



UNIVERSITAS BHAYANGKARA JAKARTA RAYA FAKULTAS ILMU KOMPUTER

Kampus I: Jl. Harsono RM No. 67, Ragunan, Pasar Minggu, Jakarta Selatan, 12550
Telepon: (021) 27808121 – 27808882
Kampus II: Jl. Raya Perjuangan, Marga Mulya, Bekasi Utara, Jawa Barat, 17142
Telepon: (021) 88955882, Fax.: (021) 88955871
Web: fasilkom.ubharajaya.ac.id, E-mail: fasilkom@ubharajaya.ac.id

SURAT TUGAS

Nomor: ST/309/III/2024/FASILKOM-UBJ

Pertimbangan : Dalam rangka mewujudkan Tri Dharma Perguruan Tinggi untuk Dosen di Universitas Bhayangkara Jakarta Raya maka dihimbau untuk melakukan penelitian.

Dasar : 1. Kalender Akademik Universitas Bhayangkara Jakarta Raya Tahun Akademik 2023/2024.
2. Rencana Kerja dan Anggaran Pembelanjaan Universitas Bhayangkara Jakarta Raya Tahun 2024.

DITUGASKAN

Kepada : Personil yang namanya tercantum dalam Surat Tugas ini.

NO.	NAMA	NIDN	JABATAN	KETERANGAN
1.	Dr. Tb. Ai Munandar, S.Kom., M.T.	0413098403	Dosen Tetap Prodi Informatika	Sebagai Penulis Pertama
2.	Ajif Yunizar Pratama Yusuf, S.Si., M.Eng.	0328068603	Dosen Tetap Prodi Informatika	Sebagai Penulis Kedua

Untuk : 1. Membuat Artikel Ilmiah dengan judul "**Regional Clustering Based on Types of Non-Communicable Diseases Using K-Means Algorithm**" dengan menerima LoA pada tanggal 05 Januari 2024 untuk dipublikasikan di media Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, Vol. 23, No. 1, Maret 2024, Hal. 285 - 296, ISSN: 2476-9843.
2. Melaksanakan tugas ini dengan penuh tanggung jawab.

Jakarta, 04 Maret 2024
DEKAN FAKULTAS ILMU KOMPUTER

Dr. Dra. Tyastuti Sri Lestari, M.M.
NIP. 1408206

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Tb Ai Munandar, Ajif Yunizar Yusuf Pratama
Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia

Article Info

Article history:

Received September 05, 2023
Revised November 13, 2023
Accepted January 05, 2024

Keywords:

Clustering
k-Means
Non-Communicable Diseases
Regional Clustering
Silhouette Index

ABSTRACT

Noncommunicable diseases (NCDs) have become a global threat to public health, necessitating a comprehensive understanding of their geographic and epidemiological distribution to devise appropriate interventions. This study aims to cluster Banten Province areas based on NCDS profiles using the unsupervised learning technique. The method used in this study is the k-means algorithm for grouping types of non-communicable diseases based on region. The processing and normalisation of NCDS prevalence data from various health sources preceded cluster analysis using the k-means clustering algorithm. This research is categorised into two scenarios: the first involves clustering data obtained from outlier analysis, while the second excludes any outliers. The objective is to observe disparities in regional clustering outcomes by categorising non-communicable diseases according to these two scenarios. The silhouette index is used to determine the validity of cluster results. These findings are analysed to determine the geographic and socioeconomic patterns associated with each cluster's NCDS profile. Based on the mean silhouette index value of 0.812, the results indicate that the sum of $k = 2$ in the k-means algorithm is the optimal cluster result. Five non-communicable diseases, namely diabetes, hypertension, obesity, stroke, and cataracts, necessitate significant focus in the first cluster (C1), where 202 regions were grouped. Six regions belong to the second cluster (C2), which includes areas that are not only susceptible to the five non-communicable diseases in cluster C1 but also to breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

Copyright ©2022 The Authors.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Tb Ai Munandar, +6281384512710
Faculty of Computer Science, Informatics Department,
Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia.
Email: tbaimunandar@gmail.com

How to Cite:

T. A. Munandar and A. Y. Y. Pratama, "Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 2, pp. 285-296, March, 2024. This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

1. INTRODUCTION

Non-communicable diseases (NCDs) comprise a collection of chronic health conditions that do not spread via direct human-to-human contact and are the leading cause of cancer-related deaths worldwide. This grouping contains a variety of illnesses, including chronic respiratory conditions, diabetes, cardiovascular disease, and cancer, which collectively impose a significant burden on global health. The gravity of non-communicable diseases (NCDs) is emphasised by the World Health Organisation (WHO), which estimates that these conditions are responsible for around 41 million deaths per year, or 71% of all casualties worldwide [1–4]. The development of non-communicable diseases (NCDs) is strongly associated with detrimental lifestyle decisions, which include unequal dietary patterns, a lack of physical activity, and the use of tobacco and alcohol. Due to urbanisation, evolving behaviours, and rapid population growth, the Indonesian province of Banten faces substantial challenges associated with non-communicable diseases (NCDs). In 2021, Banten Provincial Health Service released data that indicates Banten Province has experienced an ongoing and consistent increase in the prevalence of diabetes, hypertension, and obesity. The substantial increase in the prevalence of non-communicable diseases presents a tough obstacle for the regional health infrastructure, significantly affecting the community's overall standard of living [5, 6]. To address their complexity, an in-depth examination of the patterns and distribution of non-communicable diseases (NCDs) in Banten is necessary. The current regional grouping system predicated on NCD categories may not be ideal when developing targeted interventions.

Applying unsupervised learning techniques, specifically the cluster method, to divide Banten Province into groups based on different types of non-communicable diseases (NCDs) is the point of this research. Clustering presents a more sophisticated methodology than conventional grouping techniques, enabling the detection of inherent groupings within the data that do not require predetermined labelling. This study seeks a more precise comprehension of the distribution of non-communicable diseases (NCDs) throughout Banten Province by delineating regional clusters according to different NCD categories. For informing health policies and directing medical interventions, precise and comprehensive data on the prevalence and distribution of non-communicable diseases (NCDs) in various regions of Banten Province are indispensable. By allowing health policymakers to customise more efficacious prevention and intervention strategies, the regional clustering inferred from this study has the potential to yield significant insights regarding the spatial distribution of non-communicable disease prevalence. The resulting regional groups will provide a solid foundation for developing targeted non-communicable disease (NCD) prevention programmes, letting policymakers and medical professionals pay close attention to the specific needs of each cluster. Using unsupervised learning techniques, particularly clustering, to illustrate the intricate picture of NCD prevalence, this study's primary objective is to aid in the improvement of public health strategies in Banten Province; by employing this novel methodology, the research endeavours to establish a strong groundwork for decision-making grounded in evidence, thereby promoting a more sophisticated and focused approach to addressing the complex issues presented by non-communicable diseases in the area.

To face the challenges of NCDS in the province of Banten, a comprehensive and integrated approach to grouping areas based on the nature of NCDS is required. Currently, regional groupings are typically determined by administration or geographic location without considering each region's unique health profile. We can identify patterns and relationships in Banten Province NCDS data using unsupervised learning techniques, such as the clustering method. This phase is essential for designing more targeted interventions and enhancing health services in regions affected by NCDS. Moreover, research conducted by [7] on the epidemiology of hypertension in India suggests that accelerated urbanisation may contribute to the rising prevalence of hypertension in urban areas. The prevalence of obesity tends to be higher in urban areas than rural areas, highlighting the significance of taking regional differences into account when managing NCDs. These results indicate that the social and economic changes that result from urbanisation can affect disease patterns in the region, and it is not inconceivable that this could occur in Banten Province. Noncommunicable diseases (NCDs) are a major public health concern because they contribute to the world's elevated mortality rate and disease burden. Numerous studies have been conducted in recent years to understand the risk factors and transmission patterns of NCDS and to identify more effective prevention strategies. A study by Chen provides the most recent data regarding the prevalence and risk factors of type 2 diabetes in various populations. The findings of this study shed light on the association between diet and physical activity and the risk of type 2 diabetes. Several studies have investigated the impact of diet in reducing the risk of cardiovascular disease as part of efforts to prevent NCDs. For instance, research found a correlation between adult fast-food consumption and increased insulin resistance. Additionally, recent research has emphasised the significance of optimal sleep patterns in reducing the risk of NCDs. A study by [8] discovered that adult sleep deprivation increases the risk of adiposity. To comprehend the effect of physical activity on the prevention of NCDs, Daskalopoulou (2021) conducted a meta-analysis which revealed that higher levels of physical activity are associated with lower risks for certain NCDs. Recent research has also focused on the environment's role in NCDS. The composition of the intestinal microbiota is associated with the development of cardiovascular disease, according to [9, 10]. In terms of NCDS prevention, understanding the function of the environment is crucial. Zimmet, in 2021, investigated the global repercussions of the diabetes epidemic and other NCDs. In addition to focusing on NCDS in the adult population, researchers have also studied the

disorder in children. At the same time, Samavat 2021 examined the association between a child's adult diet and their future risk of breast cancer [11, 12]. However, as mentioned earlier, a portion of the research places greater emphasis on discerning various intervention methods and enhancing healthcare provisions by utilising assessments of individuals afflicted with non-communicable diseases. The topic of regional-specific approaches to healthcare intervention and management has not yet been addressed. The issue of non-communicable illness spreading often remains unresolved to some extent due to a lack of information regarding treatment priorities specific to different regions. It is imperative to adopt a regional cluster-based approach to enhance health services and interventions in the future.

The distinction between the present and prior studies categorises non-communicable diseases based on geographical regions. By implementing interventions and making enhancements, healthcare quality can be significantly improved. Collectively, the most recent research provides valuable insights into the effective management of NCDS. Through a greater understanding of risk factors and prevention strategies, it is anticipated that targeted preventive measures can be implemented to lessen the burden of non-communicable diseases (NCDs) and improve public health as a whole. Noncommunicable diseases (NCDs) are a significant global health burden, and it is essential to understand the patterns and patterns of dissemination of these diseases in a specific region for prevention and appropriate management. The unsupervised learning method has become a valuable instrument in analysing health data, including grouping regions by NCD type. Several recent studies have utilised unsupervised learning techniques, such as clustering and cluster analysis, to identify geographic and epidemiological patterns of NCDs in different regions. A study by [13] grouped regions based on the type of NCDS in a country using cluster analysis. This study reveals how to identify the most and least burdened counties. A related study by [14] and [15] utilised spatial clustering and unsupervised learning to categorise regions based on the pattern of cardiovascular disease distribution. The results of this study indicate certain health clusters that can be targeted by interventions to reduce the risk of cardiovascular disease. In addition, unsupervised learning techniques have been employed to determine the connection between air pollution patterns, the urban spread of respiratory diseases and health interventions. Research by [16] found that cluster patterns from air pollution data and the incidence of respiratory disease are frequently interconnected. Recent research has also investigated the use of machine learning, including unsupervised learning, for grouping regions based on other categories of diseases, such as cancer, benign and malignant tumours, diabetes, and neurodegenerative diseases [17–21]. To address the challenges posed by NCDS at the regional level, an unsupervised learning algorithm was used to identify patterns of interrelationships between specific NCDS in specific regions of a country [22]. Also, at the same time, some researchers grouped regions based on the level of exposure to certain NCDS risk factors using an unsupervised learning method [23, 24]. Overall, the application of unsupervised learning has created new opportunities to segment regions based on the type of NCDs and to gain a deeper understanding of disease transmission patterns; one of the most popular algorithms is k-means. Using this strategy, it is anticipated that prevention and intervention can be more effectively targeted based on the local community's health characteristics. The k-means algorithm is used for a variety of reasons. Apart from being frequently used in health research, it may also be used to segment promotional places in education, facilities, and teachers [25, 26] and group poverty indicators in a region [27]. This study aims to categorise non-communicable diseases based on geographical regions by analysing patient data obtained from public health institutions in Banten Province. This research is anticipated to significantly impact local government's ability to identify regional priorities for managing the development of non-communicable diseases. In addition, unsupervised learning methods, such as the k-means algorithm, can be utilised as an alternate strategy to analyse non-communicable disease data, making a valuable contribution to the health sector.

