

Analyzing schools admission performance achievement using hierarchical clustering

Tora Fahrudin¹, Ibnu Asror², Yanuar Firdaus Arie Wibowo²

¹School of Applied Sciences, Telkom University, Bandung, Indonesia

²School of Computing, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received Mar 2, 2024

Revised Jul 3, 2024

Accepted Jul 7, 2024

Keywords:

Admission data

Calinski-Harabasz

Davies-Bouldin index

Dunn index

Hierarchical clustering

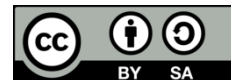
Membership movement

Silhouette score

ABSTRACT

In this study, an implementation of hierarchical clustering methods was conducted in schools' admission data. We aim to demonstrate that the hierarchical clustering method can be used to help analyze the membership changes of each cluster based on its achievement number of new students from different months period observations. This method can be used by decision-makers to make a strategy for each school which has decreasing achievement from the previous period. In this paper, we employ the hierarchical clustering method to cluster admission performance achievement from fifty Telkom Schools. Instead of clustering admission in one period directly, this paper tried to analyze the movement of clustering membership from one period to another. We observed the movement membership of the group from three categories period, such as monthly, quarterly, and semesterly. The experimental results demonstrate that the monthly scenario was the best clustering result. The monthly scenario achieves the best score for all metrics such as the Dunn index, Silhouette score, Davies-Bouldin index, and Calinski-Harabasz compared to the quarter and semester scenario. There are four schools which are consistent in the first cluster and seven schools which are consistent in the second cluster in all scenarios and all periods.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tora Fahrudin

School of Applied Sciences, Telkom University

Telekomunikasi street No. 1, Terusan Buah Batu, Bandung 40257, Indonesia

Email: torafahrudin@telkomuniversity.ac.id

1. INTRODUCTION

Nowadays, clustering has become one of the data mining methods that has been implemented across various fields. For instance, clustering techniques were applied to medical data set to build a smart healthcare system [1], financial data set to cluster stock prices [2] and to cluster churned insures and retained insures [3], educational data set to cluster student enrollment performance [4], marketing data set to segment the electric vehicles buyers [5]. Similarly, in the realm of social media, clustering has been used to group users based on their engagement patterns and behaviors [6], enabling platforms to personalize user experiences and enhance content delivery. Furthermore, other fields such as in environmental data, cluster was conducted to group the environmental and energetic impacts of Atlantic recipes [7]. Finally, cluster also used to categories technical performance and match displacement in rugby league [8].

Traditional clustering algorithms can be divided into two main groups: hierarchical and partitional. The hierarchical clustering algorithm is one of the clustering techniques which creates clusters by making the hierarchical relationship between data [9]. Unlike hierarchical clustering, where clusters are arranged in a hierarchy, partitional clustering directly assigns each data point into k clusters by optimizing some clustering

criteria [10]. Compared to partition-based clustering, hierarchical clustering has some advantages. Multilevel structure [11], no initial number of clusters required [12], better interpretability [13], decision-making flexibility [14], support for overlapping groupings [15], and resistant to noise [16].

Evaluating clustering algorithms is one of the critical aspects of clustering stages. One common approach to evaluation involves the use of internal validation metrics, which assess the quality of clustering results based on intrinsic characteristics of the data, such as cohesion and separation of clusters. Examples of these metrics include the silhouette score [17], the Davies-Bouldin index [18], the Dunn index [19], and the Calinski-Harabasz index [20].

The implementation of clustering in education is still an interesting topic for researchers. Some examples are grouping student's learning activities [21], profiling and grouping students based on their academic performance [22], grouping topic discussions thread [23], and higher education application preference [24]. To the best of our knowledge, there has been limited research focused on the application of clustering methods to categorize the performance achievements of schools. This gap in the literature underscores the need for further investigation to assess the effectiveness and potential benefits of implementing clustering techniques in analyzing school performance.

Hence, this paper proposes analyzing the movement of hierarchical cluster membership from each school in different periods of admission performance achievement. We monitored the group's membership movement throughout three different periods namely monthly, quarterly, and semesterly. By differentiate the periods, we can find which is the best period for grouping the admission performance achievement and who school which change from one cluster to the other cluster, so based on these results, the foundation can take appropriate action for schools which have a declining performance or schools that are consistently in a cluster with poor performance.

This paper is written in the following structure. A detailed methodology including data set description, and methodology explanation, is discussed in section 2. Next, section 3 discusses the results and discussion. Lastly, the conclusion is finally provided in section 4.

2. RESEARCH METHOD

2.1. Data

The data for this paper is collected from the Telkom Foundation Admission Database of fifty-one Telkom Schools, which are observed in one year of admission period which started from October 2022 until September 2023. Collecting data was performed by the system weekly and aggregated into monthly data. There are three attributes for each month, namely registered percentage achievement, accepted percentage achievement, and paid percentage achievement. Registered percentage achievement is what percentage of a new student registered from the monthly target that has been set. Accepted percentage achievement is what percentage of a new student is accepted from the accumulation target for each month. Paid percentage achievement is what percentage of a new student paid their bill. All these attributes are normalized into 0-1 scaled.

Table 1 shows a sample dataset and Table 2 shows statistical information for each attribute (in monthly). From the statistical information produced in Table 2, we can see that the average achievement of the percentage of applicants and the percentage of students accepted from all schools has reached more than 25% in the first quarter and has exceeded 100% at the end of the third quarter of the student admission period for the academic year 2023 to 2024. Furthermore, the percentage of those who made payments only reached 25% at the beginning of the second quarter, although performance increased at the end of the third quarter when the percentage of student payments had exceeded 100%.

Table 1. Sample dataset

No	School Name	M1_reg	M1_acc	M1_paid	...	M12_paid
1	Paud Bandung	0.3	0.3	0.3	...	1.15
2	Paud Makassar	0.26	0.14	0.14	...	1.65
...
51	SMK Sidoarjo	0.52	0.32	0.15	...	1.16

2.2. Methodology

Hierarchical clustering is one of the clustering methods which generates a nested sequence of partitions of objects. It proceeds successively by either merging smaller clusters into larger ones (Agglomerative) or by splitting larger clusters (Divisive). That procedure produces a dendrogram, or tree of clusters [25] which can be seen in Figure 1. A clustering of the data items into fragmented groups is produced

by cutting the dendrogram at a desired level. As we see, we can get two clusters which consist of 4-0-1 and 5-2-3, or four clusters which consist of 4, 0-1, 5, and 2-3.

Table 2. Attribute descriptive statistics

No	Attributes	Mean	Std	Min	25%	50%	75%	Max
1	m1_registered_percentage	0.221819	0.320713	0	0	0.06	0.3	1.353846
2	m1_accepted_percentage	0.170028	0.307055	0	0	0.04	0.216319	1.353846
3	m1_paid_percentage	0.105432	0.180168	0	0	0	0.152093	0.81
4	m2_registered_percentage	0.342455	0.364455	0	0	0.318182	0.519677	1.375
5	m2_accepted_percentage	0.277149	0.338936	0	0	0.16	0.403125	1.353846
6	m2_paid_percentage	0.217383	0.267642	0	0	0.16	0.334097	1.09
7	m3_registered_percentage	0.535799	0.367976	0	0.243571	0.544643	0.725385	1.428571
8	m3_accepted_percentage	0.4498	0.343164	0	0.156923	0.44	0.563711	1.401786
9	m3_paid_percentage	0.381123	0.292675	0	0.139231	0.386364	0.504444	1.133929
10	m4_registered_percentage	0.867431	0.385893	0.104167	0.630909	0.846154	1.100259	1.583333
11	m4_accepted_percentage	0.764208	0.363122	0.098958	0.412532	0.807692	1.034286	1.464286
12	m4_paid_percentage	0.70381	0.33818	0.096354	0.387143	0.733333	1.019138	1.313433
13	m5_registered_percentage	0.803448	0.395931	0	0.45	0.833333	1.049266	1.55
14	m5_accepted_percentage	0.704512	0.364926	0	0.39	0.76	1	1.464286
15	m5_paid_percentage	0.65428	0.341238	0	0.361429	0.7	0.927564	1.1875
16	m6_registered_percentage	0.803448	0.395931	0	0.45	0.833333	1.049266	1.55
17	m6_accepted_percentage	0.704512	0.364926	0	0.39	0.76	1	1.464286
18	m6_paid_percentage	0.65428	0.341238	0	0.361429	0.7	0.927564	1.1875
19	m7_registered_percentage	1.186109	0.387338	0.138021	1.014286	1.153846	1.332265	2.576923
20	m7_accepted_percentage	1.077663	0.31409	0.138021	0.866667	1.075	1.275556	1.807692
21	m7_paid_percentage	1.018222	0.305508	0.135417	0.839116	1.008547	1.216111	1.656716
22	m8_registered_percentage	1.026901	0.393396	0.109375	0.811313	1.05	1.228651	2.153846
23	m8_accepted_percentage	0.901682	0.351889	0.109375	0.62	1	1.1637	1.522388
24	m8_paid_percentage	0.84666	0.33496	0.106771	0.600633	0.866667	1.076923	1.522388
25	m9_registered_percentage	1.194064	0.384032	0.138021	1.026786	1.188235	1.332265	2.576923
26	m9_accepted_percentage	1.08235	0.311496	0.138021	0.908333	1.1	1.275556	1.807692
27	m9_paid_percentage	1.027812	0.294297	0.135417	0.847796	1.008547	1.216111	1.656716
28	m10_registered_percentage	1.194064	0.384032	0.138021	1.026786	1.188235	1.332265	2.576923
29	m10_accepted_percentage	1.08235	0.311496	0.138021	0.908333	1.1	1.275556	1.807692
30	m10_paid_percentage	1.027812	0.294297	0.135417	0.847796	1.008547	1.216111	1.656716
31	m11_registered_percentage	1.169931	0.412833	0.100427	1	1.153846	1.326667	2.576923
32	m11_accepted_percentage	1.057662	0.345999	0	0.87	1.075	1.275556	1.807692
33	m11_paid_percentage	1.002911	0.32791	0	0.832993	1.007576	1.216111	1.656716
34	m12_registered_percentage	1.169931	0.412833	0.100427	1	1.153846	1.326667	2.576923
35	m12_accepted_percentage	1.057662	0.345999	0	0.87	1.075	1.275556	1.807692
36	m12_paid_percentage	1.002911	0.32791	0	0.832993	1.007576	1.216111	1.656716

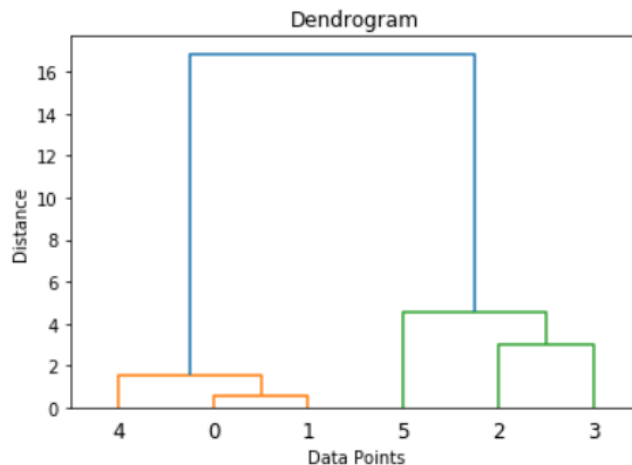


Figure 1. Example of dendrogram

Compared to divisive, agglomerative methods are far more prevalent [26]. First, agglomerative is often considered more intuitive and easier to understand [27]. Second, agglomerative clustering does not require the prior specification of the number of clusters [28], it allows users to visually inspect and choose

the number of clusters that best fits their needs from the dendrogram. Third, agglomerative clustering provides flexibility in choosing different linkage criteria (e.g., complete linkage, average linkage, Ward's method), allowing users to choose what suitable linkage method to link them until all clusters are grouped into one single cluster [29]. Finally, for small to medium-sized datasets, agglomerative clustering can be computationally efficient, this is because that algorithm has $O(n^2)$ time complexity using reciprocal nearest neighbors and reproducibility [30].

