# K-Means Binary Search Centroid with Dynamic Cluster for Java Island Health Clustering

**Muhammad Andryan Wahyu Saputra[-1*], Muhammad Faisal[-2], Ririen Kusumawati[-3]**

Magister Informatics
UIN Maulana Malik Ibrahim
Malang, Indonesia
210605210010@student.uin-malang.ac.id[1], mfaisal@ti.uin-malang.ac.id[2], ri2n.kusumawati@gmail.com[3]
(*) Corresponding Author

## Abstract

This study is focused on determining the health status of each district/city in Java using the K-means Binary Search Centroid and Dynamic Kmeans algorithms. The research data uses data on the health profile of Java Island in 2020. Comparative algorithms were tested using the Davies Bound Index and Calinski-Harabasz Index methods on the traditional k-means algorithm and dynamic binary search centroid k-means. Based on the test, 5 clusters were found in the distribution area, including 11 regions with very high health quality cluster 1, 24 regions with high health quality, 28 regions with moderate health quality, and 28 clusters 4 with low health quality, 45 regions, and cluster 5 with poor health quality is 11 regions, with the best validation value of DBI 1.8175 and CHI 67.7868. Overall optimization of the dynamic k-means algorithm based on binary search centroid results in a better average cluster quality and a smaller number of iterations than the traditional k-means algorithm. The test results can be used as one of the best methods in evaluating the level of health in the Java Island area and a reference for decision-making in determining policies for related agencies.

Keywords: Clustering; Binary Search Centroid; Dynamic K-Means; Java Island Health Profile

## Abstrak

*Penelitian ini difokuskan untuk menentukan derajat kesehatan setiap kabupaten/kota dalam Pulau Jawa menggunakan algoritma K-means Binary Search Centroid (KBSC) dan Dynamic Kmeans (DK). Data penelitian menggunakan data profil kesehatan Pulau Jawa tahun 2020. Perbandingan algoritma terbaik diuji tingkat validitasnya menggunakan metode Davies Bound Index (DBI) dan Calinski-Harabasz Index (CHI) pada algoritma k-means tradisional dan k-means dinamis binary search centroid. Berdasarkan pengujian diperoleh 5 jumlah klaster di wilayah sebaran, di antaranya klaster 1 dengan kualitas kesehatan sangat tinggi adalah 11 kabupaten/kota, klaster 2 dengan kualitas kesehatan tinggi adalah 24 kabupaten/kota, dan klaster 3 dengan kualitas kesehatan cukup adalah 28 kabupaten/kota. dan kualitas kesehatan 45 kabupaten/kota untuk klaster 4 dengan kualitas kesehatan rendah serta klaster 5 dengan kualitas kesehatan sangat rendah adalah 11 kabupaten/kota, dengan nilai validasi terbaik DBI 1.8175 dan CHI 67.7868. Secara keseluruhan optimasi algoritma k-means dinamis berbasis binary search centroid menghasilkan rata-rata kualitas klaster yang lebih baik dan jumlah iterasi yang lebih kecil dibanding algoritma k-means tradisional. Hasil pengujian dapat digunakan sebagai salah satu metode terbaik dalam evaluasi tingkat kesehatan yang ada di wilayah khususnya Pulau Jawa serta acuan pengambilan keputusan dalam menentukan kebijakan bagi instansi terkait.*

*Kata kunci: Klastering; Binary Search Centroid; Dynamic K-Means; Profil Kesehatan Pulau Jawa*

## INTRODUCTION

Indicators of the success of a country's development can be seen from the level of achievement of the country in providing health insurance. In Indonesia, the government has set several indicators through the Ministry of Health that become a benchmark for the progress of health development at all levels of Indonesian territorial units (from the province, district, to sub-district).

The Ministry of Health collects and processes population health data annually through health offices across various regions to produce provincial and district/city health rankings. These

processed results and rankings become essential references. For example, local governments (Pemda) formulate intervention plans that are more targeted as promotional materials so that local governments are more motivated to improve their regional health rankings, as the basis for the central government to formulate seriousness or especially for health problem areas (DBKBK), as the basis for determining the allocation of central government health funding to the regions also serves to support the Ministry of Vulnerable Regions (KMPDT) in regional/urban development (L. Suryani, 2021).

