence of a small denominator rather than the actual epidemiological conditions. Thus, this figure cannot be directly interpreted as an indication of a clinically high rate of HIV transmission. These two observations were retained in the analysis because they do not contain missing values and still represent the conditions of the UPKs services during the relevant reporting period. However, Cluster 1 is more appropriately understood as a group with very low screening coverage and potential data anomalies, rather than as a group with definitively high epidemiological risk. Therefore, findings in this cluster require further verification, either through examination of source data or direct confirmation with local HIV program managers. According to information from the local health authority, geographical barriers and social factors may contribute to the low testing coverage in these service units. In addition, the HIV/AIDS Information System at the Ministry of Health was recently upgraded. Therefore, continuous reminders and supervision for service-level operators regarding report completeness and data entry accuracy are important to improve data quality and program monitoring in this cluster.

Meanwhile, Cluster 2 consists of 9 observations with the highest average number of tests, namely 1,138.44, an average of 27.00 positive cases, and an average positivity rate of 0.0199. These characteristics indicate that Cluster 2 represents service units with the most active HIV screening activities compared to other clusters. Programmatically, this cluster may serve as a reference for best practices in HIV screening implementation, particularly regarding service accessibility, screening outreach, and reporting performance.

Maintaining testing capacity, ensuring linkage to care for positive cases, and strengthening monitoring systems remain important priorities for this group.

Overall, these results indicate differences in service patterns among UPKs, ranging from groups with moderate screening profiles, groups with high positivity rates but very low testing coverage, to groups with more active screening services. These findings may help support more targeted program planning, resource allocation, and monitoring strategies according to the characteristics of each cluster. To clarify the separation patterns between clusters, the K-Means results were then visualized using Principal Component Analysis (PCA), as shown in Figure 2.

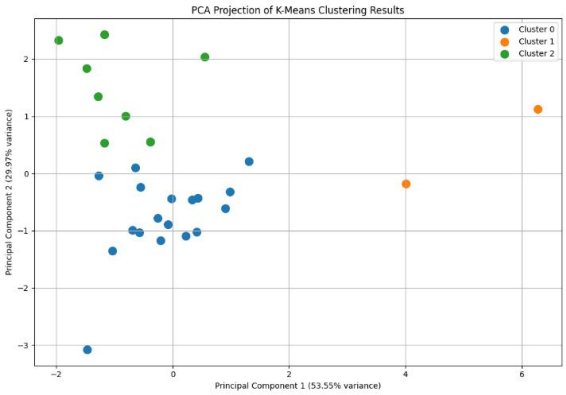

Figure 2. PCA Visualization of K-Means Clustering Results

Figure 2 shows the separation patterns of the three K-Means clusters in a two-dimensional space. Each point represents one UPK-year observation, while the color differences indicate cluster membership.