Algorithm 1. Agglomerative clustering (D, linkage)

Input:

D: distance matrix of size $n \times n$

linkage(C_1, C_2): a distance function between cluster

Output:

L (hierarchical clusters)

Process:

```

1 while |L|>1 do
2   find a pair of clusters ( $C_1, C_2$ ) in L with the smallest distance
3   merge  $C_1$  and  $C_2$  into a new cluster C
4   remove  $C_1$  and  $C_2$  from L
5   for each cluster  $C' \in L$  do
6      $d \leftarrow \text{linkage}(C, C')$ 
7     update the matrix D, set the distance between C and  $C'$  to d
8   remove the distance data which related to  $C_1$  and  $C_2$  from D
9   add C to L
return L

```

2.2.1. Experimental setting

The experimental design can be simply shown in Figure 2. There are five steps which start from collecting the dataset, deriving monthly, quarterly, and semesterly as period type of dataset, employing an agglomerative clustering algorithm, visualizing the cluster result, and lastly observing membership movement for each period type of dataset. The explanation details for each step are as follows:

- Step 1. The dataset was collected from the Telkom Foundation Admission Database. In this step, data was taken from the database and converted to comma separated value (CSV) format on a local computer. Data was taken in monthly format and transformed into tabular format as shown in Table 1.
- Step 2. Construct three types of period datasets, namely monthly, quarterly, and semesterly for the clustering process. In monthly, the clustering process uses all monthly data in one one-year period. In quarterly, the clustering process is taken in a four-quarter period separately. Finally for semesterly, the clustering process is taken in semester period separately. To easily describe the statistical information for each cluster, we derived one attribute called *total_n*, where *n* is one of these options: month/quarter/semester. For example, in the monthly scenario, to describe the cluster in the first month, we derived one column, namely *total_{m1}*. This column calculates total value for the first month of admission data from *m1_registered_percentage+m1_accepted_percentage+m1_paid_percentage*. For quarter scenario, we derived one column namely *total_{q1}* which calculates from *m1_registered_percentage+m1_accepted_percentage+m1_paid_percentage+m2_registered_percentage+m2_accepted_percentage+m2_paid_percentage+m3_registered_percentage+m3_accepted_percentage+m3_paid_percentage*. Lastly for semester scenario, *total_{s1}* was derived from *m1_registered_percentage+m1_accepted_percentage+m1_paid_percentage+m2_registered_percentage+m2_accepted_percentage+m2_paid_percentage+m3_registered_percentage+m3_accepted_percentage+m3_paid_percentage+m4_registered_percentage+m4_accepted_percentage+m4_paid_percentage+m5_registered_percentage+m5_accepted_percentage+m5_paid_percentage+m6_registered_percentage+m6_accepted_percentage+m6_paid_percentage*.

$$Total_n = \sum_{i=1}^n x_i \quad (1)$$

- Step 3. The clustering process using agglomerative clustering was conducted in this step. Python programming software, version 3.7.4, was used for cluster analysis using *sklearn.cluster.AgglomerativeClustering* package. We use default package settings for the algorithm, such as *n_clusters = 2*, *affinity&metric = 'Euclidean'*, *memory = none*, *linkage = 'ward'*, *distance_threshold = none*, and *compute_distances = false*.

- Step 4. The visualization process was done by using *matplotlib.pyplot* library and *scipy.cluster.hierarchy* respectively. Matrix linkage was built by using the linkage function and the dendrogram result was built by using the dendrogram function. These tools enabled the clear and comprehensive visualization

of hierarchical clustering results, making it easier to interpret the relationships and groupings within the data.

Step 5. The last step is the most interesting part. Observing the membership movement from one group to another group in different periods can be interesting for stakeholders to see what is going on in that phenomenon. This dynamic analysis provides valuable insights into how and who certain entities shift between clusters over time, revealing underlying trends and patterns that static analysis might miss.

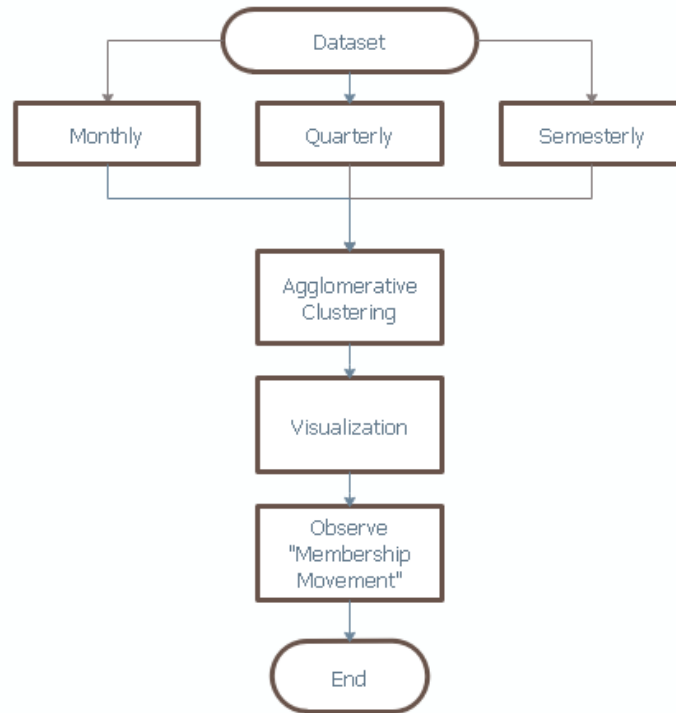


Figure 2. Experimental design

2.2.2. Evaluation metric

For the evaluation and comparison of the clustering result from three scenarios, there are three approaches to evaluate the quality of the cluster [31] such as external, internal, and relative criteria. The main difference among them is whether or not external information is used for clustering validation. External criteria use external information to measure accuracy based on given class labels. Internal criteria measure the goodness of cluster based on compactness and separation criteria [32]. Lastly, the relative criteria perform the evaluation of a clustering by comparing it to other clustering schemes. The most well-known cluster validation uses internal criteria, which are independent of external data.

In internal criteria, compactness measures how closely or how compactly related the objects in a cluster. A lower value indicates that objects in a cluster tend to be closer to each other. Within-cluster sum of squares and within-cluster variance are two samples of compactness metrics. The other name for compactness is cluster cohesion or homogeneity [33]. The objective of this measurement is to find a cluster which has a lower value of within-cluster variance which represents good compactness, and, hence, a good clustering. While separation measures how well clusters are separated from other clusters. This separation value is usually by using the minimum distance between cluster centroids or the pairwise minimum distance between objects in different clusters. So, by these metrics, the objective of this measurement is to find a cluster with maximum separation value. To tackle the trade-off between compactness and separation, some methods combine those two measures into a single score such as Silhouette coefficient (Sil), Dunn index (Dunn), Davies–Bouldin index (DBI), and Calinski–Harabasz index (CHI) [34]. Dunn index aims to show how dense and how the cluster is well-separated. Dunn index formula is expressed as:

$$Dunn = \frac{\min_{inter_cluster_distance}}{\max_{intra_cluster_distance}} \quad (2)$$

where $min_inter_cluster_distance$ is the minimum distance value (closest points) between two clusters i and j until k cluster, and $i \neq j$. While $max_intra_cluster_distance$ is the maximum distance (largest distance) between two points within the cluster k . For well-separable clusters, the inter-cluster distance is large, and their intra-cluster distance is small. Silhouette estimates how well each element in one cluster is close to another neighboring element in the same cluster compared to different clusters. Silhouette coefficient can be expressed as:

$$Sil = \frac{b_i - a_i}{\max\{b_i - a_i\}} \quad (3)$$

where b_i is the minimum average distance value between the i^{th} point and all samples, a^i is the average distance between the point and all samples. The Davies Bouldin index is defined as the ratio of the sum of the within-cluster to between-cluster distances. So, by using this index, the best cluster (lower value) is achieved when clusters are farther apart and less dispersed. This Davies Bouldin index can be formulated as:

$$DB = \frac{1}{k} \sum_{i=1}^k \max \left(\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right) \quad (4)$$

where k is the number of clusters, i and j are cluster labels, $\Delta(X_i) + \Delta(X_j)$ are all samples in clusters i and j to their respective cluster centroids, and $\delta(X_i, X_j)$ is centroid distance. Calinski-Harabasz index is a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). A greater value of CH means better clustering results. A high index value indicates that the between-cluster variance is greater than the within-cluster variance, indicating that the clusters are well separated.

$$CH = \frac{\sum_{i=1}^k n_i \cdot d(z_i, z_{tot})^2}{k-1} \frac{n-k}{\sum_{i=1}^k \sum_{x \in c_i} d(x, z_i)^2} \quad (5)$$

where n and k are the number of points and the number of clusters respectively, while z_i is the center of the cluster c_i and n_i is the number of points in c_i , x is a data point belonging to the cluster c_i .

3. RESULTS AND DISCUSSION

In this section, we present a comprehensive performance evaluation of the proposed method, validation results, and discussion about the implementation details, experimental result analysis, and limitations of the study. The results were explained in three subsections namely monthly, quarterly, semesterly. Experimental results were showed in the chart and table. In the discussion section, we delved into the implementation details, essential interpretation based on key findings, comparison and contrast with previous studies, and also highlighted the limitations.

3.1. Monthly

In the monthly scenario, the clustering was performed monthly, so there are 12 clustering processes, from 1st month, 2nd month, until 12th month. For each month, we divided into two groups of clusters, namely 1st cluster as a good performance cluster, and 2nd cluster as a fair performance cluster. Figures 3 to 14 show the dendrogram result of monthly scenarios for each month independently. Table 3 shows the clustering results for each month with additional information, such as cluster descriptions and membership of each cluster.

From the monthly results, we can see several points. First, in the first six months (January to June), the average difference between the first cluster and the second cluster was higher than the average difference between the first cluster and the second cluster in the last six months (July to December). This means that in the first six months, the performance achievements of members of the first cluster far exceeded the performance of the second cluster. Meanwhile, in the second last month, the performance achievements of members of the second cluster were getting closer to the performance of members of the first cluster. Second, the achievement of each school gets better after the first three months, which can be seen from the number of members increase from less than 10 to more than 30 in the sixth month. Another phenomenon shows that in the second and fourth quarters, the achievements of each school were at maximum performance, this can be seen from the highest number of cluster members' achievements compared to the first and third quarters. Third, there are 4 members which are consistent as the first member namely 2, 5, 37, and 44.