Regional clustering is carried out to determine the region's health level, especially Java Island. The Central Bureau of Statistics notes that Java Island has 29 regencies and nine cities. So far, the grouping of environmental health indicator data is still based on manual computing techniques, where the calculations still have several problems, especially in data consistency. In terms of improving the health services of the Health Office, a system is needed that groups healthy areas based on environmental health data so that counselling, services, and assistance can be more accurate and on target.

At the time of the condition where the regional grouping data do not want to be determined as needed, but want to produce segmentation from the distribution of the data, then to determine the assumption of the number of clusters it is not expected at the beginning, but the assumption of the number of clusters that will be formed must be sought. To handle this condition, there is the development of the K-means algorithm, namely the Dynamic Kmeans (DK) algorithm, which has a process to find the number of clusters without having to assume the assumption of the number of clusters (H. Santoso et al., 2022). In the research, Widiarina and Romi proposed a dynamic cluster algorithm on the K-means algorithm to determine the number of clusters (k) to produce optimal cluster quality and provide better and more precise potential segmentation results (W. Widiarina & R. S. Wahono, 2015). However, this algorithm has some of the same shortcomings as the K-means algorithm, namely the problem of determining the centroid point in the clustering process, which is still randomly selected. Based on research from Yugar Kumar and G. Sahoo, the problem of determining the centroid point is done by developing K-means into K-means Binary Search Centroid (KBSC), which has a process in determining the centroid point using a Binary Search technique approach (Akbari Gumilar & Yusrila Kerlooza, 2018). The results of this study prove that the K-means Binary Search Centroid (KBSC) algorithm has better intra- and

inter-cluster values than the K-means algorithm. However, the K-means Binary Search Centroid (KBSC) Algorithm has limitations in determining the number of clusters to form.

Therefore, it is proposed that combining the Dynamic Kmeans algorithm with the K-means Binary Search Centroid algorithm can complement the clustering process in the area of assuming the number of clusters and determining the cluster centroid point, to produce the best Davies-Bouldin Index value for case studies of regional clustering based on profiles. Health.

**RESEARCH METHODS**

In carrying out research, a research method is needed to increase the possibility of a research activity running smoothly. The research method acts as a framework for carrying out research so that by following this framework, research can run systematically, and it is hoped that research can be more effective and in a reasonable period (C. Kamila et al., 2021). The following is the framework of this research shown in Figure 1.
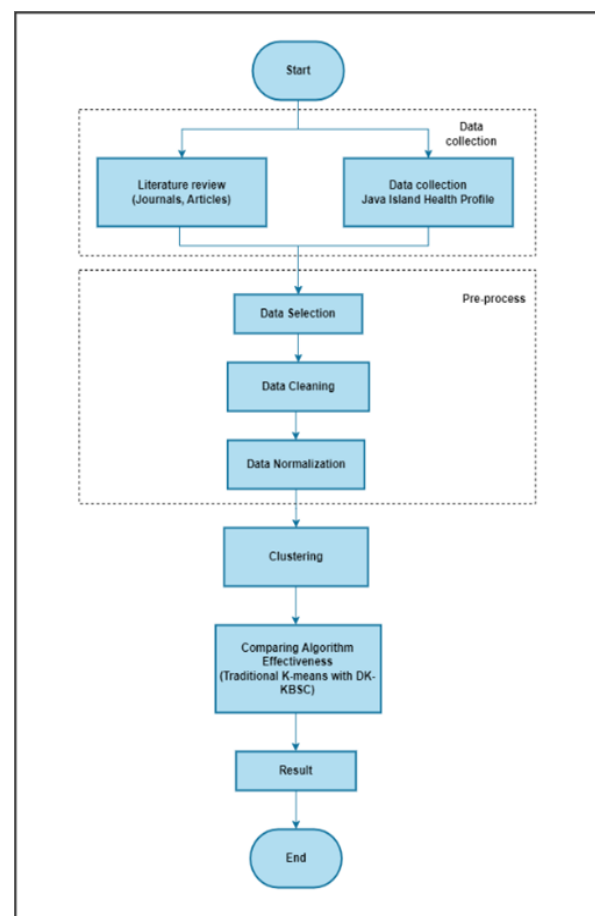