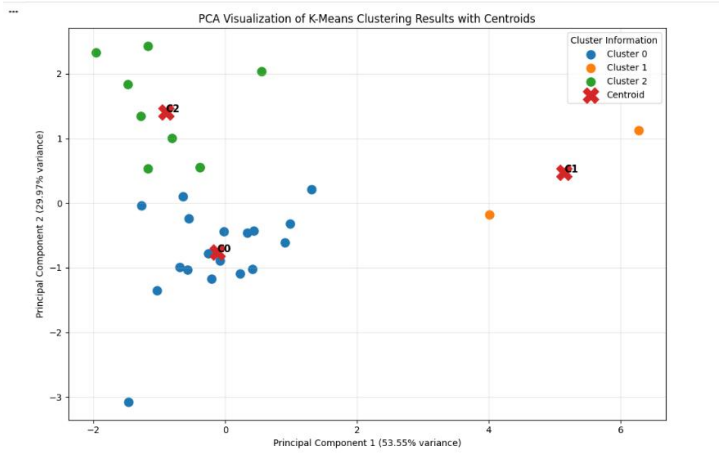

Figure 3. PCA Visualization of K-Means Clustering Results with Centroid Markers

Figure 3 shows the PCA projection of the K-Means clustering results. Principal Component 1 (PC1) explained 53.55% of the total variance, while Principal Component 2 (PC2) explained 29.97%, resulting in a cumulative explained variance of approximately 83.52%. This indicates that the two-dimensional PCA projection adequately represented the underlying clustering structure and screening pattern variability among UPKs. PC1 and PC2 were derived from the standardized HIV screening data features, and centroid markers indicate the center positions of each cluster in PCA space. The separation of clusters observed in the PCA plot demonstrates that the K-Means model was capable of distinguishing different HIV screening service patterns across UPKs, particularly between units with high screening activity and those with very low testing coverage.

The findings of this study are broadly consistent with prior research applying K-Means to HIV data in Indonesia, while offering a more operationally granular perspective. Most previous studies analyzed disease data at the provincial or district level using aggregated case counts [6],[7],[8]. The present study extends this approach by conducting the analysis at the Health Service Unit (UPK) level, allowing more detailed identification of local disease patterns and supporting targeted public health interventions. This finer granularity enables a direct characterization of service heterogeneity across individual units, which is more relevant for local intervention planning. Furthermore, the present study introduces derived variables positivity rate and sex-disaggregated ratios that provide a multidimensional view of screening performance, extending beyond absolute case numbers. This methodological innovation allows for a more comprehensive profiling of service patterns and highlights units with high positivity but low coverage that may otherwise be overlooked. In this way, the study not only confirms the applicability of K-Means in HIV data analysis but also contributes a novel operational framework for evidence-based decision support in resource-limited settings. Among the four algorithms compared, K-Means achieved the highest Silhouette Score (0.3718), produced no singleton clusters, and yielded three substantively interpretable groups. Its superiority is attributable to the preprocessing pipeline: logarithmic transformation and StandardScaler normalization reduced distributional skewness and balanced inter-observation distances, aligning the data structure with the Euclidean distance assumptions underlying K-Means. Hierarchical Clustering was a close alternative (Silhouette = 0.3626), while Fuzzy C-Means, despite its lowest Davies-Bouldin Index (0.7166), formed a singleton cluster that undermined result reliability. DBSCAN was unsuitable for this dataset, producing only one main cluster due to the absence of clear density-based separation. These results reinforce that algorithm selection should integrate validation scores, structural stability, and interpretability rather than relying on any single metric.

The three cluster profiles identified carry direct implications for HIV service planning in Teluk Bintuni Regency. Cluster 0 the largest group with 18 observations represents UPKs with an intermediate screening profile: moderate testing coverage and a low positivity rate (mean 0.98%). These units appear to have functional screening mechanisms but may benefit from outreach expansion and improved post-test counseling to increase case detection. Cluster 2, with the highest mean number of tests (1,138.44) and the greatest mean positive cases (27.00, positivity rate 1.99%), represents the most active screening units; these UPKs require strengthened referral pathways, case follow-up systems, and counseling capacity to manage their comparatively higher workload. Cluster 1, comprising only two observations with a positivity rate of 50%, must be interpreted with caution, as this extreme value more likely reflects a very small testing denominator than a genuinely elevated transmission rate. Nonetheless, its presence is analytically relevant as an early warning signal warranting field verification, as it may indicate UPKs with severely limited screening coverage, incomplete recording, or geographic concentration of at-risk populations shaped by access barriers and social stigma. Collectively, these patterns support a differentiated service delivery approach rather than a uniform program strategy to ensure that resources are allocated in proportion to the specific service characteristics of each unit.