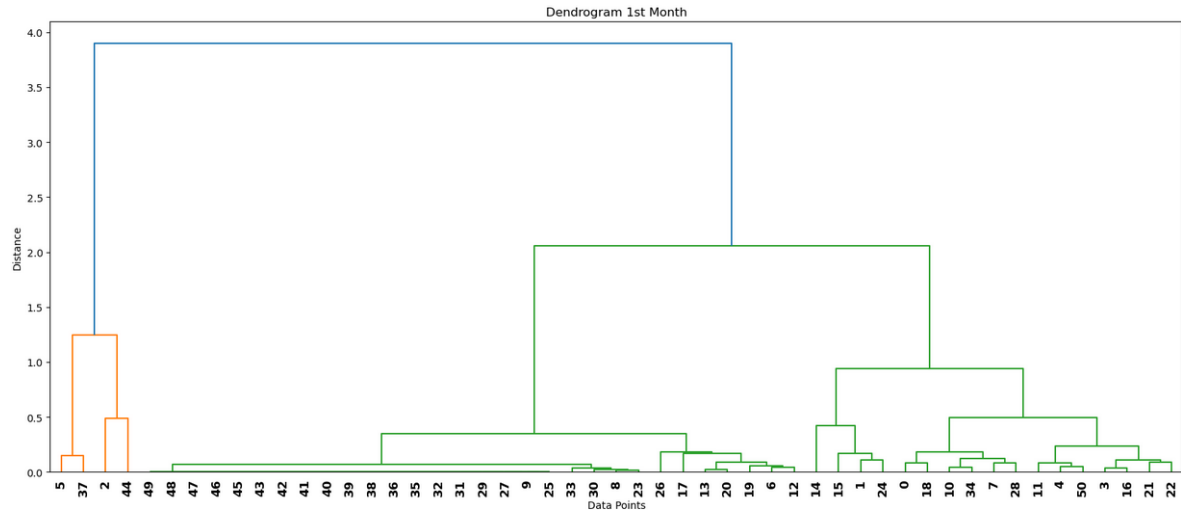


Figure 3. First month dendrogram

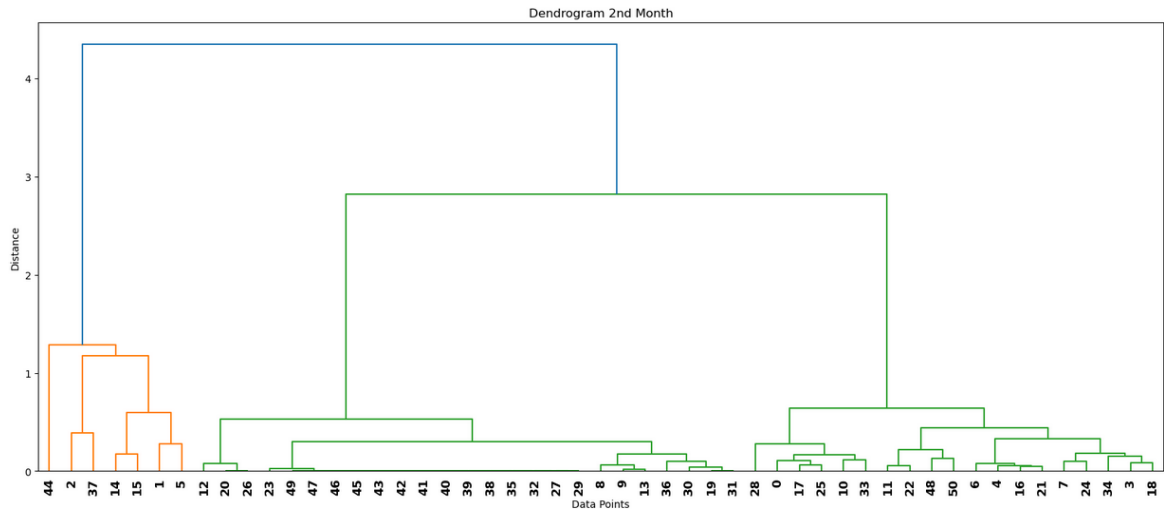


Figure 4. Second month dendrogram

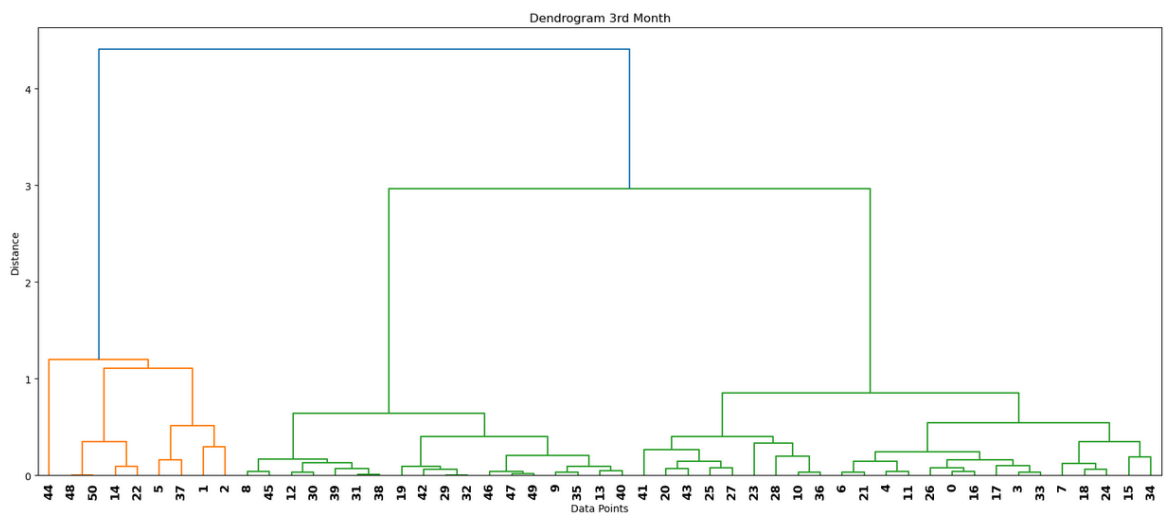


Figure 5. Third month dendrogram

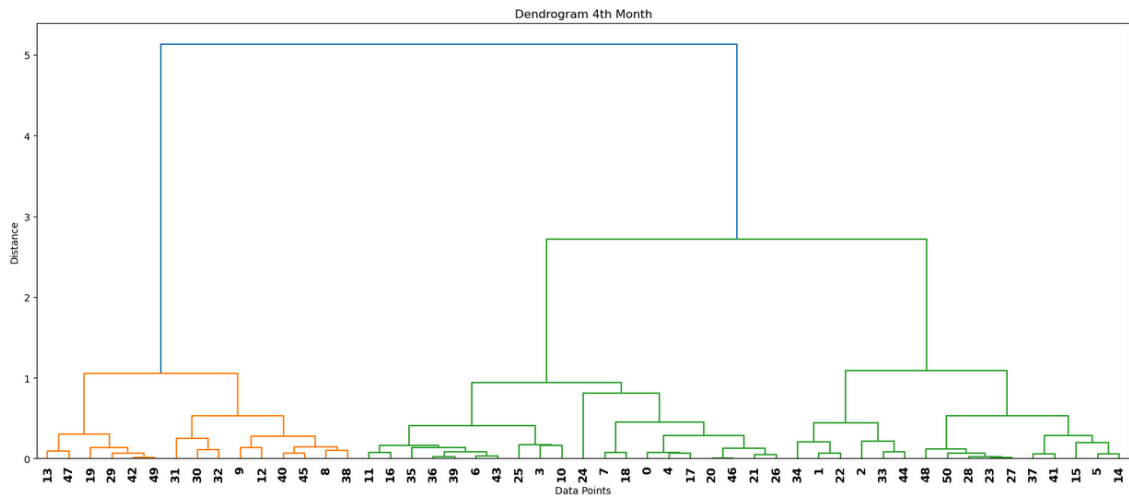


Figure 6. Fourth month dendrogram

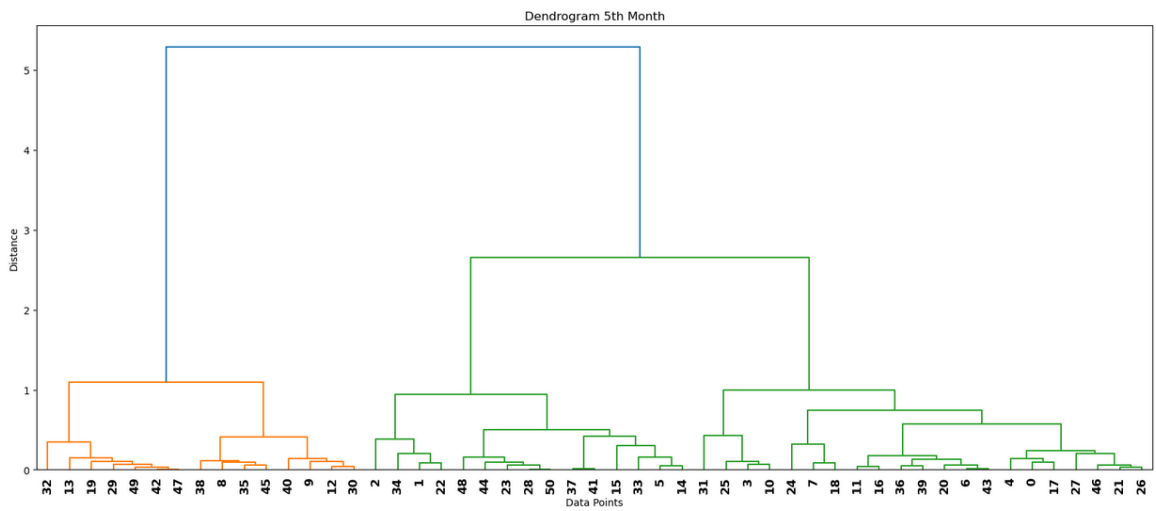


Figure 7. Fifth month dendrogram

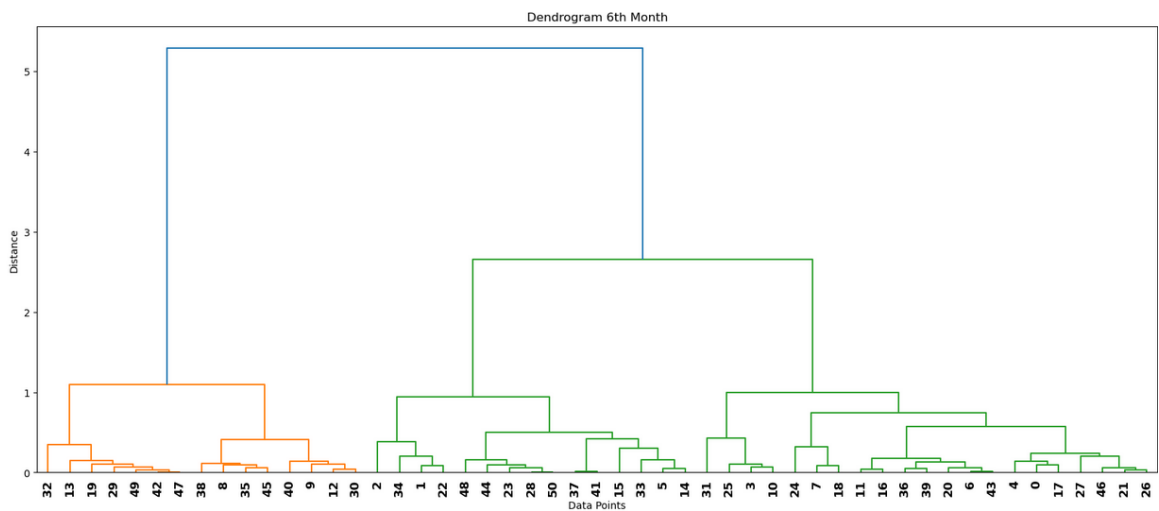


Figure 8. Sixth month dendrogram

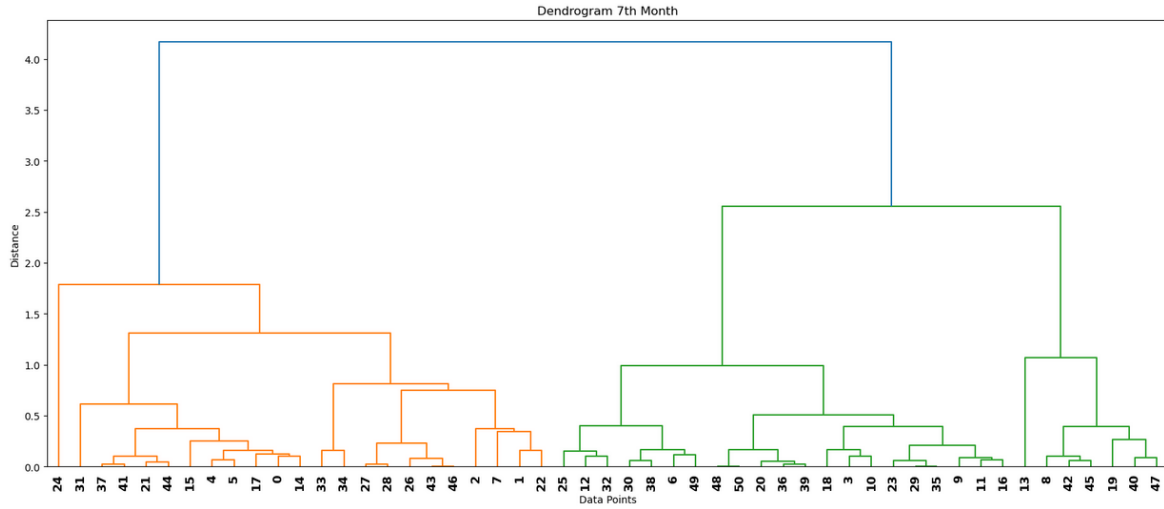


Figure 9. Seventh month dendrogram

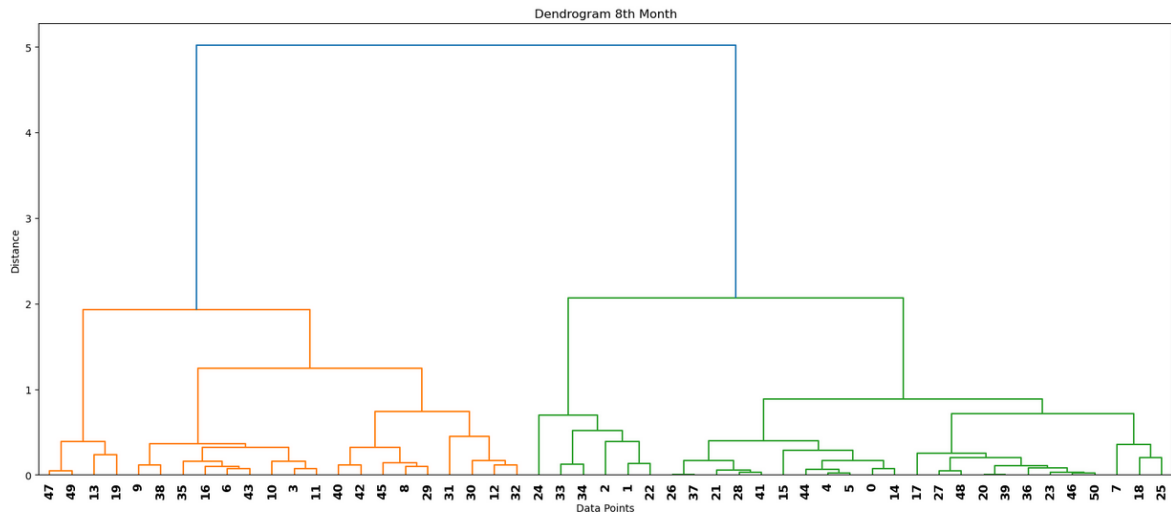


Figure 10. Eighth month dendrogram

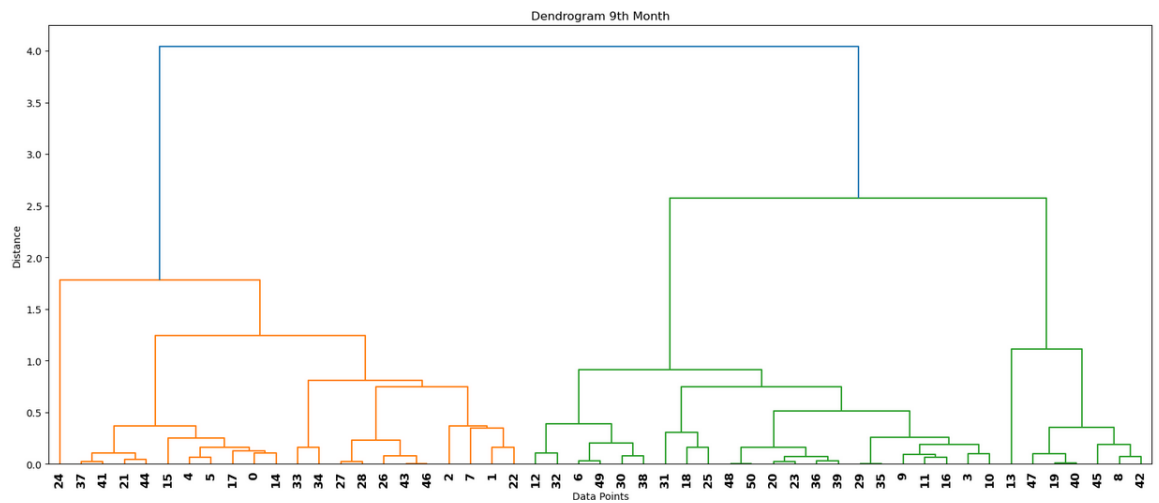


Figure 11. Ninth month dendrogram

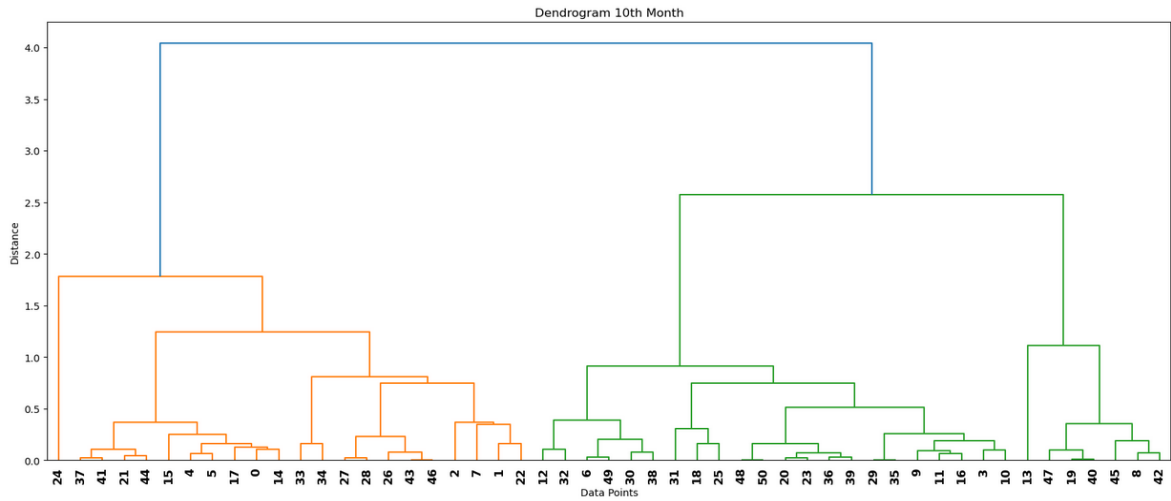


Figure 12. Tenth month dendrogram

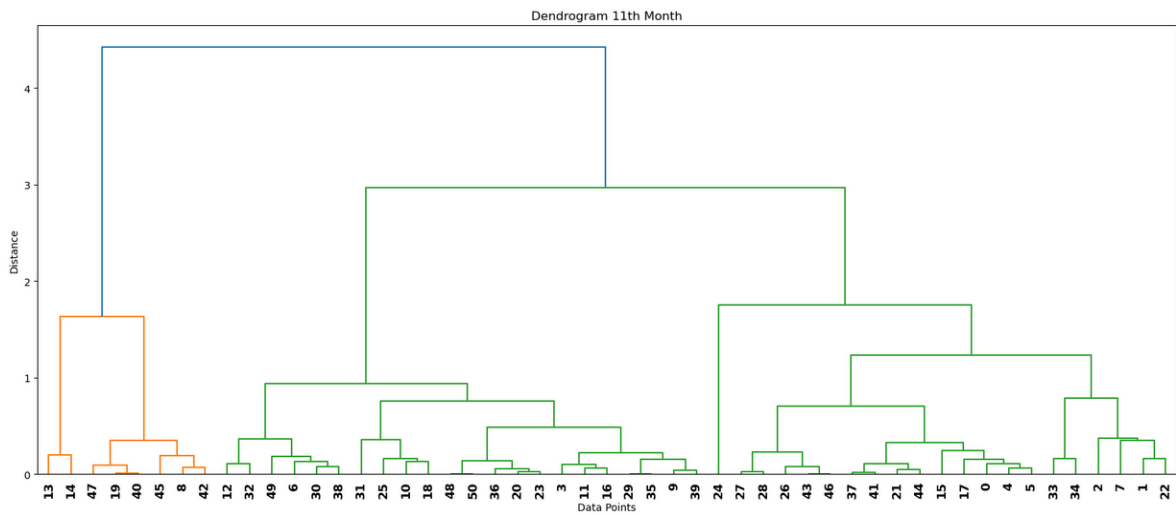


Figure 13. Eleventh month dendrogram

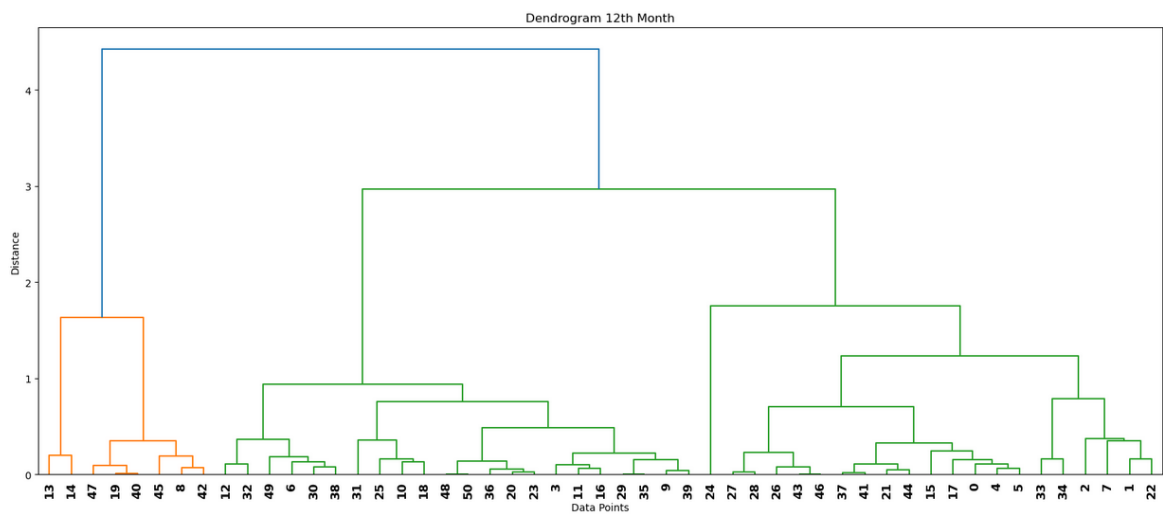


Figure 14. Twelfth month dendrogram

Table 3. Monthly result

No	Month	1 st Cluster descriptive	1 st Cluster member	2 nd Cluster descriptive	2 nd Cluster member
1	January	Count: 4	2,5,37,44	Count: 47	0,1,3,4,6,7,8,9,10,11, 12,13,14,15,16,17,18, 19,20,21,22,23,24,25,26, 27,28,29,30,31,32,33,34, 35,36,38,39,40,41,42,43,45, 46,47,48,49,50
		Mean: 0.49		Mean: 0.07	
2	February	Std: 0.35	1,2,5,14,15,37,44	Std:0.112	0,3,4,6,7,8,9,10,11,12,13,14, 16,17,18,19,20,21,22,23,24, 25,26,27,28,29,30,31,32,33, 34,35,36,38,39,40,41,42,43, 45,46,47,48,49,50
		Min: 0		Mean: 0.14	
3	March	Max: 1.35	1,2,5,14,22,37,44,48,50	Min:0	0,3,4,6,7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21,23,24, 25,26,27,28,29,30,31,32,33, 34,35,36,38,39,40,41,42,43, 45,46,47,49
		Count: 7		Mean: 0.284	
4	April	Mean: 0.67	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 25,26,27,28,33, 34,35,36,37,39,41,43, 44,46,48,50	Std:0.15	8,9,12,13,19,29,30,31,32, 38,40,42,45,47,49
		Min: 0		Min:0	
5	May	Max: 1.09	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 25,26,27,28,31,33, 34,36,37,39,41,43, 44,46,48,50	Max: 0.55	8,9,12,13,19,29,30,32,35, 38,40,42,45,47,49
		Count: 9		Mean: 0.76	
6	June	Mean: 0.83	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 25,26,27,28,31,33, 34,36,37,39,41,43, 44,46,48,50	Std:0.182	8,9,12,13,19,29,30,32,35, 38,40,42,45,47,49