This paper is divided into four sections. The first section presents the introduction, which includes the problem's background, a literature review, research gaps, and systematic information about the article. The second section explains the study methodologies used, and the third section includes the research results and commentary. Meanwhile, the fourth portion is the conclusion, which contains the investigation findings.

2. RESEARCH METHOD

This study uses unsupervised learning as its research methodology to categorise regions within Banten Province according to the specific noncommunicable disease (NCD) type. Figure 1 provides a more detailed overview of the study stages.

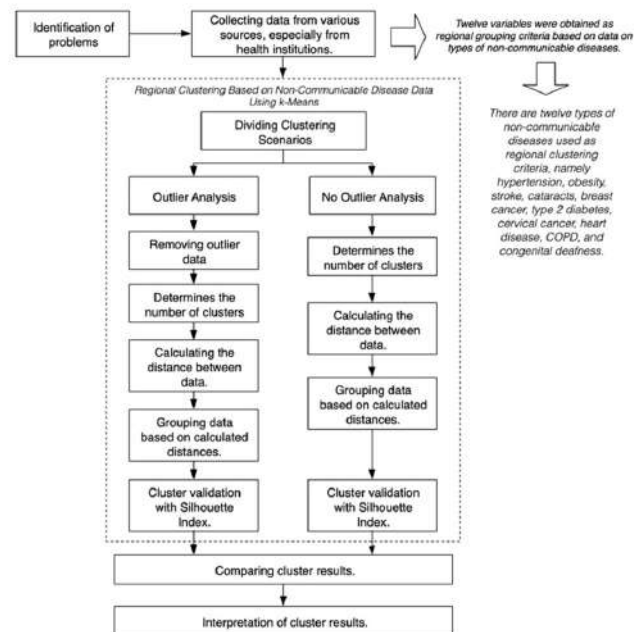


Figure 1. Research Stages

During the preliminary stage of the inquiry, data pertaining to the incidence of non-communicable diseases (NCDs) were gathered from several sources, encompassing public health institutions and hospitals situated within Banten Province. A total of 227 instances of regional information representing subdistricts were gathered due to the data-gathering process. The data is subsequently submitted to a preprocessing stage, wherein incomplete or fragmented data are repaired, and all variables are normalised to guarantee consistency in scale. Consequently, the strategy for categorising regions based on NCDs profiles involved using the k-means clustering algorithm, which falls under the umbrella of unsupervised learning techniques. Cluster analysis is conducted by determining the distance between the data points and the centroid of each cluster. This process involves combining regions that exhibit similar NCDs profiles into a cohesive cluster. The quality of the clusters was evaluated by performing validation of the cluster results using the silhouette index. The clustering criteria are determined by twelve specific non-communicable diseases, which include hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, COPD, and congenital deafness.

2.1. K-Means

The k-means algorithm is a commonly employed clustering technique that divides a dataset into separate groups according to their similarity. This is achieved through an iterative process of assigning data points to the nearest cluster centroid and adjusting the centroids to minimise the total sum of squared distances. The method's success in numerous domains can be attributed to its efficiency and simplicity despite the acknowledged drawbacks of being sensitive to initial centroid selection and prone to converging to local optima [28, 29]. The k-means algorithm is widely recognised for its prevalence in unsupervised learning. It is commonly employed in several domains, such as picture segmentation, customer segmentation, and anomaly detection. Although this method is extensive, practitioners must exercise caution when interpreting the findings. It is important to acknowledge that the outcomes can vary depending on the initialisation and scale factors [30, 31]. The phases of the k-means algorithm are as follows:

1) Initialization

- Choose the number of clusters, k .
- Initialize k centroids.

2) Assignment Step:

- For each data point in your dataset, calculate the distance (e.g., Euclidean distance) to each k centroid using Equation (1).

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (1)$$

- Assign the data point to the cluster whose centroid is closest to it. This forms k clusters.

3) Update Step:

- Calculate the new centroids of the clusters based on the data points assigned to each cluster. This is done by computing the mean of all data points within each cluster using Equation (2).

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} X_q \quad (2)$$

4) Convergence Check:

- Check if the centroids have changed significantly from the previous iteration.
- If the centroids have changed significantly, repeat steps 2 and 3 (Assignment and Update) until convergence. If not, the algorithm has converged, and the final centroids represent the cluster centres.

5) Termination:

- The algorithm terminates once the centroids no longer change significantly or after a predetermined number of iterations.

2.2. Silhouette Index

The silhouette index is a commonly employed statistic in evaluating clustering outcomes since it quantifies the degree of separation and compactness exhibited by the clusters. The metric quantifies the degree of clustering for each data point in relation to its neighbouring cluster, hence offering valuable insights about the suitability of the selected number of clusters and the success of the clustering algorithm. The silhouette index is a metric that falls within the range of -1 to 1, with higher values indicating more well-defined clusters. A number in proximity to 1 indicates the presence of distinct and independent clusters, whilst values in proximity to 0 show the existence of overlapping clusters. Negative values, on the other hand, reflect the possibility of erroneous assignment of data points to clusters. The silhouette index has become widely recognised and valued in the field primarily because of its intuitive interpretation and its effectiveness in accommodating diverse cluster shapes and densities [32–35]. The silhouette index can be derived by utilising Equation (3).

$$S_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max\{a_{(i)}, b_{(i)}\}} \quad (3)$$

Where the variable $a_{(i)}$ represents the average dissimilarity between the i th object and all other objects within the same cluster, the variable $b_{(i)}$ represents the average dissimilarity between the i th object and all other objects in the nearest cluster. The values of $a_{(i)}$ and $b_{(i)}$ can be derived by utilising Equations (4) and (5).

$$a_{(i)} = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \quad (4)$$

$$b_{(i)} = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \quad (5)$$

The variable $|C_I|$ represents the cardinality of cluster C_I , which denotes the number of data points belonging to that specific cluster. On the other hand, $d(i, j)$ represents the distance between two data points, i and j , within the cluster C_I .

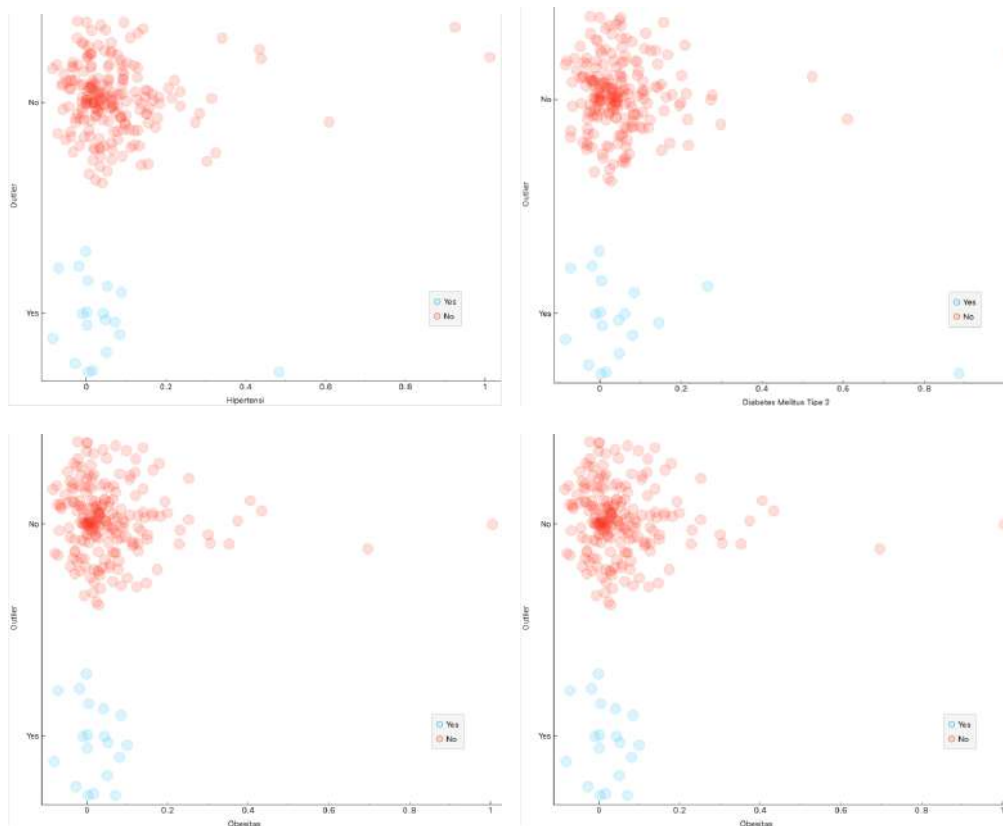


Figure 2. Results of Outlier Analysis for Some Features

3. RESULT AND ANALYSIS

This section contains two subsections. The first subsection describes the research findings, while the second section discusses the findings, including the interpretation of cluster outcomes.

3.1. Results

Outlier analysis is essential in the initial phases of data clustering, serving as a critical step to detect and assess occurrences that exhibit substantial deviation from the average. The main goal is to improve the precision and dependability of the following categorisation procedure. Within this particular context, the analysis fulfils a dual function, aiding in dividing data into two clearly defined scenarios.

In the first scenario, the clusters are carefully organised depending on the results of the outlier analysis. This strategy guarantees that the categorisation procedure is carried out on a purified dataset devoid of the impact of outliers. This strategy aims to identify and emphasise groups that demonstrate exceptional quality by carefully dealing with any exceptional data points. By isolating these clusters, researchers can extract more significant insights and provide more precise forecasts. In contrast, the second scenario entails data grouping without considering the outcomes of the outlier analysis. This methodology offers a comparison method, enabling researchers to evaluate the influence of outliers on the clustering procedure. It assists in comprehending the degree to which outliers impact the overall grouping and aids in assessing the resilience of the clustering algorithm under various circumstances. To conduct a thorough analysis of outliers, various health-related attributes such as hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness were examined closely. The selection of these attributes encompasses a wide range of health-related criteria, guaranteeing a comprehensive evaluation of exceptional cases across multiple dimensions. All of the given health characteristics were found to be outliers, according to the results of the outlier analysis. Along with Pondok Ranji, Sepatan, Banjar, Sukadiri, Cisauk, Bhakti Jaya, Kresek, Mancak, Cihara,

Ciledug, Rengas, Tunjung Teja, Cibitung, Angsana, Sindang Resmi, Cinangka, Anyar, Rajeg, and Carengang, the outliers were found in 19 different datasets or regions. Figure 2 displays a scatter plot that visualises the outliers for a particular feature.

Upon completion of the outlier analysis, the results are incorporated into the clustering procedure in the initial study scenario. This eliminates 19 data points from the entire regional dataset. The ideal number of clusters is determined using the average silhouette index value before starting the clustering process. The k-means widget in the Orange data mining framework allows for emulating an optimal number of clusters by specifying a defined range of cluster numbers. In the current study context, where dataset normalisation is lacking, the specified range extends from 2 to 8 clusters. The cluster distribution shows unevenness, as indicated by the incomplete attainment of the optimal average silhouette index for the comprehensive cluster set. The uneven distribution highlights the need for more improvement within the clustering architecture. Data normalisation plays a crucial role in improving the results of clustering. After the normalisation process, there is a noticeable increase in the average silhouette index value for each cluster. This rise demonstrates an enhanced level of unity and distinctiveness among the clusters. These findings demonstrate that normalising the data has a beneficial impact on enhancing the quality of the clustering outcomes. Despite attempts to normalise the data, the cluster arrangement that yields the highest silhouette index remains at $k = 2$. The current setup produces an average silhouette index of 0.81, indicating strong distinction and unity among the clusters. The comparative silhouette indices for different cluster topologies, with values of k equal to 3, 4, 5, 6, 7, and 8, are 0.701, 0.664, 0.649, 0.456, 0.494, and 0.514, respectively. The values provide evidence for the exceptional quality and consistency of the $k = 2$ cluster setup. To summarise, the repeated incorporation of outlier analysis and subsequent clustering procedures is crucial in extracting detailed insights from complex datasets. The methodical process of refining the data, which includes removing outliers and normalising the data, enhances the optimisation of cluster quality. Using the silhouette index as a metric helps determine the optimal number of clusters, leading to a better understanding of the patterns in the dataset.