From a policy perspective, Cluster 1 should be treated as a priority for targeted intervention because it combines very low testing coverage with an apparently high positivity rate. Local health authorities should not directly interpret this cluster as evidence of high HIV transmission, but should first conduct field verification, improve screening coverage, and review data recording practices in the related UPKs. If the pattern is confirmed, these units may require intensified outreach, mobile testing services, and stronger linkage-to-care mechanisms. Recommendations to strengthen outreach in Cluster 1 are consistent with the findings of recent studies. Previous studies have demonstrated that Mobile VCT services are effective in reaching high-risk communities and increasing HIV testing participation [24]. Likewise, community-based outreach programs have been shown to increase HIV testing uptake and promote preventive behaviors among key populations [25].

This study contributes to the literature on data mining and health informatics in several respects. First, it shifts the unit of analysis from aggregate administrative levels to individual UPKs, a perspective that remains underexplored in Indonesia, particularly in the Papua region, and produces operationally actionable findings for local program planners. Second, the use of derived variables combined with logarithmic transformation and feature normalization demonstrates that appropriate feature engineering can meaningfully improve cluster separation quality even when the dataset is small a finding with practical methodological relevance for studies working with limited local health data. Third, the multi-metric evaluation framework employed

here combining Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index alongside structural stability checks provides a more robust basis for algorithm selection than single-metric approaches, and may serve as a replicable evaluation protocol for similar health informatics studies. Fourth, PCA visualization showed that the majority of inter-UPK variance in HIV screening patterns is captured by two principal components, indicating a relatively simple latent data structure that could support the design of efficient HIV indicator monitoring systems at the district level.

This study is subject to several limitations. First, the final dataset comprised only 29 unit-year observations after 15 incomplete records were excluded because several Health Service Units did not conduct HIV screening activities during the 2024 reporting period. Therefore, the missing values were not caused by data entry errors, but reflected the absence of screening activities in those UPKs. Since key variables such as the number of individuals tested, HIV-positive cases, and gender-based screening indicators were unavailable, these observations could not be meaningfully imputed, constraining the statistical validity and generalizability of result; the findings should therefore be understood as exploratory rather than definitive. Second, no imputation was performed given the local, semi-structured nature of the source data, as imputed values on a small dataset risked distorting the actual service conditions; the resulting complete-case analysis, while methodologically conservative, further reduced the usable sample. Third, all clustering variables were quantitative screening outputs; the absence of contextual covariates such as UPK distance from the regency center, healthcare workforce availability, or area sociodemographics limits the study's capacity to causally explain cluster membership. Fourth, the study's geographic scope is confined to Teluk Bintuni Regency, which has specific geographic and demographic characteristics; direct extrapolation to other regencies in Papua or elsewhere in Indonesia requires replication with locally representative data. Future research is recommended to address these limitations by incorporating longer observation periods, a greater number of UPKs, additional contextual variables, and external validation mechanisms to improve the stability, generalizability, and practical utility of the clustering results.

4. Conclusions and Future Works

This study demonstrated that K-Means clustering can be used to group Health Service Units (UPKs) in Teluk Bintuni Regency based on HIV screening patterns. Using 29 active unit-year observations from 2024–2025, the analysis produced three interpretable clusters that reflect differences in screening coverage, positivity rate, and service workload across UPKs.

Compared with Hierarchical Clustering, Fuzzy C-Means, and DBSCAN, K-Means provided the most stable and interpretable clustering result for this dataset. Although the evaluation metrics showed only moderate cluster separation, the absence of singleton clusters and the substantive interpretability of the three groups supported the selection of K-Means as the main clustering method.

The findings indicate that HIV screening services in Teluk Bintuni should not be treated uniformly. UPKs with intermediate screening performance may require outreach expansion, highly active screening units need stronger referral and follow-up systems, while units with very low testing volume and high apparent positivity require field verification to distinguish between actual epidemiological risk and data quality issues.

This study contributes to HIV service mapping by applying clustering at the UPK level, a perspective that remains limited in Indonesian HIV data mining studies, particularly in Papua. However, the findings should be interpreted as exploratory because the dataset was small and did not include contextual variables such as geographic access, health workforce availability, and sociodemographic characteristics.