		Min:0.13		Min:0	
7	July	Max: 1.13	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 25,26,27,28,33, 34,36,37,39,41,43, 44,46,48,50	Max: 0.583	3,6,8,9,10,11,12,13, 16,18,19,20,23,25, 29,30,32,35,36, 38,39,40,42,45,47,48,49,50
		Count: 36		Mean: 1.78	
8	August	Mean: 2.90	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 25,26,27,28,33, 34,36,37,39,41,43, 44,46,48,50	Mean: 0.97	3,6,8,9,10,11,12,13, 16,19,29,30,31,32,35, 38,40,42,43,45,47,49
		Std: 0.64		Min:0.32	
9	September	Min: 1.83	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 26,27,28,33, 34,37,41,43,44,46	Min:0.299	3,6,8,9,10,11,12,13, 16,18,19,20,23,25,29,30, 31,32,35,36, 38,39,40,42,45,47,48,49,50
		Max: 4.11		Mean: 2.70	
10	October	Count: 36	0,1,2,3,4,5,6,7,10,11,14, 15,16,17,18,20,21,22,23,24, 26,27,28,33, 34,37,41,43,44,46	Count: 15	3,6,8,9,10,11,12,13, 16,18,19,20,23,25,29,30, 31,32,35,36, 38,39,40,42,45,47,48,49,50
		Mean: 2.75		Min:0.41	
11	November	Std: 0.63	0,1,2,3,4,5,6,7,9,10,11,12, 15,16,17,18,20,21,22,23,24, 25,26,27,28,29,30,31,32,33, 34,35,36,37,38,39, 41,43,44,46,48,49,50	Std:0.40	8,13,14,19, 40,42,45,47
		Min: 1.46		Max: 3.4	
12	December	Max: 3.98	0,1,2,3,4,5,6,7,9,10,11,12, 15,16,17,18,20,21,22,23,24, 26,27,28,33, 34,37,41,43,44,46	Max: 1.32	8,13,14,19, 40,42,45,47
		Count: 36		Mean: 1.48	
1	January	Mean: 2.75	0,1,2,3,4,5,6,7,9,10,11,12, 15,16,17,18,20,21,22,23,24, 26,27,28,33, 34,37,41,43,44,46	Mean: 0.76	8,13,14,19, 40,42,45,47
		Std: 0.63		Min:0.1	
2	February	Min: 1.46	0,1,2,3,4,5,6,7,9,10,11,12, 15,16,17,18,20,21,22,23,24, 26,27,28,33, 34,37,41,43,44,46	Min:0	8,13,14,19, 40,42,45,47
		Max: 3.98		Max: 2.25	

3.2. Quarterly

In the quarterly scenario, the clustering was performed quarterly, so there are 4 clustering processes, from the 1st quarter, 2nd quarter, and until 4th quarter. Following the monthly scenario, for each quarter, we divided into two groups of clusters, namely 1st cluster as a good performance cluster, and 2nd cluster as a fair performance cluster. Figures 15 to 18 show the dendrogram result of quarterly scenarios for each quarter independently. Table 4 shows the clustering results for each quarter with additional information, such as cluster descriptions and membership of each cluster.