Table 1. Distribution of Cluster Member for Each k in the First Scenario Two

silhouette-NN	0,789	0,728	0,732	0,667	0,388	0,419	0,42
silhouette-Nr	0,812	0,701	0,664	0,649	0,456	0,494	0,514
Number of k	2	3	4	5	6	7	8
C1-Nr	202	193	14	6	157	12	13
C1-NN	10	191	193	185	1	134	134
C2-Nr	6	14	189	5	5	5	2
C2-NN	198	16	12	17	5	6	5
C3-Nr	0	1	4	1	7	20	153
C3-NN	0	1	1	1	122	1	1
C4-Nr	0	0	1	184	32	6	2
C4-NN	0	0	2	2	12	2	2
C5-Nr	0	0	0	12	1	158	5
C5-NN	0	0	0	3	62	3	1
C6-Nr	0	0	0	0	6	6	9
C6-NN	0	0	0	0	6	45	45
C7-Nr	0	0	0	0	0	1	23
C7-NN	0	0	0	0	0	17	17
C8-Nr	0	0	0	0	0	0	1
C8-NN	0	0	0	0	0	0	3

Initially, the clustering method yields that 202 regions are assigned to the first cluster (C1), while the remaining 6 regions are allocated to the second cluster (C2). More precisely, there are 193 regions grouped as members of category C1, 14 regions assigned to category C2, and one region allocated to category C3 in clusters where the value of k equals 3. Table 1 presents the detailed distribution of cluster members for each k , categorised based on their normalisation state (normalised is indicated as Nr and non-normalised is denoted as NN). The data presented in Table 1 indicates that when the value of k decreases, there is a stronger inclination for individuals from one cluster to merge with individuals from other clusters with a larger number of members, reducing the number of members in the remaining clusters. This discovery highlights the correlation between the selected number of clusters and the distribution patterns of regional members. Furthermore, the arrangement of cluster participants, whether standardised or not, demonstrates the enduring presence of regional clustering. However, it is important to mention that the occurrence of regional clustering only undergoes a slight change in its location. Although there may be differences in data normalisation, the overall trend of regional clustering stays generally stable. Table 1 presents the outcomes of the clustering study, illustrating the variations in the number of clusters selected and the distribution patterns of regional members over time. Regional agglomeration is evident in both normalised and non-normalised datasets, and any disparities in their manifestation are relatively inconsequential.

In the second clustering scenario, the dataset is straightaway processed using the k-means method without any prior outlier data removal. The cluster outcomes show a pattern where, as the value of k lowers, the cluster member areas become more concentrated within one cluster or specific clusters. Significantly, the number of cluster members in C1 (with data normalisation indicated as Nr) exceeds that of C2 when k equals 2. In contrast, if the data is not normalised, a reverse correlation arises, where cluster C2 includes a greater number of members compared to cluster C1. This trend remains consistent for the majority of other k values. The distribution of cluster members shows inequality between $k = 2$ and $k = 8$. However, as k increases, there is a noticeable pattern where the distribution of cluster members becomes more spread out, although the overall pattern of members clustering together remains unchanged. In the second clustering scenario, Table 2 provides detailed information about the distribution of cluster members for each value of k . Furthermore, Figure 3 presents a graphical depiction of the clustering results, differentiating between normalised (a) and non-normalised data (b), respectively.

Table 2. Distribution of Cluster Member for Each k for the Second Scenario

silhouette-NN	0,769	0,682	0,595	0,613	0,566	0,57	0,467
silhouette-Nr	0,805	0,724	0,666	0,696	0,404	0,466	0,463
Number of k	2	3	4	5	6	7	8
C1-Nr	218	19	12	17	16	6	41
C1-NN	18	36	58	13	7	22	21
C2-Nr	9	206	205	1	1	166	2
C2-NN	209	188	151	2	50	7	7
C3-Nr	0	2	2	2	2	7	167
C3-NN	0	3	2	151	6	139	101
C4-Nr	0	0	8	204	147	1	1
C4-NN	0	0	16	55	139	49	38
C5-Nr	0	0	0	3	3	2	3
C5-NN	0	0	0	6	2	6	6
C6-Nr	0	0	0	0	58	44	5
C6-NN	0	0	0	0	23	2	2
C7-Nr	0	0	0	0	0	1	1
C7-NN	0	0	0	0	0	2	2
C8-Nr	0	0	0	0	0	0	7
C8-NN	0	0	0	0	0	0	50

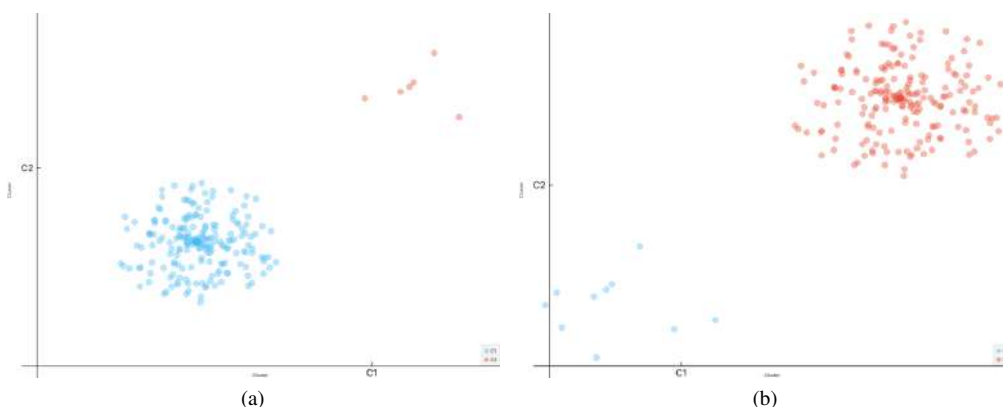


Figure 3. The difference between cluster findings with normalisation (a) and without normalisation (b)

3.2. Discussion

The cluster analysis of the two scenarios produces significant results, especially when evaluating the average silhouette index values. The initial analysis reveals that the highest average silhouette index value is observed when k is set to 2. This finding serves as the fundamental criterion for determining the value of k in the following analyses conducted in this inquiry. Each cluster member at $k = 2$ demonstrates a silhouette value near 1, suggesting strong cohesiveness and isolation within the clusters. Aligning with

previous studies conducted by [32] and [36], the Silhouette Index determines the most suitable number of clusters. According to their statement, this index can be utilised to ascertain the most suitable number of clusters before clustering. The results, especially after data normalisation, confirm that the clustering results with $k = 2$ have higher validity than other configurations. Previous studies have found that normalising data before applying machine learning approaches can enhance the effectiveness of both clustering [37, 38], and classification [39, 40], according to research findings. Hence, it is unsurprising that in this study, the application of normalisation can enhance cluster validity, thereby enabling it to ascertain the most suitable number of clusters.

Confirming the cluster results with a value of $k = 2$ shows that 202 regions are grouped under cluster C1, while the rest are assigned to cluster C2. By lining up these cluster results with infographics, it's clear that cluster C2 includes areas with almost all 10 types of noncommunicable diseases that were looked at in this study. Cluster C1 exhibits the largest frequency of individuals with non-communicable diseases, particularly for the five most common conditions: diabetes, hypertension, obesity, stroke, and cataracts. In contrast, the 202 locations in cluster C1 have a remarkably low occurrence of diseases such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

The results of this study emphasise the urgent requirement for focused attention in up to 202 regions, especially those vulnerable to the occurrence of the five illnesses: diabetes, hypertension, obesity, stroke, and cataracts. At the same time, the six places that makeup cluster C2 need immediate attention because they have high rates of the five diseases listed above, as well as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This investigation confirms the crucial significance of customised healthcare interventions based on specific regional health profiles. The focus on clusters C1 and C2 enables the implementation of specific initiatives that recognise the distinct patterns of non-communicable illnesses in each cluster. Public health programmes must prioritise addressing the distinct healthcare requirements of these clusters by customising interventions to tackle prevalent disorders and alleviate the burden of diseases in the designated locations.

4. CONCLUSION

To summarise, the research results emphasise the urgent requirement for specific healthcare interventions in the province of Banten, specifically about non-communicable diseases (NCDs). Using the k-means algorithm to do clustering analysis on NCD markers for 208 regions shows how important each part of the province has its unique health profile. The clustering results, especially when k is set to 2, indicate that 202 regions require urgent attention because of the high occurrence of diabetes, hypertension, obesity, stroke, and cataracts. Besides the main diseases listed above, the province's other 28 regions deal with a wider range of noncommunicable diseases (NCDs), such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This extensive analysis of regional health profiles establishes a basis for focused public health initiatives, highlighting the imperative need for government action in specified locations.

Nevertheless, it is imperative to recognise the constraints of the clustering methodology utilised in this research. Although the k-means algorithm efficiently classifies regions into two main clusters based on NCD markers, it does not provide information about the spatial proximity of NCD sufferers across regions. Given this constraint, it is clear that further research using other methods is necessary to thoroughly investigate the pattern of closeness among individuals with non-communicable diseases in different areas. This necessitates an intricate method of comprehending the frequency of particular illnesses in each group and the interaction of geographical elements that impact healthcare dynamics.

Future research needs to concentrate on improving techniques for capturing the spatial linkages and proximity patterns among individuals with non-communicable diseases (NCDs). This deeper comprehension will enable the implementation of more accurate and customised healthcare plans, guaranteeing that interventions are customised to the individual requirements of each location. Integrating spatial analyses into future research will enhance the effectiveness of tackling the intricate terrain of non-communicable diseases in the province of Banten.

5. DECLARATIONS

AUTHOR CONTRIBUTION

All authors contributed to the writing of this article.

FUNDING STATEMENT

This research was self-funded, and the authors did not receive any external financial support for the design, data collection, analysis, or interpretation of the study. All expenses related to this research were borne by the authors personally.

COMPETING INTEREST

The authors declare no conflict of interest in this article.