Future research should extend the observation period, include more UPKs or regencies, add contextual service variables, and conduct external validation with local health practitioners. Further comparison with other clustering methods suitable for small-scale health datasets may also improve the robustness of future analysis.

5. References

[1] R. Kementerian Kesehatan, “Profil Kesehatan Indonesia 2023,” Jakarta, 2024.

[2] A. Kasimur, Y. Ruru, S. Rantetoding, A. Zainuri, and M. Akbar, "Analisis Pembiayaan Layanan Kesehatan HIV Dan AIDS Berbasis Standar Pelayanan Minimal Dengan Pendekatan District Health Account Di Kabupaten Teluk Bintuni Provinsi Papua Barat," *Journal of Innovative and Creativity*, vol. 5, no. 2, pp. 5908–5929, 2025.

[3] M. N. Kadafi and A. Finandhita, "Application of Data Mining in Disease Clustering at Klinik Keluarga," *Komputa : Jurnal Ilmiah Komputer dan Informatika*, vol. 13, no. 1, pp. 33–42, Apr. 2024, doi: 10.34010/komputa.v13i1.11273.

[4] R. Sebastian and C. Juliane, "Application of Data Mining to Group the Spread of Covid-19 in West Java Province, Indonesia Using the K-Means Algorithm," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 5, no. 2, pp. 83–90, 2022.

[5] B. D. Cahyani, I. A. Putri, P. Utomo, and D. Sofia, "Penerapan Data Mining Dalam Analisis Klasterisasi Penyebaran Penyakit Hiv Menggunakan Algoritma K-Means," *Edutech*, vol. 24, no. 2, pp. 1239–1255, 2025, doi: 10.17509/e.v24i2.84914.

[6] R. I. L. Sinaga, W. Saputra, and H. Qurniawan, "Pengelompokan Jumlah Kasus Penyakit Aids Berdasarkan Provinsi Menggunakan Metode K-Means," *Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, vol. 2, no. 2, pp. 99–107, 2021, doi: https://doi.org/10.30645/kesatria.v2i2.64.

[7] H. Noor, A. Dharmawati, and T. W. Qur'ana, "Penerapan Algoritma K-Means Clustering Analysis Pada Kasus Penderita Hiv/Aids (Studi Kasus Kabupaten Banjar)," *Technologia: Jurnal Ilmiah*, vol. 12, no. 2, p. 72, 2021, doi: 10.31602/tji.v12i2.4573.

[8] W. Rosida and Y. A. Wijaya, "Klasterisasi Penyakit HIV/AIDS di Jawa Barat Menggunakan Algoritma K-Means Clustering," *Blend Sains Jurnal Teknik*, vol. 1, no. 4, pp. 306–315, 2023, doi: 10.56211/blendsains.v1i4.235.

[9] K. Kodratul Munawar and A. Irma Purnamasari, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING PADA KLASTERISASI KASUS HIV DI JAWA BARAT," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 2, pp. 1092–1099, Aug. 2023, doi: 10.36040/jati.v7i2.6372.

[10] M. B. Gesicho, M. C. Were, and A. Babic, "Evaluating performance of health care facilities at meeting HIV-indicator reporting requirements in Kenya: an application of K-means clustering algorithm," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 6, Dec. 2021, doi: 10.1186/s12911-020-01367-9.

[11] M. Y. Matdoan, L. Igo, R. Rumeon, R. Fadhilah, and N. S. Laamena, "Penerapan Algoritma K-Means Untuk Klusterisasi Kabupaten/Kota Berdasarkan Tingkat Kemiskinan Di Kepulauan Maluku Dan Papua," *Jurnal Sains Matematika dan Statistika*, vol. 10, no. 1, p. 1, 2024, doi: 10.24014/jsms.v10i1.21260.

[12] P. Liu, Z. Wang, N. Liu, and M. A. Peres, "A scoping review of the clinical application of machine learning in data-driven population segmentation analysis," *Journal of the American Medical Informatics Association*, vol. 30, no. 9, pp. 1573–1582, Aug. 2023, doi: 10.1093/jamia/ocad111.