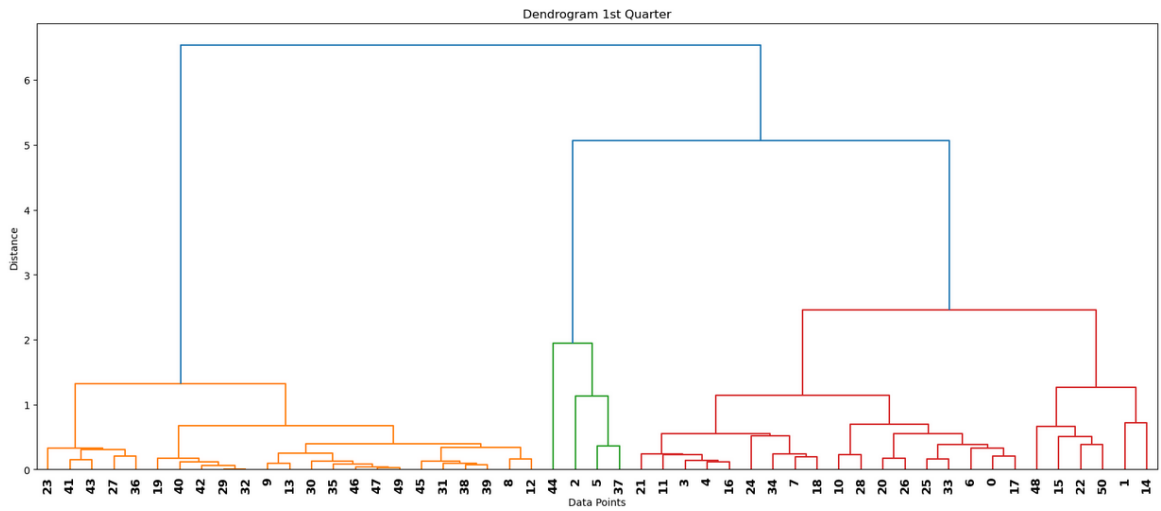


Figure 15. First quarter dendrogram

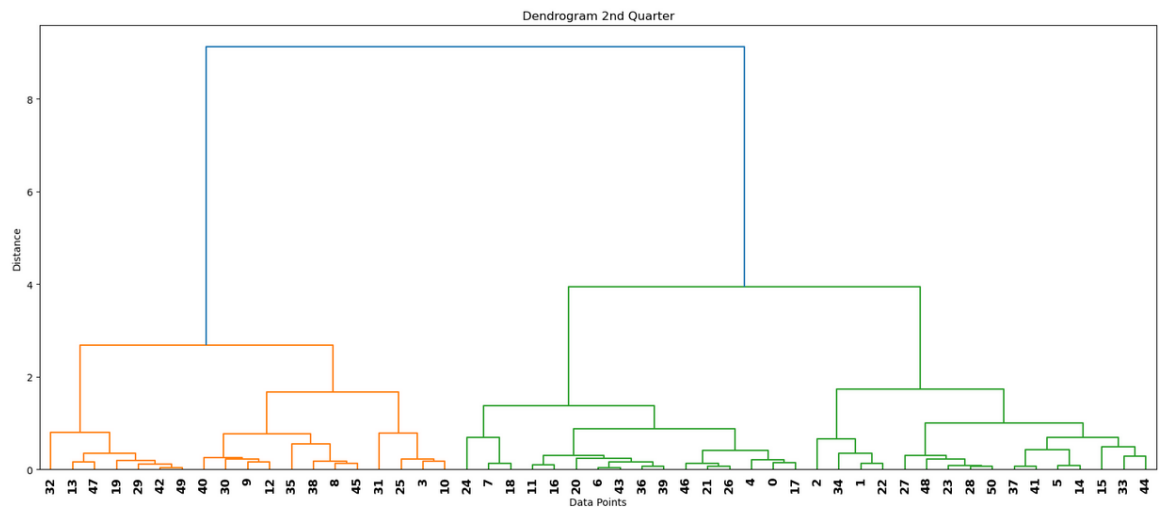


Figure 16. Second quarter dendrogram

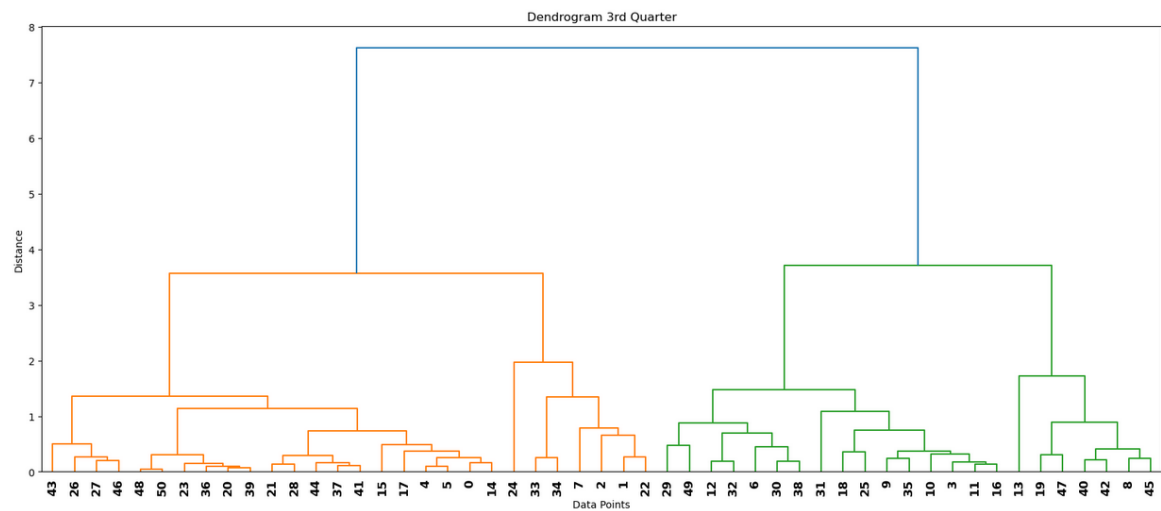


Figure 17. Third quarter dendrogram

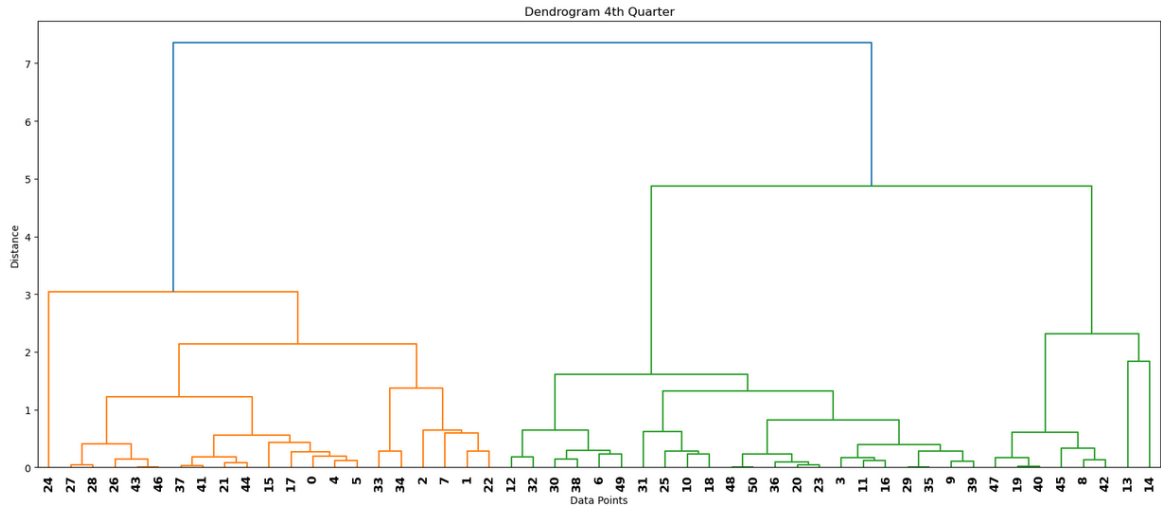


Figure 18. Fourth quarter dendrogram

Table 4. Quarterly result

No	Quarter	1 st Cluster Descriptive	1 st Cluster Member	2 nd Cluster Descriptive	2 nd Cluster Member
1	1 st Quarter (Jan-Mar)	Count: 28	0,1,2,3,4,5,6,7,10,11,	Count: 23	8,9,12,13,19,23,27,29,30,
		Mean: 4.39	14,15,16,17,18,	Mean: 0.64	31,32,35,36,38,39,40,41,
		Std: 2.31	20,21,22,24,25,26,	Std: 0.42	42,43,45,46,47,49
		Min: 1.98	28,33,34,	Min: 0	
2	2 nd Quarter (Apr-Jun)	Max: 10.70	37,44,48,50	Max: 1.42	
		Count: 32	0,1,2,4,5,6,7,11,14,	Count: 19	3,8,9,10,12,13,19,
		Mean: 8.74	15,16,17,18,20,21,22,23,24,	Mean: 3.14	25,29,30,31,32,35,
		Std: 1.70	26,27,28,33,	Std: 1.61	38,40,42,45,47,49
3	3 rd Quarter (Jul-Sep)	Min: 6.46	34,36,37,39,41,43,	Min: 0.88	
		Max: 11.94	44,46,48,50	Max: 6	
		Count: 28	0,1,2,4,5,7,14,	Count: 23	3,6,8,9,10,11,12,13,
		Mean: 11.4	15,17,20,21,22,23,24,	Mean: 6.87	16,19,25,29,30,31,32,
4	4 th Quarter (Oct-Dec)	Std: 1.68	26,27,28,33,	Std: 1.97	35,38,40,42,45,47,49
		Min: 9.47	34,36,37,39,41,43,	Min: 1.14	
		Max: 16.74	44,46,48,50	Max: 9.16	
		Count: 21	0,1,2,4,5,7,	Count: 30	3,6,8,9,10,11,12,13,14,
4	4 th Quarter (Oct-Dec)	Mean: 12.37	15,17,21,22,24,	Mean: 7.94	16,18,19,20,23,25,29,30,31,
		Std: 1.74	26,27,28,33,	Std: 2.19	32,35,36,38,39,40,42,
		Min: 10.36	34,37,41,43,	Min: 1.23	45,47,48,49,50
		Max: 17.8	44,46	Max: 10.38	

The cluster mean value exhibited no significant deviation compared to earlier cases. The mean values of both the first and second clusters exhibited a continuous increase from the first quarter to the fourth quarter. By examining the membership trends of each cluster from the first quarter to the fourth quarter, it becomes evident that there is a progressive shift in membership from the first cluster (positive cluster) to the second cluster (negative cluster). After the fourth quarter, the membership count of the second cluster exceeds that of the first cluster. This phenomenon suggests that the members who performed well in the first quarter have maintained their high level of performance until the end of the quarter. Although certain members of the cluster experienced an initial improvement in performance around the middle of the quarter, their performance subsequently declined towards the conclusion of the quarter. The first cluster consists of 4 members: 0, 1, 2, 4, 5, 7, 15, 17, 21, 22, 24, 26, 28, 33, 34, 37, and 44. There are currently 18 members who have moved from the 1st cluster to the 2nd cluster. These members are numbered 3, 6, 10, 11, 14, 16, 18, 20, 23, 25, 27, 36, 39, 41, 43, 46, 48, and 50.