REFERENCES

- [1] S. Bhattacharya, P. Heidler, and S. Varshney, "Incorporating neglected non-communicable diseases into the national health program: A review," *Frontiers in Public Health*, vol. 10, no. January, pp. 1–9, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.1093170/full>
- [2] M. F. Owusu, J. Adu, B. A. Dorte, S. Gyamfi, and E. Martin-Yeboah, "Exploring health promotion efforts for non-communicable disease prevention and control in Ghana," *PLOS Global Public Health*, vol. 3, no. 9, pp. 1–14, 2023. [Online]. Available: <https://dx.plos.org/10.1371/journal.pgph.0002408>
- [3] A. Odunyemi, T. Rahman, and K. Alam, "Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19," *BMC Public Health*, vol. 23, no. 1, pp. 1–12, 2023. [Online]. Available: <https://bmcpublihealth.biomedcentral.com/articles/10.1186/s12889-023-16498-7>
- [4] H. G. A. S. Samarasinghe, D. A. T. D. S. Ranasinghe, W. R. Jayasekara, S. A. A. D. Senarathna, J. D. P. M. Jayakody, P. M. Kalubovila, M. D. Edirisuriya, and N. S. A. S. N. Senarath, "Barriers to Accessing Medical Services and Adherence to Recommended Drug Regimens among Patients with Non-Communicable Diseases: A Study at Divisional Hospital Thalagama, Sri Lanka," in *IECN 2023*. MDPI, 2023, pp. 1–6. [Online]. Available: <https://www.mdpi.com/2673-9976/29/1/14>
- [5] K. S. Maliangkay, U. Rahma, S. Putri, and N. D. Istanti, "Analisis Peran Promosi Kesehatan Dalam Mendukung Keberhasilan Program Pencegahan Penyakit Tidak Menular Di Indonesia," *Jurnal Medika Nusantara*, vol. 1, no. 2, pp. 108–122, 2023.
- [6] H. B. H. Akbar, and S. Sarman, "Pencegahan Penyakit Tidak Menular Melalui Edukasi Cerdik Pada Masyarakat Desa Moyag Kotamobagu," *Abdimas Universal*, vol. 3, no. 1, pp. 83–87, 2021. [Online]. Available: <http://abdimasuniversal.uniba-bpn.ac.id/index.php/abdimasuniversal/article/view/94>
- [7] R. Gupta, K. Gaur, and C. V. S. Ram, "Emerging trends in hypertension epidemiology in India," *Journal of Human Hypertension*, vol. 33, no. 8, pp. 575–587, 2019. [Online]. Available: <https://www.nature.com/articles/s41371-018-0117-3>
- [8] C. Antza, G. Kostopoulos, S. Mostafa, K. Nirantharakumar, and A. Tahrani, "The links between sleep duration, obesity and type 2 diabetes mellitus," *Journal of Endocrinology*, vol. 252, no. 2, pp. 125–141, 2022. [Online]. Available: <https://joe.bioscientifica.com/view/journals/joe/252/2/JOE-21-0155.xml>
- [9] L. Wang, S. Wang, Q. Zhang, C. He, C. Fu, and Q. Wei, "The role of the gut microbiota in health and cardiovascular diseases," *Molecular Biomedicine*, vol. 3, no. 1, pp. 1–50, 2022. [Online]. Available: <https://link.springer.com/10.1186/s43556-022-00091-2>
- [10] A. A. Samarraie, M. Pichette, and G. Rousseau, "Role of the Gut Microbiome in the Development of Atherosclerotic Cardiovascular Disease," *International Journal of Molecular Sciences*, vol. 24, no. 6, pp. 1–17, 2023. [Online]. Available: <https://www.mdpi.com/1422-0067/24/6/5420>
- [11] R. T. Chlebowski, J. Luo, G. L. Anderson, W. Barrington, K. Reding, M. S. Simon, J. E. Manson, T. E. Rohan, J. Wactawski-Wende, D. Lane, H. Strickler, Y. Mosaver-Rahmani, J. L. Freudenheim, N. Saquib, and M. L. Stefanick, "Weight loss and breast cancer incidence in postmenopausal women," *Cancer*, vol. 125, no. 2, pp. 205–212, 2019. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.31687>
- [12] M. Ellingjord-Dale, S. Christakoudi, E. Weiderpass, S. Panico, L. Dossus, A. Olsen, A. Tjønneland, R. Kaaks, M. B. Schulze, G. Masala, I. T. Gram, G. Skeie, A. H. Rosendahl, M. Sund, T. Key, P. Ferrari, M. Gunter, A. K. Heath, K. K. Tsilidis, and E. Riboli, "Long-term weight change and risk of breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) study," *International Journal of Epidemiology*, vol. 50, no. 6, pp. 1914–1926, 2022. [Online]. Available: <https://academic.oup.com/ije/article/50/6/1914/6182058>
- [13] E. Kričković, T. Lukić, and D. Jovanović-Popović, "Geographic Medical Overview of Noncommunicable Diseases (Cardiovascular Diseases and Diabetes) in the Territory of the AP Vojvodina (Northern Serbia)," *Healthcare*, vol. 11, no. 1, pp. 1–33, 2022. [Online]. Available: <https://www.mdpi.com/2227-9032/11/1/48>

- [14] T. B. Darikwa and S. O. Manda, "Spatial Co-Clustering of Cardiovascular Diseases and Select Risk Factors among Adults in South Africa," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, pp. 1–16, 2020. [Online]. Available: <https://www.mdpi.com/1660-4601/17/10/3583>
- [15] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning," *Applied Sciences*, vol. 13, no. 3, pp. 1–15, 2023. [Online]. Available: <https://www.mdpi.com/2076-3417/13/3/1509>
- [16] L. Zhang, G. Yang, and X. Li, "Mining sequential patterns of PM2.5 pollution between 338 cities in China," *Journal of Environmental Management*, vol. 262, no. March, pp. 1–8, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0301479720302760>
- [17] D. Majcherek, M. A. Weresa, and C. Ciecierski, "A Cluster Analysis of Risk Factors for Cancer across EU Countries: Health Policy Recommendations for Prevention," *International Journal of Environmental Research and Public Health*, vol. 18, no. 15, pp. 1–14, 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/15/8142>
- [18] M. A. Emon, A. Heinson, P. Wu, D. Domingo-Fernández, M. Sood, H. Vrooman, J.-C. Corvol, P. Scordis, M. Hofmann-Apitius, and H. Fröhlich, "Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020. [Online]. Available: <https://www.nature.com/articles/s41598-020-76200-4>
- [19] J. Prakash, V. Wang, R. E. Quinn, and C. S. Mitchell, "Unsupervised Machine Learning to Identify Separable Clinical Alzheimer's Disease Sub-Populations," *Brain Sciences*, vol. 11, no. 8, pp. 1–21, 2021. [Online]. Available: <https://www.mdpi.com/2076-3425/11/8/977>
- [20] S. Bhattacharjee, Y.-B. Hwang, R. I. Sumon, H. Rahman, D.-W. Hyeon, D. Moon, K. S. Carole, H.-C. Kim, and H.-K. Choi, "Cluster Analysis: Unsupervised Classification for Identifying Benign and Malignant Tumors on Whole Slide Image of Prostate Cancer," in *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*. IEEE, 2022, pp. 1–5. [Online]. Available: <https://ieeexplore.ieee.org/document/10052952/>
- [21] Y. Jiang, Z.-G. Yang, J. Wang, R. Shi, P.-L. Han, W.-L. Qian, W.-F. Yan, and Y. Li, "Unsupervised machine learning based on clinical factors for the detection of coronary artery atherosclerosis in type 2 diabetes mellitus," *Cardiovascular Diabetology*, vol. 21, no. 1, pp. 1–10, 2022. [Online]. Available: <https://cardiab.biomedcentral.com/articles/10.1186/s12933-022-01700-8>
- [22] G. Sarveswaran, V. Kulothungan, and P. Mathur, "Clustering of noncommunicable disease risk factors among adults (1869 years) in rural population, South-India," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1005–1014, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1871402120301624>
- [23] S. V. Rocha, S. C. de Oliveira, H. L. R. Munaro, C. F. R. Squarcini, B. M. P. Ferreira, F. de Oliveira Mendonça, and C. A. dos Santos, "Cluster analysis of risk factors for chronic non-communicable diseases in elderly Brazilians: population-based cross-sectional studies in a rural town," *Research, Society and Development*, vol. 10, no. 17, pp. 1–10, 2021. [Online]. Available: <https://rsdjournal.org/index.php/rsd/article/view/24202>
- [24] R. Uddin, E.-Y. Lee, S. R. Khan, M. S. Tremblay, and A. Khan, "Clustering of lifestyle risk factors for non-communicable diseases in 304,779 adolescents from 89 countries: A global perspective," *Preventive Medicine*, vol. 131, no. December, pp. 1–8, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0091743519304384>
- [25] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 337–350, 2022. [Online]. Available: <https://journal.universitatumigora.ac.id/index.php/matrik/article/view/1411>
- [26] H. Hairani, D. Susilowati, I. P. Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 275–282, 2022. [Online]. Available: <https://journal.universitatumigora.ac.id/index.php/matrik/article/view/1542>

- [27] S. Annas, B. Poerwanto, S. Sapriani, and M. F. S, "Implementation of K-Means Clustering on Poverty Indicators in Indonesia," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 257–266, 2022. [Online]. Available: <https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/1289>
- [28] Y. Zhao and X. Zhou, "K-means Clustering Algorithm and Its Improvement Research," *Journal of Physics: Conference Series*, vol. 1873, no. 1, pp. 1–5, 2021. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1873/1/012074>
- [29] M. Darwis, L. H. Hasibuan, M. Firmansyah, N. Ahady, and R. Tiaharyadini, "Implementation of K-Means clustering algorithm in mapping the groups of graduated or dropped-out students in the Management Department of the National University," *JISA(Jurnal Informatika dan Sains)*, vol. 4, no. 1, pp. 1–9, 2021. [Online]. Available: <http://trilogi.ac.id/journal/ks/index.php/JISA/article/view/848>
- [30] A. R. Danurisa and J. Heikal, "Customer Clustering Using the K-Means Clustering Algorithm in the Top 5 Online Marketplaces in Indonesia," *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, vol. 5, no. 3, 2022.
- [31] A. Chaerudin, D. T. Murdiansyah, and M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Indonesia Journal on Computing (Indo-JC)*, vol. 6, no. 1, pp. 53–60, 2021.
- [32] A. Dudek, *Silhouette Index as Clustering Evaluation Tool*, 2020, pp. 19–33. [Online]. Available: http://link.springer.com/10.1007/978-3-030-52348-0_{_}2
- [33] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/6/759>
- [34] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Techno.Com*, vol. 20, no. 2, pp. 186–197, 2021. [Online]. Available: <http://publikasi.dinus.ac.id/index.php/technoc/article/view/4556>
- [35] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 6, no. 2, pp. 48–54, 2021. [Online]. Available: <https://ojs.unanda.ac.id/index.php/jiit/article/view/659>
- [36] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 174–182, 2023. [Online]. Available: <https://online-journals.org/index.php/i-joe/article/view/37059>
- [37] T. Li, Y. Ma, and T. Endoh, "Normalization-Based Validity Index of Adaptive K-Means Clustering for Multi-Solution Application," *IEEE Access*, vol. 8, pp. 9403–9419, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/8952702/>
- [38] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *Journal of Physics: Conference Series*, vol. 1566, no. 1, pp. 1–7, 2020. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1566/1/012112>
- [39] H. A. Ahmed, P. J. M. Ali, A. K. Faeq, and S. M. Abdullah, "An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method," *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 2, pp. 29–37, 2022. [Online]. Available: <https://aro.koyauniversity.org/index.php/aro/article/view/970>
- [40] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, vol. 10, no. 11, p. 1942, 2022. [Online]. Available: <https://www.mdpi.com/2227-7390/10/11/1942>



MATRIK

Jurnal : Manajemen, Teknik Informatika dan Rekayasa Komputer
e-issn 2476-9843 | p-issn 1858-4144

tbaimunandar

[HOME](#) [EDITORIAL TEAM](#) [PEER REVIEWERS](#) [CURRENT](#) [ARCHIVES](#) [ANNOUNCEMENT](#)

[SEARCH](#)

[HOME](#) / [ARCHIVES](#) / Vol 23 No 2 (2024)



We are pleased to announce that Volume 23, Issue 2 of MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, has recently been published. This latest issue features contributions from 66 authors from 4 different countries (**Indonesia, Malaysia, Uzbekistan, and Iraq**). There are 4 overseas, 7 public, and 15 private institutions, namely **Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, Northern Technical University, Mosul, Iraq, Samarkand Institute of Economics and Service, Samarkand, Uzbekistan, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia, Politeknik Angkatan Darat, Malang, Indonesia, Universitas Malikussaleh, Aceh, Indonesia, Universitas Negeri Malang, Malang, Indonesia, Universitas Islam Negeri Sunan Ampel, Surabaya, Indonesia, PLN Nusantara Power, Surabaya, Indonesia, Universitas Negeri Semarang, Semarang, Indonesia, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia, Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia, Universitas Teknologi Yogyakarta, Yogyakarta,**

Indonesia, Universitas Bumigora, Mataram, Indonesia, Universitas Dian Nuswantoro, Semarang, Indonesia, Telkom University, Bandung, Indonesia, Universitas Muhammadiyah Purworejo, Purworejo, Indonesia, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia, Universitas Nusa Putra, Sukabumi, Indonesia, Sampoerna University, Jakarta, Indonesia, Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia, Institut Teknologi dan Bisnis Asia Malang, Malang, Indonesia, Institut Keuangan-Perbankan dan Informatika Asia Perbanas, Jakarta, Indonesia, Universitas Kristen Satya Wacana University, Salatiga, Indonesia, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, and Universitas Muhammadiyah Malang, Malang, Indonesia.