[13] Z. W. Alexiou *et al.*, "Using country-level variables to discover country clusters beyond traditional health policy and performance metrics: An unsupervised machine learning approach for HIV healthcare delivery and financing," *PLOS Global Public Health*, vol. 5, no. 12, p. e0004583, Dec. 2025, doi: 10.1371/journal.pgph.0004583.

[14] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, Apr. 2022, doi: 10.1016/j.engappai.2022.104743.

[15] M. Verhoeff *et al.*, "Clusters of medical specialties around patients with multimorbidity – employing fuzzy c-means clustering to explore multidisciplinary collaboration," *BMC Health Services Research*, vol. 23, no. 1, p. 975, Sep. 2023, doi: 10.1186/s12913-023-09961-z.

[16] F. Han, Z. Zhang, H. Zhang, J. Nakaya, K. Kudo, and K. Ogasawara, "Extraction and Quantification of Words Representing Degrees of Diseases: Combining the Fuzzy C-Means Method and Gaussian Membership," *JMIR Formative Research*, vol. 6, no. 11, p. e38677, Nov. 2022, doi: 10.2196/38677.

[17] D. Chicco, L. Oneto, and D. Cangelosi, "DBSCAN and DBCV application to open medical records heterogeneous data for identifying clinically significant clusters of patients with neuroblastoma," *BioData Mining*, vol. 18, no. 1, p. 40, Jun. 2025, doi: 10.1186/s13040-025-00455-8.

[18] D. P. S. Dewi, "Analisis Clustering Penyebaran Hiv Di Karawang Berdasarkan Kecamatan Dengan Algoritma K-Means," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4878.

[19] S. S. Putro, M. Syarief, and E. M. S. Rochman, "Penerapan Algoritma K-Means, Dbscan, Dan Ahc Untuk Clustering Kualitas Garam Pada Pt. Garam (Persero)," *Multitek Indonesia*, vol. 18, no. 2, pp. 114–125, 2025, doi: 10.24269/mtkind.v18i2.8501.

[20] A. Ripai, Y. Baskoro, and Imelda, "Perbandingan Algoritma K-means, Fuzzy C-means, dan Hierarchical clustering pada Klasterisasi Tingkat Pemahaman Siswa dalam Pembelajaran Berdiferensiasi," *Jurnal Algoritma*, vol. 22, no. 2, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2890.

[21] M. Kossakov, A. Mukasheva, G. Balbayev, S. Seidazimov, D. Mukammejanova, and M. Sydybayeva, "Quantitative Comparison of Machine Learning Clustering Methods for Tuberculosis Data Analysis," in *CIEES 2023*, Basel Switzerland: MDPI, Jan. 2024, p. 20. doi: 10.3390/engproc2024060020.

[22] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.

[23] I. P. Sari, F. Ramadhani, O. K. Sulaiman, A. Satria, and A. A. Manurung, "Clustering Hiv / Aids Disease Using K-Means," *Proceeding International Seminar of Islamic Studies*, vol. 5, no. 1, pp. 1668–1676, 2024.

[24] C. M. Sari, N. K. A. Wulandari, and K. P. Sedana, "Intervensi Mobile VCT untuk Meningkatkan Akses Layanan Tes HIV di Wilayah Lokalisasi," *Jurnal Pengabdian Masyarakat Sundaram*, vol. 3, no. 1, pp. 53–66, Jun. 2025, doi: 10.52073/jpms.v3i1.62.

[25] L. Wahyuniar, A. Sutrisna, D. B. Wicaksana, and R. Habsi, "Developing The Community Outreach Effectiveness Model (COEM) for HIV Prevention Among Key Populations in Indonesia: Evidence from The IBBS," *Media Kesehatan Masyarakat Indonesia*, vol. 21, no. 4, pp. 359–369, Dec. 2025, doi: 10.65844/mkmi.v21i4.45782.