Figure 1. Research Flow Design

## A. Data Collection Stage

Data collection for this study cites publications from the Central Statistics Agency and the Java Island Health Service. This study uses data from 2020. This study will use datasets from 29 districts and nine cities on the island of Java. The variables used in this study are (Dinkes & BPS Pulau Jawa, 2020):

– Life expectancy
– Health Center Ratio
– Hospital Ratio
– Clinic Ratio
– Percentage of Eligible Sanitation RT
– Percentage of Drinking Water Facilities
– Percentage of Babies with LBW
– Complete Basic Immunization Percentage
– Percentage of Babies Who Get Exclusive Breastfeeding
– Percentage of Toddlers with Malnutrition
– Percentage of Tuberculosis Treatment Success
– Diarrhea Pain Rate (per 1k population)
– Covid-19 Healing Rate
– Covid-19 Death Rate
– Percentage of the Population Who Smokes
– Percentage of Services for Diabetes Mellitus Patients
– Percentage of Public Places Meet Health Requirements
–

## B. Data Pre-processing Stage

This stage consists of a data selection stage: selecting data from a set of operational data for extracting information, such as summarizing data for each city for each aspect. Then perform data cleaning, namely data containing data, removing data duplication, checking inconsistent data, and correcting errors in data. After that, normalize the data, which is the process of making several variables have the same value range, none of which is too big or too small so that it can make statistical analysis easier (C. Satria and A. Anggrawan, 2021).

## C. Clustering Stage

In this study, we will combine the DK and KBSC algorithms with the advantages of each algorithm to find the best number of clusters and determine the initial centroid point for each formation of the number of clusters. The combination of the two algorithms is shown in Figure 2.
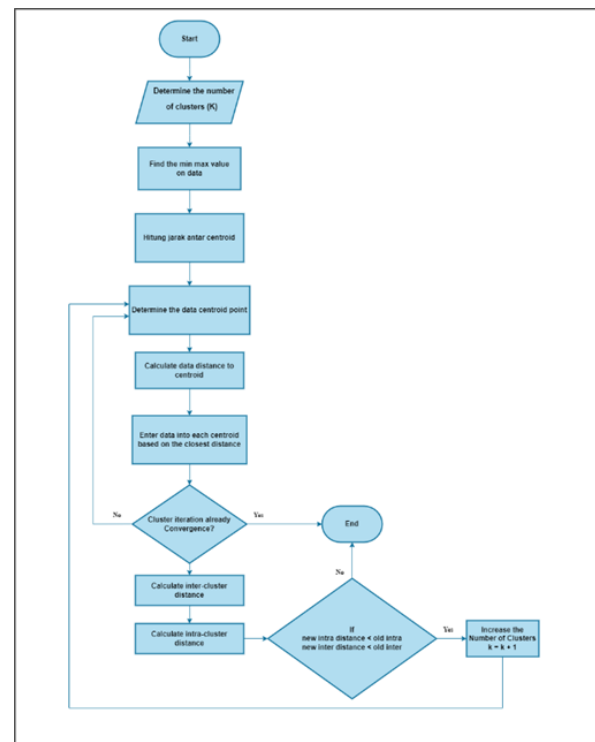


Figure 2. The Flow of Merging the Two Algorithms

Figure 2 describes the merging flow of the two algorithms. Flow chart. This centroid point will use a data search technique with a Binary Search approach (K. Ariasa, 2020). The process will divide the data area based on the number of clusters entered. The following is an explanation of the flowchart flow:

1. Enter the number of clusters as much as $k$ (Supriyatna et al., 2020).
2. Calculate the maximum and minimum values in the data for each data attribute.
3. Calculate the range between the centroid points, calculate the range between the centroid points in data in the following way:

$$M = \frac{\max(a_i) - min(a_i)}{k} \quad \text{......................................................... (1)}$$

4. Based on formula (1), it is used to calculate a value for the variable M which is the specific distance between the data centroid points before giving the results of the data centroid points. $min(a_i)$ the minimum value of each data attribute, $max(a_i)$ is the maximum value of each data attribute, and $k$ is the number of clusters to be formed.
5. Determine the data centroid point. To generate the centroid point, use the following formula:
$$C_k = min(a_i) + (k-1)M \quad (2)$$

6. Based on formula (2), CK K is the centroid for cluster k, $min(a_i)$ the minimum data value, and M range between data centroid points.