3.3. Semesterly

In the scenario when the semesters were used, the clustering was carried out in semesters. Two clustering procedures take place in the first and second semesters. Similarly, to the previous situation, we classified each semester into two groups of clusters: the first cluster representing students with good

performance, and the other cluster representing students with poor performance. Figure 19 and Figure 20 display the dendrograms representing the results of semesterly scenarios for each month individually. Table 5 displays the clustering outcomes for each month, including further details like cluster descriptions and the composition of each cluster.

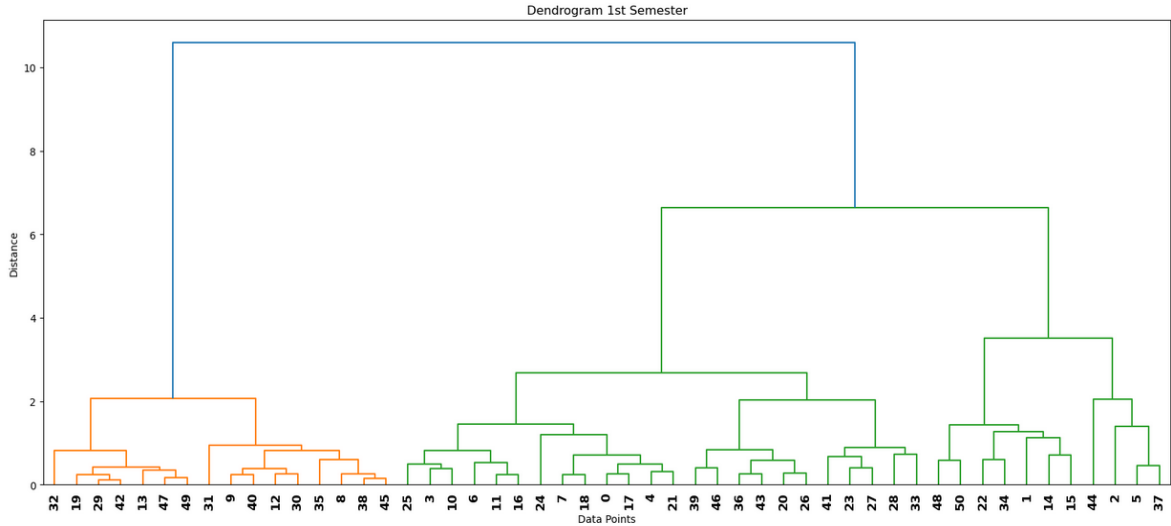


Figure 19. First semester dendrogram

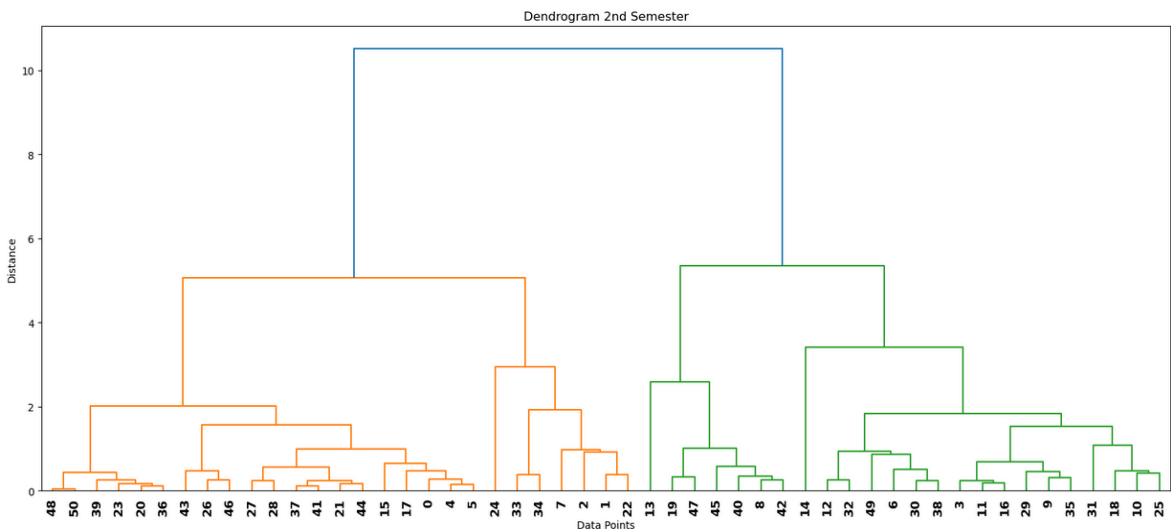


Figure 20. Second semester dendrogram

Table 5. Semesterly result

No	Month	1 st Cluster descriptive	1 st Cluster member	2 nd Cluster descriptive	2 nd Cluster member
1	1 st Semester (Jan-Jun)	Count: 35	0,1,2,3,4,5,6,7,10,11, 14,15,16,17,18, 20,21,22,23,24,25,26,27, 28,33,34,36,	Count: 16	8,9,12,13,19,29,30, 31,32,35,38,40, 42,45,47,49
		Mean: 12.21		Mean: 3.12	
	Std: 3.85		Std: 1.37		
	Min: 7.07		Min: 1.24		
2	2 nd Semester (Jul-Dec)	Max: 22.65	37,39,41,43,44,46,48,50	Max: 5.07	
		Count: 27	0,1,2,4,5,7, 15,17,20,21,22,23,24, 26,27,28,33,34,36,	Count: 24	3,6,8,9,10,11,12,13,14, 16,18,19,25,29,30, 31,32,35,38,40, 42,45,47,49
	Mean: 23.28		Mean: 3.14		
	Std: 3.5		Std: 1.61		
		Min: 18.9	37,39,41,43,44,46,48,50	Min: 0.88	
		Max: 34.54		Max: 6	

In the semesterly scenario, many schools were grouped into 1st cluster due to their performance in the first semester. But in the second semester, some schools move to 2nd cluster which causes the number of members in the cluster to be balanced between the first and second clusters (52.94% vs 47.06%). When observing the average value of each cluster, there is a very large difference between the first cluster and the second cluster. This shows that the performance achieved by members in cluster one far exceeds the achievements of members in cluster two. Schools which are consistently in 1st cluster are 0, 1, 2, 4, 5, 7, 15, 17, 20, 21, 22, 23, 24, 26, 27, 28, 33, 34, 36, 37, 39, 41, 43, 44, 46, 48, 50. Other schools which move from 1st cluster to 2nd cluster are 3, 6, 10, 11, 14, 16, 18, 25.