DOI: <https://doi.org/10.30812/matrik.v23i2>

PUBLISHED: 2024-03-30

ARTICLES

Hostage Liberation Operations using Wheeled Robots Based on LIDAR (Light Detection and Ranging) Sensors

Kasiyanto Kasiyanto, Aripriharta Aripriharta, Dekki Widiatmoko, Dodo Irmanto, Muhammad Cahyo Bagaskoro

243-258

[PDF](#)

Abstract Viewed : 44 | pdf Download : 60 | DOI <https://doi.org/10.30812/matrik.v23i2.3493>



Comparing Long Short-Term Memory and Random Forest Accuracy for Bitcoin Price Forecasting

Munirul Ula, Veri Ilhadi, Zailani Mohamed Sidek

259-272

[PDF](#)

Abstract Viewed : 48 | pdf Download : 58 | DOI <https://doi.org/10.30812/matrik.v23i2.3267>

Quick Menu

[History](#)

[Focus and Scope](#)

[Plagiarism Policy](#)

[Open Access Policy](#)

[Contacts](#)

[Copyright Notice](#)

[Author Guidelines](#)

[Publication Charge](#)

[Publication Ethics](#)

[Peer Review Process](#)

[Visitor Matrik](#)

[Abstracting and Indexing](#)

[Statement of the Letter](#)

accredited sinta 2

Tools

[Article LATEX Template](#)

[Article Template](#)

MENDELEY

Plagiarism Checker

Crossref Similarity Check
Powered by iThenticate

grammarly



Power Efficiency using Bank Capacitor Regulator on Field Service Shoes with Fast Charge Method

Dekki Widiatmoko, Aripriharta Aripriharta, Kasiyanto Kasiyanto, Dodo Irmanto, Muchamad Wahyu Prasetyo

273-284

[PDF](#)

Abstract Viewed : 45 | pdf Download : 36 | DOI <https://doi.org/10.30812/matrik.v23i2.3494>



Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Tb Ai Munandar, Ajif Yunizar Yusuf Pratama

285-296

[PDF](#)

Abstract Viewed : 52 | pdf Download : 39 | DOI <https://doi.org/10.30812/matrik.v23i2.3352>



Learning Accuracy with Particle Swarm Optimization for Music Genre Classification Using Recurrent Neural Networks

Muhammad Rizki, Arief Hermawan, Donny Avianto

297-308

[PDF](#)

Abstract Viewed : 49 | pdf Download : 44 | DOI <https://doi.org/10.30812/matrik.v23i2.3037>



Modeling the Farmer Exchange Rate in Indonesia Using the Vector Error Correction Model Method

Yuniar Farida, Afanin Hamidah, Silvia Kartika Sari, Lutfi Hakim

309-322

[PDF](#)

Abstract Viewed : 66 | pdf Download : 21 | DOI <https://doi.org/10.30812/matrik.v23i2.3407>



Optimizing Treatment of Herbal Plant Using SOPHERBAL Android Application Forward Chaining Method

Muhamad Azwar, Eka Nurul Qomaliyah, Nurul Indriani

323-332

[PDF](#)

Abstract Viewed : 24 | pdf Download : 20 | DOI <https://doi.org/10.30812/matrik.v23i2.3371>



Abstracting and Indexing



Scopus Citedness



Contact Support



Supervised



KEYWORDS



Comparison of DenseNet-121 and MobileNet for Coral Reef Classification

Heru Pramono Hadi, Eko Hari Rachmawanto, Rabei Raad Ali

333-342

 [PDF](#)Abstract Viewed : 41 | pdf Download : 25 | DOI <https://doi.org/10.30812/matrik.v23i2.3683>**Improving Performance Convolutional Neural Networks Using Modified Pooling Function**

Achmad Lukman, Wahyu Tjahjo Saputro, Erni Seniwati

343-352

 [PDF](#)Abstract Viewed : 41 | pdf Download : 22 | DOI <https://doi.org/10.30812/matrik.v23i2.3763>**Enhancing Predictive Models: An In-depth Analysis of Feature Selection Techniques Coupled with Boosting Algorithms**

Neny Sulistianingsih, Galih Hendro Martono

353-364

 [PDF](#)Abstract Viewed : 34 | pdf Download : 19 | DOI <https://doi.org/10.30812/matrik.v23i2.3788>**Detecting Hidden Illegal Online Gambling on .go.id Domains Using Web Scraping Algorithms**

Muchlis Nurseno, Umar Aditiawarman, Haris Al Qodri Maarif, Teddy Mantoro

365-378

 [PDF](#)Abstract Viewed : 139 | pdf Download : 78 | DOI <https://doi.org/10.30812/matrik.v23i2.3824>**Educational Data Mining: Multiple Choice Question Classification in Vocational School**

Sucipto Sucipto, Didik Dwi Prasetya, Triyanna Widiyaningtyas

379-388

 [PDF](#)Abstract Viewed : 30 | pdf Download : 24 | DOI <https://doi.org/10.30812/matrik.v23i2.3499>**Unsafe Conditions Identification Using Social Networks in Power Plant Safety Reports**

Annisa'ul Mubarakah, Rita Ambarwati, Dedy Dedy, Mashhura Toirxonovna Alimova

389-404

 [PDF](#)Abstract Viewed : 41 | pdf Download : 19 | DOI <https://doi.org/10.30812/matrik.v23i2.3883>

Google Scholar Citation : Journal MATRIK

	All	Since 201
Citations	3130	3079
h-index	24	24
i10-index	93	92

Journal MATRIK



Enhancing Accuracy in Stock Price Prediction: The Power of Optimization Algorithms

Vivi Aida Fitria, Lilis Widayanti

405-418

[PDF](#)

Abstract Viewed : 26 | pdf Download : 20 | DOI <https://doi.org/10.30812/matrik.v23i2.3785>



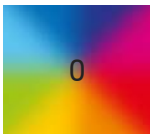
Optimization of SVM and Gradient Boosting Models Using GridSearchCV in Detecting Fake Job Postings

Rofik Rofik, Roshan Aland Hakim, Jumanto Unjung, Budi Prasetyo, Much Aziz Muslim

419-430

[PDF](#)

Abstract Viewed : 56 | pdf Download : 34 | DOI <https://doi.org/10.30812/matrik.v23i2.3566>



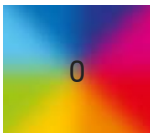
Forecasting the Poverty Rates using Holt's Exponential Smoothing

Riza Prapascatama Agusdin, Sylvert Prian Tahalea, Vynska Amalia Permadi

431-440

[PDF](#)

Abstract Viewed : 33 | pdf Download : 13 | DOI <https://doi.org/10.30812/matrik.v23i2.2672>



Sentiment Analysis of e-Government Service Using the Naive Bayes Algorithm

Winy purbaratri, Hindriyanto Dwi Purnomo, Danny Manongga, Iwan Setyawan, Hendry Hendry

441-452

[PDF](#)

Abstract Viewed : 36 | pdf Download : 22 | DOI <https://doi.org/10.30812/matrik.v23i2.3272>



DenseNet Architecture for Efficient and Accurate Recognition of Javanese Script Hanacaraka Character

Egi Dio Bagus Sudewo, Muhammad Kunta Biddinika, Abdul Fadlil

453-464

[PDF](#)

Abstract Viewed : 28 | pdf Download : 19 | DOI <https://doi.org/10.30812/matrik.v23i2.3855>



Implementation of Neural Machine Translation in Translating from Indonesian to Sasak Language

Helna Wardhana, I Made Yadi Dharma, Khairan Marzuki, Ibjan Syarif Hidayatullah

465-476

 [PDF](#)

Abstract Viewed : 47 | pdf Download : 24 | DOI <https://doi.org/10.30812/matrik.v23i2.3465>



Stroke Prediction with Enhanced Gradient Boosting Classifier and Strategic Hyperparameter

Dela Ananda Setyarini, Agnes Ayu Maharani Dyah Gayatri, Christian Sri Kusuma Aditya, Didih Rizki Chandranegara

477-490

 [PDF](#)

Abstract Viewed : 96 | pdf Download : 35 | DOI <https://doi.org/10.30812/matrik.v23i2.3555>



Abstracting and Indexing



Platform
workflow b
OJS / PK

MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer

Published by: LPPM Universitas Bumigora

Address: Jl. Ismail Marzuki-Cilinaya-Cakranegara-Mataram 83127, Indonesia

Phone: +6285-933-083-240

e-mail: matrik@universitasbumigora.ac.id

[View Matrik Stats](#)



Journal : Manajemen, Teknik Informatika dan Rekayasa Komputer
e-issn 2476-9843 | p-issn 1858-4144

tbaimunandar

[HOME](#) [EDITORIAL TEAM](#) [PEER REVIEWERS](#) [CURRENT](#) [ARCHIVES](#) [ANNOUNCEMENT](#)

[SEARCH](#)

Editorial Team

Editor in Chief

Hairani, Universitas Bumigora, Indonesia

[Google Scholar](#) | SINTA ID : [6100285](#) | SCOPUS ID: [57222080524](#) | Orchid ID: [0000-0002-6756-5896](#) | [Curriculum Vitae](#)

Managing Editor

Khairan Marzuki, Universitas Bumigora, Indonesia

[Google Scholar](#) | GARUDA ID: [1568552](#) | SINTA ID: [6680270](#) | SCOPUS: [57280266600](#)

Editorial Board

Regional Editor for Asia and Pasific

Anthony Anggrawan, Universitas Bumigora, Indonesia

[Google Scholar](#) | SINAT ID: [6114236](#) | SCOPUS ID: [57194213003](#)

Lalu Ganda Rady Putra, Universitas Bumigora, Indonesia

[Google Scholar](#) | SINTA ID: [6707978](#) | SCOPUS ID: [57540061000](#)

Irwan Oyong, Universitas AMIKOM Yogyakarta, Indonesia

[Google Scholar](#) | ORCID ID: [0000-0001-6365-0826](#) | SCOPUS ID: [57200211469](#) | SINTA ID: [6721323](#)

Choerun Asnawi, Universitas Jendral Ahmat Yani Yogyakarta, Indonesia

[Google Scholar](#) | SINTA ID : [5981941](#) | SCOPUS ID:

Didi Haryono, Institut Bisnis dan Keuangan Nitro, Makassar, Indonesia

[Google Scholar](#) | ORCID ID: [0000-0001-8957-2336](#) | SINTA ID: [5985592](#) | SCOPUS ID: [57216611534](#)

Andi Sunyoto, Universitas AMIKOM, Yogyakarta, Indonesia

[Google Scholar](#) | SCOPUS ID: [57202197381](#) | SINTA ID: [6008849](#) | WoS Clarivate : [501895](#)

Hengki Tamando Sihotang, Institute of Computer Science, Medan, Indonesia

[Google Scholar](#) | SINTA ID: [6154823](#) | SCOPUS ID: [57211266124](#)

Erick Fernando, Universitas Multimedia Nusantara, Serpong, Indonesia

[Google Scholar](#) | ORCID ID: [0000-0003-2428-0484](#) | SINTA ID: [207171](#) | SCOPUS ID: [57189355900](#) | WoS: [1840902](#)

Che Ku Nuraini, Universiti Teknikal Malaysia Melaka, Malaysia

[Google Scholar](#) | SCOPUS ID: [35759291800](#) | ORCID ID: [9823-5315](#)

Siti Soraya, Universitas Bumigora, Mataram, Indonesia

[Google Scholar](#) | SINTA ID: [5994189](#) | SCOPUS ID: [57209226434](#)

Sirojul Hadi, Universitas Bumigora, Mataram, Indonesia

[Google Scholar](#) | SINTA ID: [6708110](#) | SCOPUS ID: [57207596263](#)