7. Group all data or objects with similarities by calculating the distance to the nearest centroid point. This distance is defined as the similarity of a data object with a predetermined centroid point. The proximity of two objects is determined based on the calculation of the distance between the objects. The distance theory can be used to calculate the distance of all data to each cluster centroid point. Euclidean Distance coordinates are formulated as follows (Bagus Muhammad et al., 2021) :

$$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x - x_{2j})^2 + \cdots + (x_{ni} - x_{nj})^2} \quad \text{...... (3)}$$

8. Equation (3) explains that $D(i,j)$ is the distance of the data to the center of clusters, then $x_{ni}$ is the data to the nth attribute of data and $x_{nj}$ is the center point or centroid of the Ken attribute. In the K-means dynamic cluster algorithm, intercluster is the minimum distance between the centroid points or the cluster's center with other cluster centers. This Inter is usually used to measure the minimum level of similarity in other clusters or the separation between clusters shown in the equation (C. Sreedhar et al., 2017).

$$inter = min(\sum_{i=0,j=0}^{N-1} \quad )dist(x_k, x_j) \forall k = 1,2,\dots,k-1 \text{ and } k = k+1,\dots,k_n \dots\dots (4)$$

9. Equation (4) explains that $x_k$ is the distance to the center of the previous cluster, while $x_j$ is the distance to the center of the next cluster, and $k$ is the distance to the cluster's center. Next, in the dynamic cluster algorithm on K-means, there is also an intra-term: the distance between the data and the centroid point in a cluster. Intra is used to measure a cluster's maximum similarity level, which is indicated by equation (5) (R. M. Alguliyev et al., 2019).

$$\sqrt{\frac{1}{n-1} \sum_{i-1}^{n} (x_i, x_m)^2} \dots\dots (5)$$

10. Equation (5) explains that $n$ is the amount of data, $x_i$ is data, and $x_m$ is the centroid

### D. Analysis Stage
The analysis was carried out since the determination of the data, the pre-processing process, and the implementation of the algorithm

calculations. The evaluation of clustering was tested for validity using several methods, namely Cluster Variance (V), Davies Bound Index (DBI), and Calinski-Harabasz Index (CHI). The equation used by the Davies-Bouldin Index will produce the most optimum value, namely the smallest value produced by equations (6) and (7) (Y. A. Wijaya et al., 2021).

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq 1}\{D_{i,j}\} \dots\dots (6)$$

$$D_{i,j} = \frac{\bar{d}_i + \bar{d}_j}{d_{i,j}} \dots\dots (7)$$

Based on equations (6) and (7), DBI is the value of the Davies-Bouldin Index, $k$ is the number of clusters, $D_{i,j}$ is the value of the distance between clusters $i$ and $j$. $di$ and $d$j are the average distance between the data in the temporary cluster $i$ and the cluster $j$. The smaller the Davies Bouldin Index (DBI) value obtained (non-negative $>= 0$), the better the cluster obtained from grouping using the clustering algorithm. The Calinski-Harabasz (CH) validity index calculates the comparison between the Sum of Squares. Values between clusters (SSB) as separation and the Sum of Square within-cluster (SSW) value as compactness multiplied by the normalization factor, which is the difference between the number of data and the number of clusters divided by the number of clusters minus one. The more excellent CH value indicates the number of the best clusters [(Luthfi et al., 2021). Suppose there is a data set with $k$ clusters and $N$ points data. For example $C_i$ is the *l*-th cluster where $x_i$ is the *i*-th point in the *l*-th cluster, $N_l$ is the number of points in the *l*-cluster, and $\bar{x}_i$ If the center point of the *l*-cluster, then the calculation of the CH validity index can be seen in the formula (8), (9), and (10) (A. F. Khairati et al., 2019).

$$SSW = \sum_{i=1}^{k}\sum_{x_i \in C_i}(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \dots\dots (8)$$
$$SSB = \sum_{i=1}^{k} N_i(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \dots\dots (9)$$
$$CH = \frac{trace(SSB)}{trace(SSW)} \times \frac{N-k}{k-1} \dots\dots (10)$$

To determine the evaluation of the model used, several experimental scenarios were carried out to determine the most suitable and accurate algorithm for evaluating algorithm performance. After all cluster results are formed, the algorithms are compared, and conclusions are drawn about which algorithms work optimally, the best algorithm accuracy, and the most optimal number of clusters based on the test data criteria.

Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

## RESULTS AND DISCUSSION

### A. Result

In knowing the comparison of algorithm performance, the test scenarios are carried out alternately. The traditional k-means test was carried out with clusters ranging from 3 to 6. Meanwhile, the Dynamic K-means algorithm with Binary Search Centroid uses the results of the ideal cluster number based on intra and inter-cluster calculations of each dynamic cluster algorithm.

The assessment of the final evaluation results is determined by the acquisition of experiments that have been carried out previously, then comparing 3 to 6 the number of clusters formed. Figure 3 and Figure 4 are graphs evaluating the DBI and CHI values from testing the two algorithms.



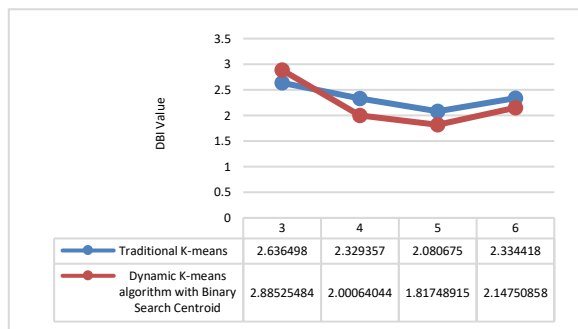| | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| Traditional K-means | 2.636498 | 2.329357 | 2.080675 | 2.334418 |
| Dynamic K-means algorithm with Binary Search Centroid | 2.88525484 | 2.00064044 | 1.81748915 | 2.14750858 |

Figure 3. Comparison of the average DBI values of the two algorithms

Overall the results of the DBI evaluation obtained the best value on the dynamic k-means algorithm with a DBI value of 1.81748915. The addition of the number of clusters affects the results of the DBI evaluation, where the decrease in the value of DBI on the dynamic cluster algorithm is better. It is known that the optimal cluster results with the smallest DBI value or close to the value (non-negative >= 0) indicate the cluster is getting better (Siagian et al., 2022).



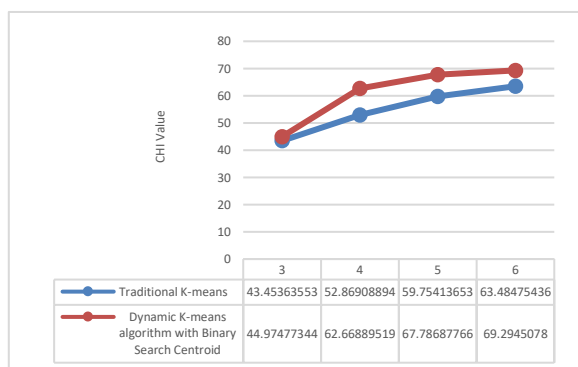| | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| Traditional K-means | 43.45363553 | 52.86908894 | 59.75413653 | 63.48475436 |
| Dynamic K-means algorithm with Binary Search Centroid | 44.97477344 | 62.66889519 | 67.78687766 | 69.2945078 |

Figure 4. Comparison of the average Calinski-Harabasz Index (CHI) values of the two algorithms

In the evaluation test using the Calinski-Harabasz Index (CHI), the Dynamic K-means algorithm with Binary Search Centroid performs better than the traditional K-means algorithm. This can be seen from the acquisition of the CHI value with the best number of clusters indicated by the more excellent CHI value (Harani et al., 2020). Based on Figure 4, the Calinski-Harabasz Index (CHI) evaluation value, the traditional k-means algorithm, is unstable for each number of clusters formed.

Based on the evaluation of DBI and CHI, it is clear that the change in the value of each algorithm depends on the initiation of the centroid in determining the initial cluster used. The Dynamic K-means algorithm with Binary Search Centroid has the advantage of constant iterations in each test. So that every time the data is grouped, the Dynamic K-means method with Binary Search Centroid always produces the same cluster. Initialization of the initial cluster center point value randomly causes the traditional k-means algorithm to obtain different validation values and the number of clusters in each test, making it very difficult to obtain individual initial cluster results. The test results do not necessarily represent a good cluster.