3.4. Metrics results

To evaluate the cluster result performances, we need to conduct metric evaluations such as Dunn index, Silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. Figure 21 shows the Dunn index and Silhouette score for clustering results in monthly scenarios. From this figure, we can see that according to the evaluation with the Silhouette index and Dunn index, the best cluster results are the cluster results in the first month. Meanwhile, the cluster results with the lowest Silhouette and Dunn index values were in the seventh, ninth and tenth months. This demonstrates that most schools did not meet their target within the first month. Only a small percentage, or about four institutions, performed well. In the evaluation of cluster results based on Calinski-Harabasz and Davies-Bouldin index (DBI) as shown in Figure 22, the best Calinski-Harabasz values were obtained in the fifth and sixth months and months while the Davies-Bouldin index (DBI) values were in the first month.

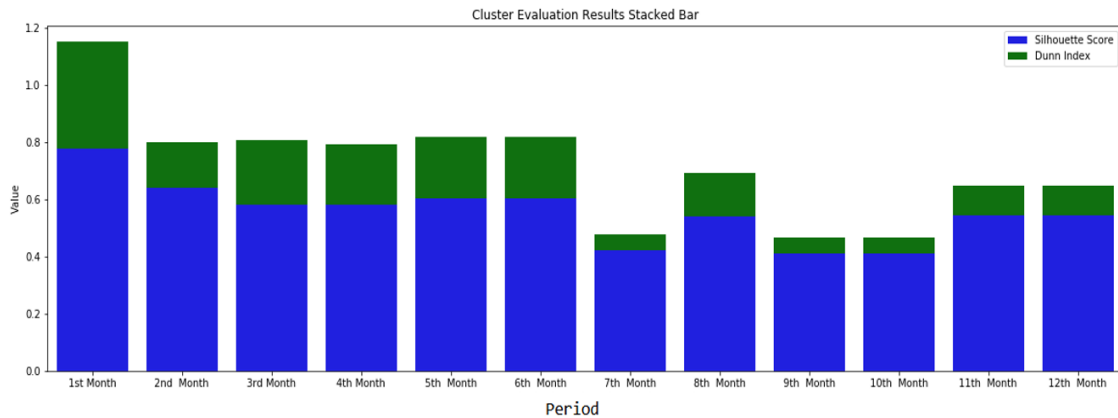


Figure 21. Dunn index and Silhouette index for monthly scenarios

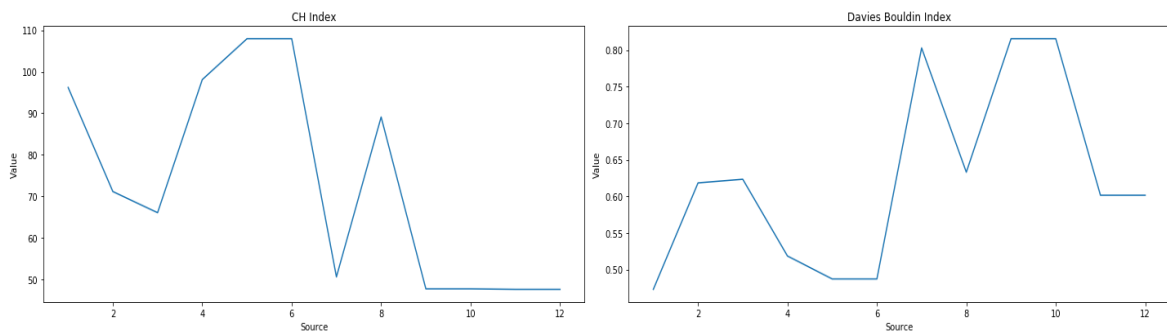


Figure 22. Calinski-Harabasz index and Davies-Bouldin index for monthly scenarios

The best cluster was found in the second quarter for the quarterly scenario depicted in Figure 23, which involved evaluating the cluster outcomes using the two matrices: the Silhouette index and the Dunn index. The results obtained through the evaluation of the two previous matrices were also obtained in the Calinski-Harabasz and Davies-Bouldin (DBI) matrices in Figure 24. The second quarter's index scores change significantly from those of the first, third, and fourth quarters.

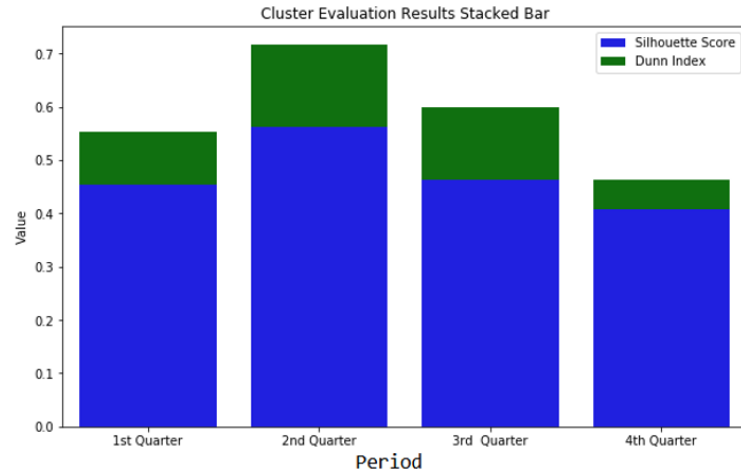


Figure 23. Dunn index and Silhouette index for quarterly scenarios

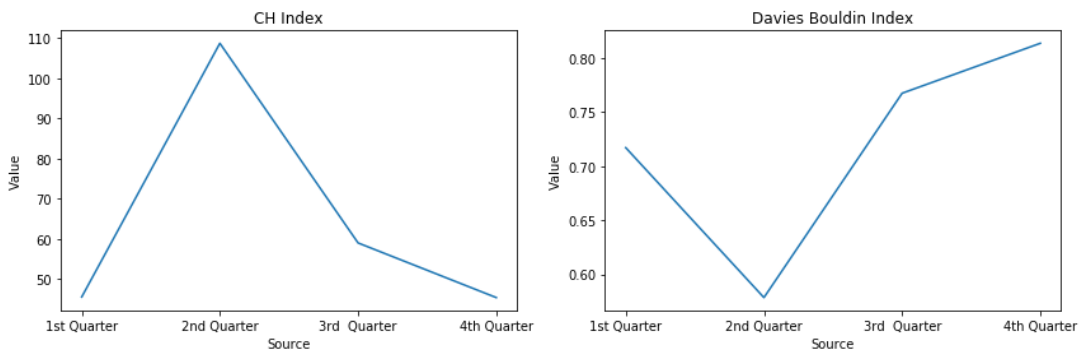


Figure 24. Calinski-Harabasz index and Davies-Bouldin index for quarterly scenarios

Lastly, Figure 25 and Figure 26 in the semesterly scenario show that the Silouhete index, the Dunn index, the Calinski-Harabasz, and the Davies-Bouldin index (DBI) matrix shown the same result, which show that first semester achieve the better cluster result compared to the second semester. If we compared all the metric from three kinds of scenarios, monthly scenario gives better performance results. This indicates that the evaluation of decision maker can be conducted effectively for each month.

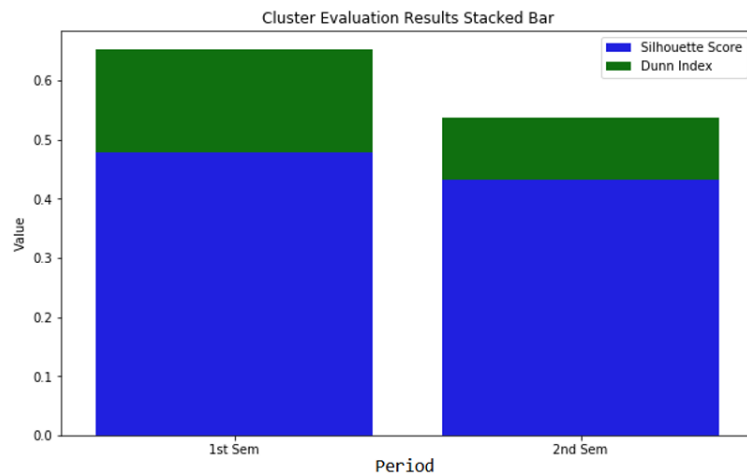


Figure 25. Dunn index and Silhouette index for semesterly scenarios

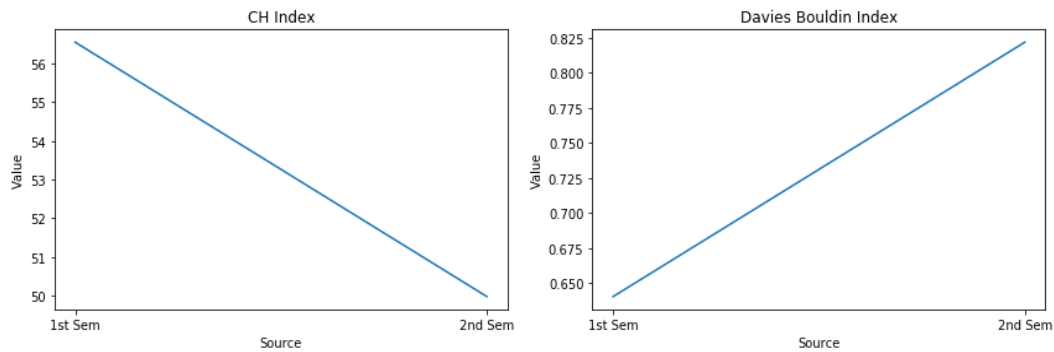


Figure 26. Calinski-Harabasz index and Davies-Bouldin index for semesterly scenarios

3.5. Discussion

Based on the results in previous sub section, the implementation of hierarchical clustering methods which focus on analyze the membership changes of each cluster based on its achievement number of new students from different months period observations was successfully illustrates that movements. This movement finds that monthly scenarios achieved the best metrics evaluation results compared to quarter and semesterly scenarios. In the first six months, the performance achievements of members of the first cluster far exceeded the performance of the second cluster. This conclusion is supported by our finding that in the 1st month until 6th month, the Dunn index, Silhouette index, CH index and also Davies Bouldin index value higher than 1st and 2nd quarter and also 1st semester. These results indicate that the decision maker can evaluate the admission performance earlier with monthly scenarios, so the school's strategy in first cluster can be imitated for other schools in second cluster. This shows that there is a tendency that 4 schools to have consistently good performance from the first month to the twelfth month.

Our findings align with previous studies that have reported the clustering can effectively monitor the student admission achievement performance by using clustering method. For instance, a study by [4] demonstrated k-means clustering for grouping student enrollment performance by using k-Means. However, our study extends these findings to investigate the student enrollment achievement performance in different periods of observation such as monthly, quarterly, and semesterly to identify the moving cluster member for each period. Despite these strengths, our study has limitations which determine fixed number of cluster in single value ($k=2$) and did not discuss the sub-clusters found through the agglomerative clustering method in more depth.

In summary, our study aimed to evaluate the effects of membership changes of each cluster in hierarchical clustering methods in three different periods. The significant results show that clustering using monthly scenarios can be best practice for decision makers to evaluate and make appropriate policy about that achievement. However, our findings also raise important questions did this finding is consistent if we add number cluster. Future research should focus on how number of clusters is optimal for monthly scenarios.