Muhammad Zulfikri, Universitas Bumigora, Indonesia

[Google Scholar](#) | SINAT ID : [6719678](#) | SCOPUS ID : [57222187881](#) | Publon : [4174051](#)

Quick Menu

[History](#)

[Focus and Scope](#)

[Plagiarism Policy](#)

[Open Access Policy](#)

[Contacts](#)

[Copyright Notice](#)

[Author Guidelines](#)

[Publication Charge](#)

[Publication Ethics](#)

[Peer Review Process](#)

[Visitor Matrik](#)

[Abstracting and Indexing](#)

[Statement of the Letter](#)

[accredited sinta 2](#)

Tools

[Article LATEX Template](#)

[Article Template](#)

MENDELEY

Plagiarism Checker

Crossref Similarity Check
Powered by iThenticate

grammarly

Regional Editor for Europe

Gerhard-Wilhelm Weber, Poznan University of Technology, Poland
[Google Scholar](#) | ORCID ID: [0000-0003-0849-7771](#) | SCOPUS ID: [55634220900](#)

Abdellah Salhi, University of Essex, United Kingdom
[Google Scholar](#) | SCOPUS ID: [8857392100](#)

Regional Editor West Asia

Anton Abdulbasah Kamil, Nisantasi University, Turkey
 ORCID ID: [0000-0001-5410-812X](#) | SCOPUS ID: [24481107300](#)

Language Editor

Diah Supatmiwati, SINTA ID: [5994570](#) | SCOPUS ID: [57211635097](#)

Zainudin Abdussamad, SINTA ID: [6163190](#)

Technical Support

- Sulistianti
- Dinda Lesta

Abstracting and Indexing



Scopus Citedness



Contact Support



Supervised



KEYWORDS



Google Scholar Citation : Journal MATRIK

	All	Since 201
Citations	3130	3079
h-index	24	24
i10-index	93	92

Journal MATRIK

Abstracting and Indexing

Platform
workflow b
OJS / PK

MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer

Published by: LPPM Universitas Bumigora

Address: Jl. Ismail Marzuki-Cilinaya-Cakranegara-Mataram 83127, Indonesia

Phone: +6285-933-083-240

e-mail: matrik@universitasbumigora.ac.id

[View Matrik Stats](#)



MATRIK

Jurnal : Manajemen, Teknik Informatika dan Rekayasa Komputer
e-issn 2476-9843 | p-issn 1858-4144

tbaimunandar

[HOME](#) [EDITORIAL TEAM](#) [PEER REVIEWERS](#) [CURRENT](#) [ARCHIVES](#) [ANNOUNCEMENT](#)

[SEARCH](#)

Reviewers

Dorien J Detombe, International Research Society on Methodology of Societal Complexity, Amsterdam, Netherlands.

Orcid Id: [0000-0001-9031-2451](#) | Scopus Id: [6603137867](#)

Milagros Baldemor, Mindanao State University, Philippines

Orcid Id: [0000-0002-2095-8959](#) | [Profile](#)

Mohamed Arezki Mellal, M'Hamed Bougara University, Algeria

[Google Scholar](#) | Scopus Id: | Orcid Id: | [Profile](#)

Teddy Mantoro, Universitas Nusa Putra, Sukabumi, Indonesia

[Google Scholar](#) | Sopos Id: [22735122000](#) | Sinta Id: [119641](#)

Media Anugerah Ayu, Universitas Sampoerna, Indonesia

[Google Scholar](#) | Scopus Id: [35589381300](#) | Sinta Id: [5999218](#)

Arief Setyanto, Universitas AMIKOM Yogyakarta, Indonesia

[Google Scholar](#) | Scopus Id: [55523640300](#) | Sinta Id: [6024270](#)

Zulfian Azmi, STMIK Triguna Dharma Medan, Indonesia

[Google Scholar](#) | Scopus Id: [57200726694](#) | Sinta Id: [6010159](#)

Dwiza Riana, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Indonesia

[Google Scholar](#) | Scopus Id: [56242366200](#) | Sinta Id: [125329](#)

Suwanto Raharjo, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia

[Google Scholar](#) | Scopus Id: [56192446500](#) | Sinta Id: [152252](#)

Agus Perdana Windarto, STIKOM Tunas Bangsa Pematangsiantar Medan Sumatra Utara, Indonesia

[Google Scholar](#) | Scopus Id: [57197780326](#) | Sinta Id: [257474](#)

Nur Uddin, Pembangunan Jaya, Indonesia

[Google Scholar](#) | Scopus Id: [54948199600](#) | Sinta Id: [6031360](#)

Aji Supriyanto, Universitas Stikubank Semarang, Indonesia

[Google Scholar](#) | Scopus Id: [57208599527](#) | Sinta Id: [6026244](#)

Dadang Priyanto, Universitas Bumigora, Indonesia

[Google Scholar](#) | Scopus Id: [57211320355](#) | Sinta Id: [5978450](#)

Deny Jollyta, Institut Bisnis dan Teknologi Pelita Indonesia, Indonesia

[Google Scholar](#) | Scopus Id: [57200089307](#) | Sinta Id: [6096202](#)

Widodo, Universitas Negeri Jakarta, Indonesia

[Google Scholar](#) | Scopus Id: [56592813500](#) | WoS: [2205906](#) | Sinta Id: [5976955](#)

Erick Fernando, Universitas Multimedia Nusantara, Tangerang, Indonesia

[Google Scholar](#) | Scopus Id: [57189355900](#) | WoS: [1840902](#) | Sinta Id: [207171](#)

Quick Menu

[History](#)

[Focus and Scope](#)

[Plagiarism Policy](#)

[Open Access Policy](#)

[Contacts](#)

[Copyright Notice](#)

[Author Guidelines](#)

[Publication Charge](#)

[Publication Ethics](#)

[Peer Review Process](#)

[Visitor Matrik](#)

[Abstracting and Indexing](#)

[Statement of the Letter](#)

accredited sinta 2

Tools

 [Article LATEX Template](#)

 [Article Template](#)

 **MENDELEY**

 **Plagiarism Checker**

 **Crossref Similarity Check**
Powered by iThenticate

 **grammarly**

Google Scholar Citation : Journal MATRIK

	All	Since 201
Citations	3130	3079
h-index	24	24
i10-index	93	92

Journal MATRIK

Abstracting and Indexing

Platform
workflow b
OJS / PK



MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer

Published by: LPPM Universitas Bumigora

Address: Jl. Ismail Marzuki-Cilinaya-Cakranegara-Mataram 83127, Indonesia

Phone: +6285-933-083-240

e-mail: matrik@universitasbumigora.ac.id

[View Matrik Stats](#)

Author's Comment for Reviewer

Title : Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

19 Dec 2023

From editor :

No.	Reviewer's Comment	Author's Comment
1	focused on explaining the differences or uniqueness of this research with several studies in the state-of-the-art section, so that the novelty is clearly visible. Beginning with the words "The difference between this research and the previous one is..."	<ul style="list-style-type: none">• We have added two sentences to show the differences between current and previous research.• Mark as yellow• See page 2
2	Add the research objectives explicitly and the contribution of this research to the development of science or its benefits.	We've already add the research objective and the benefit (contribution). <ul style="list-style-type: none">• Mark as yellow on page 3
3	Move Figure 1 to the methods section	<ul style="list-style-type: none">• We've moved Figure 1 as your sugesstion.• See Page 3
4	1. findings of this research explicitly, starting with the words "The findings of this research are..." 2. There needs to be justification from previous research that strengthens or contradicts the results of this research which begins with the words "The results of this research are in line with or supported by..."	<ul style="list-style-type: none">• We've adjusted the paragraph based on your sugesstion and adding some reference to justified the finding.• Mark as yellow on page 7-8
5	References marked yellow should be replaced with journals from the last 5 years, from 2019 - 2023	<ul style="list-style-type: none">• We already replace the marked yellow reference and add 4 references (ref 1-4) and also reference 6. All references already replaced with journal from the last 5 years (2019-2023) as your suggestion• See references section

Author's Comment for Reviewer

Title : Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Reviewer #1 :

No.	Reviewer's Comment	Author's Comment
1	What is the purpose of this study? The abstract's final results do not match with the conclusions.	<ul style="list-style-type: none"> We have stated the purpose of the research in the abstract in the second sentence. We have adjusted the final results of the research, both the abstract and conclusions.
2	Many typos were found	We have corrected all typos related to classification terms to suit the research objectives
3	Clearly state the research purpose in the introduction	We have stated the purpose of the research in the second paragraph of the introduction.
4	The purpose of this study is to compile data on noncommunicable diseases. The author should explain the criteria for grouping. The author must ensure that the process in question is clustering, not classification, because the final results of the two processes are different.	<ul style="list-style-type: none"> We have stated the clustering criteria in the research methodology section. There are twelve criteria that are the basis for grouping regions based on types of non-communicable diseases, namely hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, COPD, and congenital deafness. Apart from that, we also stated in the research results. We have also confirmed and checked every term typo. Which should refer to clustering.
5	The research steps necessary to reach the aim are not yet visible. It is preferable to create research stages in chart design.	The research stages have been created in the form of a flowchart.
6	In result section :	
	<ul style="list-style-type: none"> Why is the conversation focusing on outlier data? According to the title or abstract, outliers are not the main cause 	In the abstract section, we have stated the outliers that the reviewer is concerned about in the results and discussion sections.

	of the problem. The discussion here is not acceptable.	
	<ul style="list-style-type: none"> • What are the criteria for grouping non-communicable diseases? 	<ul style="list-style-type: none"> • We have stated the clustering criteria in the research methodology section. There are twelve criteria that are the basis for grouping regions based on types of non-communicable diseases, namely hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, COPD, and congenital deafness. Apart from that, we also stated in the research results. • We have also confirmed and checked every term typo. Which should refer to clustering.
	<ul style="list-style-type: none"> • The author must explain the occurrence of outliers 	We have stated this in the results and discussion.
	<ul style="list-style-type: none"> • The author must explain how outliers and the silhouette index are related. 	We have stated this in the results and discussion.
7	Use references from the last 5 years (2019-2023)	We have corrected the bibliography according to suggestions.

Reviewer #2 :

No.	Reviewer's Comment	Author's Comment
1	<p>The problem statement is not clearly defined in the paper. What are the limitations of the existing methods?</p> <p>The authors should mention their contribution and novelty compared to the previous work in the introduction section in a paragraph (or more).</p>	<ul style="list-style-type: none"> • We have stated the problems statement by rewriting the introductory part of the first paragraph. Meanwhile, the contribution we offer is more of an alternative regional grouping based on unsupervised learning by extracting hidden patterns from twelve types of non-communicable disease data.
2	Results were not explained properly. Please write the Findings of the Study and immediately after the findings, please write the implications of the findings.	We have restated the research findings and added some information by extending the results of previous research.