As a result, the obtained centroid is not a centroid that is dominated by the grouping of health indicators that have more similarities. The iteration process after the initial cluster generation will produce an inaccurate centroid. Determining initial cluster initiation dramatically affects the number of iterations of the clustering algorithm, and the estimated time is not constant.

Based on the test results, the Dynamic K-means algorithm with Binary Search Centroid produces better cluster quality than traditional K-means. To find out the best number of clusters in the dynamic kmeans algorithm with cluster initialization, testing is carried out on the level of validity of each number of clusters. Table 1 is a comparison of each evaluation in each number of clusters.

Table 1. Validation value of Dynamic K-means cluster k-means with Binary Search Centroid

| Number of Clusters | VW | VB | DBI | CHI |
|---|---|---|---|---|
| 3 | 3.7282 | 2.8909 | 2.8852 | 44.9747 |
| 4 | 3.6566 | 4.2032 | 2.0006 | 62.6688 |
| 5 | 3.4085 | 4.0360 | 1.8174 | 67.7868 |
| 6 | 3.7151 | 4.5564 | 2.1475 | 69.2945 |

Based on Table 1, the VW value of the number of 4 clusters is smaller than the number of clusters 3, while the VB value of the number of clusters 4 is greater than the VB value of the number of clusters

3. This causes the DK-KBSC algorithm to add one more cluster to cluster 5 according to the results of the cluster evaluation.

When viewed from the DBI and CHI values, the Dynamic K-means algorithm with Binary Search Centroid has several clusters 5, which is better than the number of clusters 3, 4, and 6.

**B.    Analysis**

Based on the results of clustering of Java Island health profile data in 2020 using the dynamic k-means algorithm with mean-based initial cluster initialization, the results obtained are 5 clusters in the distribution area, of which cluster 1 with very high health quality is 11 districts/cities, cluster 2 with High health quality is 28 regencies/cities, and cluster 3 with adequate health quality is 24 regencies or cities. Moreover, the health quality of 45 districts/cities for Cluster 4 with low health quality and cluster 5 with poor health quality in 11 districts or cities. The percentage distribution of clusters is shown in Figure 5.
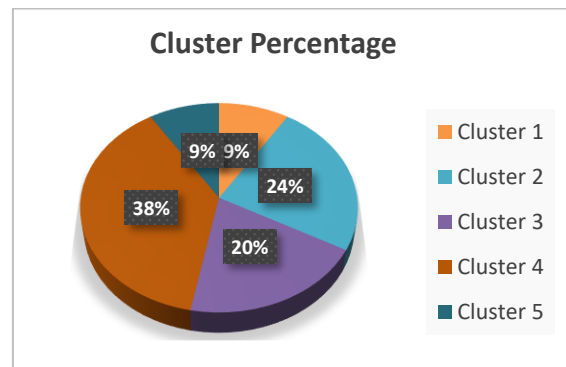


Figure 5. Percentage of Dynamic K-means Clustering Results with Binary Search Centroid

The result of the Dynamic K-means clustering area distribution with Binary Search Centroid from health profile data is shown in Table 2. The clustering results can then be used as a reference to determine the distribution of regional health levels and as evaluation material by the relevant agencies.