4. CONCLUSION

In conclusion, this study emphasizes exploring the hierarchical clustering process in the third observation period including the movement of clustering membership from one cluster to the other and analyzes what is the best cluster using four metrics namely the Silhouette index, the Dunn index, the Calinski-Harabasz, and the Davies-Bouldin index. The result showed that there is a movement membership from one period to another period. By focusing on dividing the cluster into 2 clusters in each observation period, the dendrogram shows that there is a massive movement starting in the 4th month to the 12th month for the monthly period, while for the quarterly scenario, the movement is smaller and in semesterly scenarios was the smallest. This is supported by the average value of each cluster showing a contrast between the first cluster and the second cluster from the 5th month to 12th month for the monthly scenario, the first quarter to the fourth quarter for the quarterly scenario, and the first and second semesters for the semesterly scenario.

The study also found that based on metric evaluation, it was found that the largest value was achieved when the different number of members in the first cluster compared to the second cluster was big. This can be seen in the monthly scenario where the number of members in the first cluster is only four while the number of members in the second cluster is forty-seven. The same thing can also be seen in the quarterly and semester scenarios, where the best cluster is when the difference in the number of members of the first

and fourth clusters is the largest, namely thirty-two in the first cluster and nine teens in the second cluster in the quarterly scenario and also thirty-five members of the first cluster and six-teen members of the second cluster in the semester scenario.

Moreover, the study also finds that there are any movement groups from one cluster to another cluster from one period to another period. This can be used by decision-makers to make new marketing strategies for schools which move from the first cluster to the second cluster or for schools which consistent in the second cluster from all periods. In future research, it is possible to expand the number of clusters not only for two clusters but also for three, four as the foundation needs.

ACKNOWLEDGEMENTS

This research was funded by Directorate of Research and Community Service at Telkom University through Grant No: 014/PNLT1/PPM/2023. We would like to express our sincere gratitude to DPSE YPT for providing the data that made this research possible. Their support and collaboration have been invaluable to our research and its successful completion. Without their generous provision of data, this work would not have been possible.




REFERENCES

- [1] W.-C. Yang, J.-P. Lai, Y.-H. Liu, Y.-L. Lin, H.-P. Hou, and P.-F. Pai, "Using medical data and clustering techniques for a smart healthcare system," *Electronics*, vol. 13, no. 1, p. 140, Dec. 2023, doi: 10.3390/electronics13010140.
- [2] H. W. Aqsari, D. D. Prastyo, and S. Puteri Rahayu, "Clustering stock prices of financial sector using k-means clustering with dynamic time warping," in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, Dec. 2022, doi: 10.1109/icitisee57756.2022.10057714.
- [3] K. Shahroodi, S. Avakh Darestani, S. Soltani, and A. Eisazadeh Saravani, "Developing strategies to retain organizational insurers using a clustering technique: Evidence from the insurance industry," *Technological Forecasting and Social Change*, vol. 201, p. 123217, Apr. 2024, doi: 10.1016/j.techfore.2024.123217.
- [4] T. Fahrudin, I. Asror, and Y. F. A. Wibowo, "Student enrollment performance of Telkom schools in 23/24 schoolyear using k-means clustering," in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, IEEE, Dec. 2023, doi: 10.1109/icitic60109.2023.10381939.
- [5] F. A. Bhat, M. Verma, and A. Verma, "Who will buy electric vehicles? Segmenting the young Indian buyers using cluster analysis," *Case Studies on Transport Policy*, vol. 15, p. 101147, Mar. 2024, doi: 10.1016/j.cstp.2024.101147.
- [6] A. Kanavos, I. Karamitsos, and A. Mohasseb, "Exploring clustering techniques for analyzing user engagement patterns in twitter data," *Computers*, vol. 12, no. 6, p. 124, Jun. 2023, doi: 10.3390/computers12060124.
- [7] C. Cambeses-Franco *et al.*, "A clustering approach to analyse the environmental and energetic impacts of Atlantic recipes - A Galician gastronomy case study," *Journal of Cleaner Production*, vol. 383, p. 135360, Jan. 2023, doi: 10.1016/j.jclepro.2022.135360.
- [8] N. Dalton-Barron *et al.*, "Clustering of match running and performance indicators to assess between- and within-playing position similarity in professional rugby league," *Journal of Sports Sciences*, vol. 40, no. 15, pp. 1712–1721, Aug. 2022, doi: 10.1080/02640414.2022.2100781.
- [9] L. Yin, M. Li, H. Chen, and W. Deng, "An improved hierarchical clustering algorithm based on the idea of population reproduction and fusion," *Electronics*, vol. 11, no. 17, p. 2735, Aug. 2022, doi: 10.3390/electronics11172735.
- [10] S. Shalileh, "An effective partitionial crisp clustering method using gradient descent approach," *Mathematics*, vol. 11, no. 12, p. 2617, Jun. 2023, doi: 10.3390/math11122617.
- [11] L. M. C. Cabezas, R. Izbicki, and R. B. Stern, "Hierarchical clustering: Visualization, feature importance and model selection," *Applied Soft Computing*, vol. 141, p. 110303, Jul. 2023, doi: 10.1016/j.asoc.2023.110303.
- [12] Z. Luo, L. Zhang, N. Liu, and Y. Wu, "Time series clustering of COVID-19 pandemic-related data," *Data Science and Management*, vol. 6, no. 2, pp. 79–87, Jun. 2023, doi: 10.1016/j.dsm.2023.03.003.
- [13] W.-B. Xie, Z. Liu, D. Das, B. Chen, and J. Srivastava, "Scalable clustering by aggregating representatives in hierarchical groups," *Pattern Recognition*, vol. 136, p. 109230, Apr. 2023, doi: 10.1016/j.patcog.2022.109230.
- [14] H.-J. Mucha, "On validation of hierarchical clustering," in *Advances in Data Analysis*, Springer Berlin Heidelberg, 2007, pp. 115–122, doi: 10.1007/978-3-540-70981-7_14.
- [15] P. Dráždilová, P. Prokop, J. Platoš, and V. Snášel, "A hierarchical overlapping community detection method based on closed trail distance and maximal cliques," *Information Sciences*, vol. 662, p. 120271, Mar. 2024, doi: 10.1016/j.ins.2024.120271.
- [16] S. K. Popat and Emmanuel M., "Review and comparative study of clustering techniques," in *International Journal of Computer Science and Information Technologies*, pp. 805–812, 2014.
- [17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
- [18] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/tpami.1979.4766909.
- [19] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, Jan. 1974, doi: 10.1080/01969727408546059.
- [20] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.
- [21] I. Pasina, G. Bayram, W. Labib, A. Abdelhadi, and M. Nurunnabi, "Clustering students into groups according to their learning style," *MethodsX*, vol. 6, pp. 2189–2197, 2019, doi: 10.1016/j.mex.2019.09.026.
- [22] O. Darcan and B. Badur, "Student profiling on academic performance using cluster analysis," *Journal of e-Learning & Higher Education*, pp. 1–8, Jan. 2012, doi: 10.5171/2012.622480.
- [23] M. G. Nitin, S. Gottipati, and V. Shankararaman, "Clustering models for topic analysis in graduate discussion forums," in *27th International Conference on Computers in Education*, 2019.




- [24] Z. T. Kosztyán, É. Orbán-Mihálykó, C. Mihálykó, V. V. Csányi, and A. Teles, “Analyzing and clustering students’ application preferences in higher education,” *Journal of Applied Statistics*, vol. 47, no. 16, pp. 2961–2983, Jan. 2020, doi: 10.1080/02664763.2019.1709052.
- [25] M. Halkidi, “Hierarchical clustering,” in *Encyclopedia of Database Systems*, Springer US, 2009, pp. 1291–1294, doi: 10.1007/978-0-387-39940-9_604.
- [26] E. Burghardt, D. Sewell, and J. Cavanaugh, “Agglomerative and divisive hierarchical Bayesian clustering,” *Computational Statistics & Data Analysis*, vol. 176, p. 107566, Dec. 2022, doi: 10.1016/j.csda.2022.107566.
- [27] X. Cai and Q. Wang, “Educational tool and active-learning class activity for teaching agglomerative hierarchical clustering,” *Journal of Statistics Education*, vol. 28, no. 3, pp. 280–288, Aug. 2020, doi: 10.1080/10691898.2020.1799727.
- [28] J. M. John, O. Shobayo, and B. Ogunleye, “An exploration of clustering algorithms for customer segmentation in the UK retail market,” *Analytics*, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi: 10.3390/analytics2040042.
- [29] A. Gere, “Recommendations for validating hierarchical clustering in consumer sensory projects,” *Current Research in Food Science*, vol. 6, p. 100522, 2023, doi: 10.1016/j.crf.2023.100522.
- [30] Q. Xu, Q. Zhang, J. Liu, and B. Luo, “Efficient synthetical clustering validity indexes for hierarchical clustering,” *Expert Systems with Applications*, vol. 151, p. 113367, Aug. 2020, doi: 10.1016/j.eswa.2020.113367.
- [31] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, Oct. 2001, doi: 10.1023/A:1012801612483.
- [32] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal clustering validation measures,” in *2010 IEEE International Conference on Data Mining*, IEEE, Dec. 2010, doi: 10.1109/icdm.2010.35.
- [33] A. Pita, F. J. Rodriguez, and J. M. Navarro, “Analysis and evaluation of clustering techniques applied to wireless acoustics sensor network data,” *Applied Sciences*, vol. 12, no. 17, p. 8550, Aug. 2022, doi: 10.3390/app12178550.
- [34] L. E. Ekemeyong Awong and T. Zielinska, “Comparative analysis of the clustering quality in self-organizing maps for human posture classification,” *Sensors*, vol. 23, no. 18, p. 7925, Sep. 2023, doi: 10.3390/s23187925.

BIOGRAPHIES OF AUTHORS






Tora Fahrudin    received the bachelor’s degree in informatics engineering from STT Telkom Bandung, Indonesia, in 2007, the master’s degree in management telecommunication from the Telkom Institute of Technology, Bandung, in 2010, and the Ph.D. degree in computer science from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 2019. Currently, he is a lecturer and a researcher with the School of Applied Science, Telkom University, Bandung, where he is also an associate professor. His current research interests include data science applied in various fields, such as finance, agriculture, and restaurant. He can be contacted at email: torafahrudin@telkomuniversity.ac.id.



Ibnu Asror    received the bachelor’s degree in informatics engineering from the Telkom Institute of Technology, Bandung, Indonesia, in 2009, the master’s degree in information technologies from the Telkom University, Bandung, in 2011, and currently Ph.D. degree in information technology from Telkom University, Bandung, Indonesia. His current research interests include data science and machine learning. He can be contacted at email: iasror@telkomuniversity.ac.id.



Yanuar Firdaus Arie Wibowo    holds a Bachelor of Science in informatics engineering from the Islamic University of Indonesia and a Master of Technology in information technology from Gadjah Mada University. Currently, he serves as a digital transformation expert consultant at PT Telkom Indonesia and PT Bhakti Unggul Teknovasi, while also holding the position of Director of Information Technology at the Indonesia International Islamic University (ex-officio). He has contributed to digital transformation in numerous higher education institutions in Indonesia, strategic planning, university rankings, digital journey formulation, and smart campus initiatives, as well as the implementation of information technology governance. His research interests include not only digital transformation but also sustainable smart campus development, IT strategic planning, IT governance, and information systems. He can be contacted at email: yanuar@telkomuniversity.ac.id.