3	Lack of comparison results with existing methods. The authors should compare their proposed method with existing ones and explain its advantageous and disadvantageous.	As stated in the research problem statement in the introduction, we did not compare it with other cluster methods. However, we compared the cluster results based on the two scenarios that we determined. The first cluster is based on the results of outlier analysis, the second is without outliers.
---	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Tb Ai Munandar, Ajif Yunizar Yusuf Pratama
Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia

Article Info

Article history:

Received September 05, 2023
Revised November 13, 2023
Accepted January 05, 2024

Keywords:

Clustering
k-Means
Non-Communicable Diseases
Regional Clustering
Silhouette Index

ABSTRACT

Noncommunicable diseases (NCDs) have become a global threat to public health, necessitating a comprehensive understanding of their geographic and epidemiological distribution to devise appropriate interventions. This study aims to cluster Banten Province areas based on NCDS profiles using the unsupervised learning technique. The method used in this study is the k-means algorithm for grouping types of non-communicable diseases based on region. The processing and normalisation of NCDS prevalence data from various health sources preceded cluster analysis using the k-means clustering algorithm. This research is categorised into two scenarios: the first involves clustering data obtained from outlier analysis, while the second excludes any outliers. The objective is to observe disparities in regional clustering outcomes by categorising non-communicable diseases according to these two scenarios. The silhouette index is used to determine the validity of cluster results. These findings are analysed to determine the geographic and socioeconomic patterns associated with each cluster's NCDS profile. Based on the mean silhouette index value of 0.812, the results indicate that the sum of $k = 2$ in the k-means algorithm is the optimal cluster result. Five non-communicable diseases, namely diabetes, hypertension, obesity, stroke, and cataracts, necessitate significant focus in the first cluster (C1), where 202 regions were grouped. Six regions belong to the second cluster (C2), which includes areas that are not only susceptible to the five non-communicable diseases in cluster C1 but also to breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

Copyright ©2022 The Authors.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Tb Ai Munandar, +6281384512710
Faculty of Computer Science, Informatics Department,
Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia.
Email: tbaimunandar@gmail.com

How to Cite:

T. A. Munandar and A. Y. Y. Pratama, "Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 2, pp. 285-296, March, 2024. This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

1. INTRODUCTION

Non-communicable diseases (NCDs) comprise a collection of chronic health conditions that do not spread via direct human-to-human contact and are the leading cause of cancer-related deaths worldwide. This grouping contains a variety of illnesses, including chronic respiratory conditions, diabetes, cardiovascular disease, and cancer, which collectively impose a significant burden on global health. The gravity of non-communicable diseases (NCDs) is emphasised by the World Health Organisation (WHO), which estimates that these conditions are responsible for around 41 million deaths per year, or 71% of all casualties worldwide [1–4]. The development of non-communicable diseases (NCDs) is strongly associated with detrimental lifestyle decisions, which include unequal dietary patterns, a lack of physical activity, and the use of tobacco and alcohol. Due to urbanisation, evolving behaviours, and rapid population growth, the Indonesian province of Banten faces substantial challenges associated with non-communicable diseases (NCDs). In 2021, Banten Provincial Health Service released data that indicates Banten Province has experienced an ongoing and consistent increase in the prevalence of diabetes, hypertension, and obesity. The substantial increase in the prevalence of non-communicable diseases presents a tough obstacle for the regional health infrastructure, significantly affecting the community's overall standard of living [5, 6]. To address their complexity, an in-depth examination of the patterns and distribution of non-communicable diseases (NCDs) in Banten is necessary. The current regional grouping system predicated on NCD categories may not be ideal when developing targeted interventions.

Applying unsupervised learning techniques, specifically the cluster method, to divide Banten Province into groups based on different types of non-communicable diseases (NCDs) is the point of this research. Clustering presents a more sophisticated methodology than conventional grouping techniques, enabling the detection of inherent groupings within the data that do not require predetermined labelling. This study seeks a more precise comprehension of the distribution of non-communicable diseases (NCDs) throughout Banten Province by delineating regional clusters according to different NCD categories. For informing health policies and directing medical interventions, precise and comprehensive data on the prevalence and distribution of non-communicable diseases (NCDs) in various regions of Banten Province are indispensable. By allowing health policymakers to customise more efficacious prevention and intervention strategies, the regional clustering inferred from this study has the potential to yield significant insights regarding the spatial distribution of non-communicable disease prevalence. The resulting regional groups will provide a solid foundation for developing targeted non-communicable disease (NCD) prevention programmes, letting policymakers and medical professionals pay close attention to the specific needs of each cluster. Using unsupervised learning techniques, particularly clustering, to illustrate the intricate picture of NCD prevalence, this study's primary objective is to aid in the improvement of public health strategies in Banten Province; by employing this novel methodology, the research endeavours to establish a strong groundwork for decision-making grounded in evidence, thereby promoting a more sophisticated and focused approach to addressing the complex issues presented by non-communicable diseases in the area.

To face the challenges of NCDS in the province of Banten, a comprehensive and integrated approach to grouping areas based on the nature of NCDS is required. Currently, regional groupings are typically determined by administration or geographic location without considering each region's unique health profile. We can identify patterns and relationships in Banten Province NCDS data using unsupervised learning techniques, such as the clustering method. This phase is essential for designing more targeted interventions and enhancing health services in regions affected by NCDS. Moreover, research conducted by [7] on the epidemiology of hypertension in India suggests that accelerated urbanisation may contribute to the rising prevalence of hypertension in urban areas. The prevalence of obesity tends to be higher in urban areas than rural areas, highlighting the significance of taking regional differences into account when managing NCDs. These results indicate that the social and economic changes that result from urbanisation can affect disease patterns in the region, and it is not inconceivable that this could occur in Banten Province. Noncommunicable diseases (NCDs) are a major public health concern because they contribute to the world's elevated mortality rate and disease burden. Numerous studies have been conducted in recent years to understand the risk factors and transmission patterns of NCDS and to identify more effective prevention strategies. A study by Chen provides the most recent data regarding the prevalence and risk factors of type 2 diabetes in various populations. The findings of this study shed light on the association between diet and physical activity and the risk of type 2 diabetes. Several studies have investigated the impact of diet in reducing the risk of cardiovascular disease as part of efforts to prevent NCDs. For instance, research found a correlation between adult fast-food consumption and increased insulin resistance. Additionally, recent research has emphasised the significance of optimal sleep patterns in reducing the risk of NCDs. A study by [8] discovered that adult sleep deprivation increases the risk of adiposity. To comprehend the effect of physical activity on the prevention of NCDs, Daskalopoulou (2021) conducted a meta-analysis which revealed that higher levels of physical activity are associated with lower risks for certain NCDs. Recent research has also focused on the environment's role in NCDS. The composition of the intestinal microbiota is associated with the development of cardiovascular disease, according to [9, 10]. In terms of NCDS prevention, understanding the function of the environment is crucial. Zimmet, in 2021, investigated the global repercussions of the diabetes epidemic and other NCDs. In addition to focusing on NCDS in the adult population, researchers have also studied the

disorder in children. At the same time, Samavat 2021 examined the association between a child's adult diet and their future risk of breast cancer [11, 12]. However, as mentioned earlier, a portion of the research places greater emphasis on discerning various intervention methods and enhancing healthcare provisions by utilising assessments of individuals afflicted with non-communicable diseases. The topic of regional-specific approaches to healthcare intervention and management has not yet been addressed. The issue of non-communicable illness spreading often remains unresolved to some extent due to a lack of information regarding treatment priorities specific to different regions. It is imperative to adopt a regional cluster-based approach to enhance health services and interventions in the future.

The distinction between the present and prior studies categorises non-communicable diseases based on geographical regions. By implementing interventions and making enhancements, healthcare quality can be significantly improved. Collectively, the most recent research provides valuable insights into the effective management of NCDS. Through a greater understanding of risk factors and prevention strategies, it is anticipated that targeted preventive measures can be implemented to lessen the burden of non-communicable diseases (NCDs) and improve public health as a whole. Noncommunicable diseases (NCDs) are a significant global health burden, and it is essential to understand the patterns and patterns of dissemination of these diseases in a specific region for prevention and appropriate management. The unsupervised learning method has become a valuable instrument in analysing health data, including grouping regions by NCD type. Several recent studies have utilised unsupervised learning techniques, such as clustering and cluster analysis, to identify geographic and epidemiological patterns of NCDs in different regions. A study by [13] grouped regions based on the type of NCDS in a country using cluster analysis. This study reveals how to identify the most and least burdened counties. A related study by [14] and [15] utilised spatial clustering and unsupervised learning to categorise regions based on the pattern of cardiovascular disease distribution. The results of this study indicate certain health clusters that can be targeted by interventions to reduce the risk of cardiovascular disease. In addition, unsupervised learning techniques have been employed to determine the connection between air pollution patterns, the urban spread of respiratory diseases and health interventions. Research by [16] found that cluster patterns from air pollution data and the incidence of respiratory disease are frequently interconnected. Recent research has also investigated the use of machine learning, including unsupervised learning, for grouping regions based on other categories of diseases, such as cancer, benign and malignant tumours, diabetes, and neurodegenerative diseases [17–21]. To address the challenges posed by NCDS at the regional level, an unsupervised learning algorithm was used to identify patterns of interrelationships between specific NCDS in specific regions of a country [22]. Also, at the same time, some researchers grouped regions based on the level of exposure to certain NCDS risk factors using an unsupervised learning method [23, 24]. Overall, the application of unsupervised learning has created new opportunities to segment regions based on the type of NCDs and to gain a deeper understanding of disease transmission patterns; one of the most popular algorithms is k-means. Using this strategy, it is anticipated that prevention and intervention can be more effectively targeted based on the local community's health characteristics. The k-means algorithm is used for a variety of reasons. Apart from being frequently used in health research, it may also be used to segment promotional places in education, facilities, and teachers [25, 26] and group poverty indicators in a region [27]. This study aims to categorise non-communicable diseases based on geographical regions by analysing patient data obtained from public health institutions in Banten Province. This research is anticipated to significantly impact local government's ability to identify regional priorities for managing the development of non-communicable diseases. In addition, unsupervised learning methods, such as the k-means algorithm, can be utilised as an alternate strategy to analyse non-communicable disease data, making a valuable contribution to the health sector.

This paper is divided into four sections. The first section presents the introduction, which includes the problem's background, a literature review, research gaps, and systematic information about the article. The second section explains the study methodologies used, and the third section includes the research results and commentary. Meanwhile, the fourth portion is the conclusion, which contains the investigation findings.

2. RESEARCH METHOD

This study uses unsupervised learning as its research methodology to categorise regions within Banten Province according to the specific noncommunicable disease (NCD) type. Figure 1 provides a more detailed overview of the study stages.

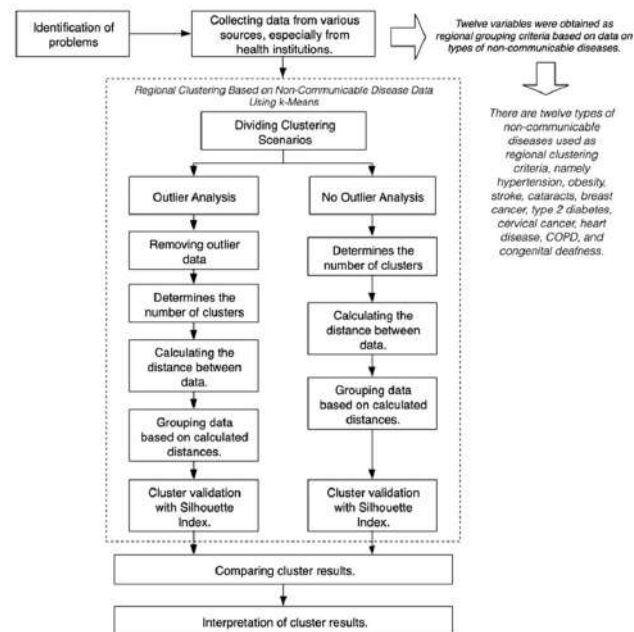


Figure 1. Research Stages

During the preliminary stage of the inquiry, data pertaining to the incidence of non-communicable diseases (NCDs) were gathered from several sources, encompassing public health institutions and hospitals situated within Banten Province. A total of 227 instances of regional information representing subdistricts were gathered due to the data-gathering process. The data is subsequently submitted to a preprocessing stage, wherein incomplete or fragmented data are repaired, and all variables are normalised to guarantee consistency in scale. Consequently, the strategy for categorising regions based on NCDs profiles involved using the k-means clustering algorithm, which falls under the umbrella of unsupervised learning techniques. Cluster analysis is conducted by determining the distance between the data points and the centroid of each cluster. This process involves combining regions that exhibit similar NCDs profiles into a cohesive cluster. The quality of the clusters was evaluated by performing validation of the cluster results using the silhouette index. The clustering criteria are determined by twelve specific non-communicable diseases, which include hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, COPD, and congenital deafness.