Table 2. Clustering Results

| Cluster | Regency/City | |
|---|---|---|
| Cluster 1 | Sukabumi City | North Jakarta City |
| | Cirebon City | West Jakarta City |
| | Tangerang City | South Jakarta City |
| | Cilegon City | East Jakarta City |
| | South Tangerang City | Seribu Islands Regency |
| | Central Jakarta City | |
| Cluster 2 | Pacitan Regency | Purwakarta Regency |
| | Bogor Regency | Karawang Regency |
| | Sukabumi Regency | Pangandaran Regency |
| | Cianjur Regency | Bogor City |
| | Bandung Regency | Bandung City |
| | Garut Regency | Bekasi City |
| | Tasikmalaya Regency | Depok City |
| | Ciamis Regency | Cimahi City |
| | Kuningan Regency | Tasikmalaya City |
| | Cirebon Regency | Lebak Regency |
| | Majalengka Regency | Pandeglang Regency |
| | Sumedang Regency | Serang Regency |
| | Indramayu Regency | Tangerang Regency |
| | Subang Regency | Serang City |
| Cluster 3 | Kediri Regency | Bangkalan Regency |
| | Lumajang Regency | Sampang Regency |
| | Jember Regency | Pamekasan Regency |
| | Banyuwangi Regency | Kediri City |
| | Bondowoso Regency | Malang city |
| | Situbondo Regency | Probolinggo City |
| | Probolinggo Regency | Pasuruan City |
| | Pasuruan Regency | Grobogan Regency |
| | Jombang Regency | Rembang Regency |
| | Nganjuk Regency | Pekalongan Regency |
| | Madiun Regency | Brebes Regency |
| | Tuban Regency | Banjar City |
| Cluster 4 | Ponorogo Regency | Boyolali Regency |

| Cluster | | Regency/City |
|---|---|---|
| | Trenggalek Regency | Klaten Regency |
| | Tulungagung Regency | Sukoharjo Regency |
| | Blitar Regency | Wonogiri Regency |
| | Malang Regency | Karanganyar Regency |
| | Sidoarjo Regency | Sragen Regency |
| | Mojokerto Regency | Blora Regency |
| | Magetan Regency | Pati Regency |
| | Ngawi Regency | Holy Regency |
| | Bojonegoro Regency | Jepara Regency |
| | Lamongan Regency | Demak Regency |
| | Gresik Regency | Semarang Regency |
| | Sumenep Regency | Temanggung Regency |
| | Surabaya City | Kendal Regency |
| | Batu City | Batang Regency |
| | Cilacap Regency | Pemalang Regency |
| | Banyumas Regency | Tegal Regency |
| | Purbalingga Regency | Surakarta City |
| | Banjarnegara Regency | Salatiga City |
| | Kebumen Regency | Semarang city |
| | Purworejo Regency | Pekalongan City |
| | Wonosobo Regency | Tegal City |
| | Magelang Regency | |
| Cluster 5 | Blitar City | Kulon Progo Regency |
| | Mojokerto City | Bantul Regency |
| | Madiun City | Gunung Kidul Regency |
| | Magelang City | Sleman Regency |
| | Kuningan Regency | Yogyakarta City |
| | Bekasi Regency | |

## CONCLUSIONS

The dynamic cluster algorithm on kmeans with binary search centroid initiation produces a better average cluster quality than the traditional kmeans algorithm in evaluating the Davies Bound Index (DBI) and Calinski-Harabasz Index (CHI). The algorithm's success can be seen from the resulting 5 clusters with the smallest VW value of 3.408591352 compared to the traditional k-means algorithm.

The advantage of the dynamic k-means algorithm with binary search centroid initiation is its constant number of iterations in each test. So that every time the data is grouped, the dynamic k-means method with binary search centroid initiation always produces the same cluster. Meanwhile, the traditional k-means algorithm obtains different validation values and the number of clusters in each test because the initialization of the initial cluster center point value is carried out randomly, so obtaining individual initial cluster results and a small number of iterations is challenging.

The number of clusters affects the DBI and CHI values. If the number of clusters increases, the CHI value will increase. Evaluation of the validity of the Calinski-Harabasz Index (CHI) the more excellent CHI value indicates the best value. The more the number of clusters and the data used, the better the quality of the cluster or grouping. Adding the number of clusters affects the results of the DBI evaluation in a better direction. Although other algorithms have improved clustering results, the decrease in DBI in the traditional k-means algorithm has not been significant.

## REFERENCES

A. F. Khairati, A. A. Adlina, G. F. Hertono, and B. D. Handari, "Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 2, pp. 161–170, 2019.

Akbari Gumilar and Yusrila Kerlooza. "Peningkatan Hasil Cluster Menggunakan Algoritma Dynamic K-means dan K-means Binary Search Centroid." *Jurnal Tata Kelola dan Kerangka Kerja Teknologi Informasi,* vol. 4, no. 2, pp. 25-33, 2018.

Badan Pusat Statistik Pulau Jawa, "Statistik Kesehatan Pulau Jawa 2020", Available: https://jatim.bps.go.id/publication/2021/08/05/a70cbc1ca224552d5e0f5000/statistik-kesehatan-provinsi-jawa-timur-2020.html. [Online]. [Accessed: 05-August-2022].

C. Kamila, M. Adiyatma, G. R. Namang, and R. R. F.

Syah, "Systematic Literature Review: Penggunaan Algoritma K-Means untuk Clustering di Indonesia dalam Bidang Pendidikan," *Informatika dan Teknologi (Intech)*, vol. 2, no. 1, pp. 19–24, 2021.

C. Satria and A. Anggrawan, "Aplikasi K-Means Berbasis Web untuk Klasifikasi Kelas Unggulan," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 1, pp. 111 – 124, nov 2021.

C. Sreedhar, N. Kasiviswanath, and P. Chenna Reddy, "Clustering Large Datasets Using K-Means Modified Inter and Intra Clustering (KM-I2C) in Hadoop," *Journal of Big Data*, vol. 4, no. 27, pp. 1 – 19, 2017.

Dinas Kesehatan Pulau Jawa, "Profil Kesehatan Pulau Jawa 2020", Available: https://dinkes.jatimprov.go.id/userfile/doku men/PROFIL%20KESEHATAN%202020.pdf. [Online]. [Accessed: 05-August-2022].

H. Santoso, H. Magdalena, H. Wardhana, and I. Artikel, "Aplikasi Dynamic Cluster pada K-Means Berbasis Web untuk Klasifikasi Data Industri Rumahan," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 541–554, 2022.

Harani, Nisa Hanum, Cahyo Prianto, and Fikri Aldi Nugraha. "Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python." *Jurnal Manajemen Informatika (JAMIKA),* Vol. 10 No. 2, pp. 133-146, 2020.

Islami, Bagus Muhammad, Cepy Sukmayadi, and Tesa Nur Padilah. "Clustering of Health Facilities Based on Districts in Karawang With the K-Means Algorithm." *BINA INSANI ICT JOURNAL*, Vol. 8, No. 1, pp. 83-92, 2021.

K. Ariasa, I. G. A. Gunadi, and I. M. Candiasa, "Optimasi Algoritma Klaster Dinamis pada K-Means dalam Pengelompokkan Kinerja Akademik Mahasiswa (Studi Kasus: Universitas Pendidikan Ganesha)," *Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI*, vol. 9, no. 2, pp. 181–193, 2020.

L. Suryani, "Evaluasi Sistem Informasi Kesehatan dengan Pendekatan Health Metrics Network (HMN) di Dinas Kesehatan Kota Pagar Alam Tahun 2021," *J. Kesehat. Saelmakers PERDANA*, vol. 5, pp. 97–103, 2022.

Luthfi, Emir, and Arie Wahyu Wijayanto. "Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia." *INOVASI,* Vol. 17, No. 4, pp. 761-773, 2021.

R. M. Alguliyev, R. M. Aliguliyev, and F. J. Abdullayeva, "PSO+K-Means Algorithm for Anomaly Detection in Big Data,"*Statistics, Optimization and Information Computing*, vol. 7, no. 2, pp. 348–359, 2019.

Siagian, Romadansyah, Pahala Sirait, and Arwin Halim. "The Implementation of K-Means dan K-Medoids Algorithm for Customer Segmentation on E-commerce Data Transactions." *Sistemasi: Jurnal Sistem Informasi*, Vol. 11 No.2, pp. 260-270, 2022.

Supriyatna, Adi, et al. Rice productivity analysis by province using K-means cluster algorithm. In: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, Vol. 771, No. 1, pp. 012025, 2020.

W. Widiarina and R. S. Wahono, "Algoritma Cluster Dinamik untuk Optimasi Cluster pada Algoritma K-Means dalam Pemetaan Nasabah Potensial," *Journal of Intelligent Systems*, vol. 1, no. 1, pp. 33–36, 2015.

Y. A. Wijaya, D. A. Kurniady, E. Setyanto, W. S. Tarihoran, D. Rusmana, and R. Rahim, "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities," *TEM Journal*, vol. 10, no. 3, pp. 1099–1103, 2021.