2.1. K-Means

The k-means algorithm is a commonly employed clustering technique that divides a dataset into separate groups according to their similarity. This is achieved through an iterative process of assigning data points to the nearest cluster centroid and adjusting the centroids to minimise the total sum of squared distances. The method's success in numerous domains can be attributed to its efficiency and simplicity despite the acknowledged drawbacks of being sensitive to initial centroid selection and prone to converging to local optima [28, 29]. The k-means algorithm is widely recognised for its prevalence in unsupervised learning. It is commonly employed in several domains, such as picture segmentation, customer segmentation, and anomaly detection. Although this method is extensive, practitioners must exercise caution when interpreting the findings. It is important to acknowledge that the outcomes can vary depending on the initialisation and scale factors [30, 31]. The phases of the k-means algorithm are as follows:

1) Initialization

- Choose the number of clusters, k .
- Initialize k centroids.

2) Assignment Step:

- For each data point in your dataset, calculate the distance (e.g., Euclidean distance) to each k centroid using Equation (1).

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (1)$$

- Assign the data point to the cluster whose centroid is closest to it. This forms k clusters.

3) Update Step:

- Calculate the new centroids of the clusters based on the data points assigned to each cluster. This is done by computing the mean of all data points within each cluster using Equation (2).

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} X_q \quad (2)$$

4) Convergence Check:

- Check if the centroids have changed significantly from the previous iteration.
- If the centroids have changed significantly, repeat steps 2 and 3 (Assignment and Update) until convergence. If not, the algorithm has converged, and the final centroids represent the cluster centres.

5) Termination:

- The algorithm terminates once the centroids no longer change significantly or after a predetermined number of iterations.

2.2. Silhouette Index

The silhouette index is a commonly employed statistic in evaluating clustering outcomes since it quantifies the degree of separation and compactness exhibited by the clusters. The metric quantifies the degree of clustering for each data point in relation to its neighbouring cluster, hence offering valuable insights about the suitability of the selected number of clusters and the success of the clustering algorithm. The silhouette index is a metric that falls within the range of -1 to 1, with higher values indicating more well-defined clusters. A number in proximity to 1 indicates the presence of distinct and independent clusters, whilst values in proximity to 0 show the existence of overlapping clusters. Negative values, on the other hand, reflect the possibility of erroneous assignment of data points to clusters. The silhouette index has become widely recognised and valued in the field primarily because of its intuitive interpretation and its effectiveness in accommodating diverse cluster shapes and densities [32–35]. The silhouette index can be derived by utilising Equation (3).

$$S_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max\{a_{(i)}, b_{(i)}\}} \quad (3)$$

Where the variable $a_{(i)}$ represents the average dissimilarity between the i th object and all other objects within the same cluster, the variable $b_{(i)}$ represents the average dissimilarity between the i th object and all other objects in the nearest cluster. The values of $a_{(i)}$ and $b_{(i)}$ can be derived by utilising Equations (4) and (5).

$$a_{(i)} = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \quad (4)$$

$$b_{(i)} = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \quad (5)$$

The variable $|C_I|$ represents the cardinality of cluster C_I , which denotes the number of data points belonging to that specific cluster. On the other hand, $d(i, j)$ represents the distance between two data points, i and j , within the cluster C_I .

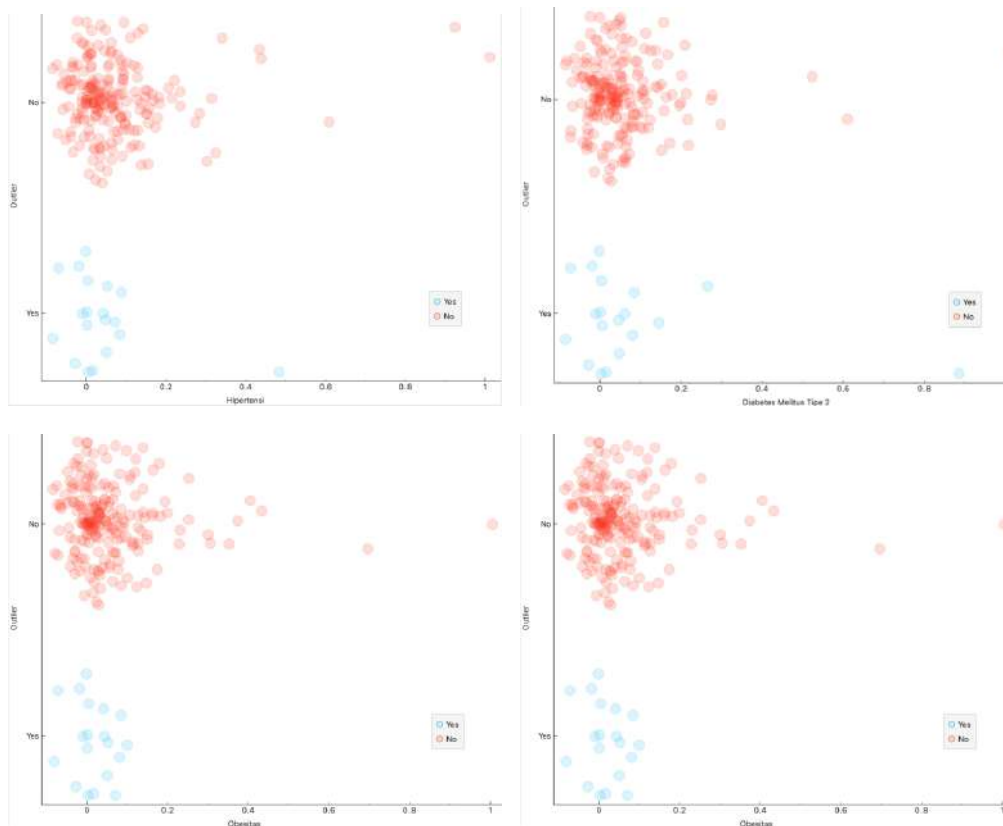


Figure 2. Results of Outlier Analysis for Some Features

3. RESULT AND ANALYSIS

This section contains two subsections. The first subsection describes the research findings, while the second section discusses the findings, including the interpretation of cluster outcomes.

3.1. Results

Outlier analysis is essential in the initial phases of data clustering, serving as a critical step to detect and assess occurrences that exhibit substantial deviation from the average. The main goal is to improve the precision and dependability of the following categorisation procedure. Within this particular context, the analysis fulfils a dual function, aiding in dividing data into two clearly defined scenarios.

In the first scenario, the clusters are carefully organised depending on the results of the outlier analysis. This strategy guarantees that the categorisation procedure is carried out on a purified dataset devoid of the impact of outliers. This strategy aims to identify and emphasise groups that demonstrate exceptional quality by carefully dealing with any exceptional data points. By isolating these clusters, researchers can extract more significant insights and provide more precise forecasts. In contrast, the second scenario entails data grouping without considering the outcomes of the outlier analysis. This methodology offers a comparison method, enabling researchers to evaluate the influence of outliers on the clustering procedure. It assists in comprehending the degree to which outliers impact the overall grouping and aids in assessing the resilience of the clustering algorithm under various circumstances. To conduct a thorough analysis of outliers, various health-related attributes such as hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness were examined closely. The selection of these attributes encompasses a wide range of health-related criteria, guaranteeing a comprehensive evaluation of exceptional cases across multiple dimensions. All of the given health characteristics were found to be outliers, according to the results of the outlier analysis. Along with Pondok Ranji, Sepatan, Banjar, Sukadiri, Cisauk, Bhakti Jaya, Kresek, Mancak, Cihara,

Ciledug, Rengas, Tunjung Teja, Cibitung, Angsana, Sindang Resmi, Cinangka, Anyar, Rajeg, and Carengang, the outliers were found in 19 different datasets or regions. Figure 2 displays a scatter plot that visualises the outliers for a particular feature.

Upon completion of the outlier analysis, the results are incorporated into the clustering procedure in the initial study scenario. This eliminates 19 data points from the entire regional dataset. The ideal number of clusters is determined using the average silhouette index value before starting the clustering process. The k-means widget in the Orange data mining framework allows for emulating an optimal number of clusters by specifying a defined range of cluster numbers. In the current study context, where dataset normalisation is lacking, the specified range extends from 2 to 8 clusters. The cluster distribution shows unevenness, as indicated by the incomplete attainment of the optimal average silhouette index for the comprehensive cluster set. The uneven distribution highlights the need for more improvement within the clustering architecture. Data normalisation plays a crucial role in improving the results of clustering. After the normalisation process, there is a noticeable increase in the average silhouette index value for each cluster. This rise demonstrates an enhanced level of unity and distinctiveness among the clusters. These findings demonstrate that normalising the data has a beneficial impact on enhancing the quality of the clustering outcomes. Despite attempts to normalise the data, the cluster arrangement that yields the highest silhouette index remains at $k = 2$. The current setup produces an average silhouette index of 0.81, indicating strong distinction and unity among the clusters. The comparative silhouette indices for different cluster topologies, with values of k equal to 3, 4, 5, 6, 7, and 8, are 0.701, 0.664, 0.649, 0.456, 0.494, and 0.514, respectively. The values provide evidence for the exceptional quality and consistency of the $k = 2$ cluster setup. To summarise, the repeated incorporation of outlier analysis and subsequent clustering procedures is crucial in extracting detailed insights from complex datasets. The methodical process of refining the data, which includes removing outliers and normalising the data, enhances the optimisation of cluster quality. Using the silhouette index as a metric helps determine the optimal number of clusters, leading to a better understanding of the patterns in the dataset.

Table 1. Distribution of Cluster Member for Each k in the First Scenario Two

silhouette-NN	0,789	0,728	0,732	0,667	0,388	0,419	0,42
silhouette-Nr	0,812	0,701	0,664	0,649	0,456	0,494	0,514
Number of k	2	3	4	5	6	7	8
C1-Nr	202	193	14	6	157	12	13
C1-NN	10	191	193	185	1	134	134
C2-Nr	6	14	189	5	5	5	2
C2-NN	198	16	12	17	5	6	5
C3-Nr	0	1	4	1	7	20	153
C3-NN	0	1	1	1	122	1	1
C4-Nr	0	0	1	184	32	6	2
C4-NN	0	0	2	2	12	2	2
C5-Nr	0	0	0	12	1	158	5
C5-NN	0	0	0	3	62	3	1
C6-Nr	0	0	0	0	6	6	9
C6-NN	0	0	0	0	6	45	45
C7-Nr	0	0	0	0	0	1	23
C7-NN	0	0	0	0	0	17	17
C8-Nr	0	0	0	0	0	0	1
C8-NN	0	0	0	0	0	0	3

Initially, the clustering method yields that 202 regions are assigned to the first cluster (C1), while the remaining 6 regions are allocated to the second cluster (C2). More precisely, there are 193 regions grouped as members of category C1, 14 regions assigned to category C2, and one region allocated to category C3 in clusters where the value of k equals 3. Table 1 presents the detailed distribution of cluster members for each k , categorised based on their normalisation state (normalised is indicated as Nr and non-normalised is denoted as NN). The data presented in Table 1 indicates that when the value of k decreases, there is a stronger inclination for individuals from one cluster to merge with individuals from other clusters with a larger number of members, reducing the number of members in the remaining clusters. This discovery highlights the correlation between the selected number of clusters and the distribution patterns of regional members. Furthermore, the arrangement of cluster participants, whether standardised or not, demonstrates the enduring presence of regional clustering. However, it is important to mention that the occurrence of regional clustering only undergoes a slight change in its location. Although there may be differences in data normalisation, the overall trend of regional clustering stays generally stable. Table 1 presents the outcomes of the clustering study, illustrating the variations in the number of clusters selected and the distribution patterns of regional members over time. Regional agglomeration is evident in both normalised and non-normalised datasets, and any disparities in their manifestation are relatively inconsequential.

In the second clustering scenario, the dataset is straightaway processed using the k-means method without any prior outlier data removal. The cluster outcomes show a pattern where, as the value of k lowers, the cluster member areas become more concentrated within one cluster or specific clusters. Significantly, the number of cluster members in C1 (with data normalisation indicated as Nr) exceeds that of C2 when k equals 2. In contrast, if the data is not normalised, a reverse correlation arises, where cluster C2 includes a greater number of members compared to cluster C1. This trend remains consistent for the majority of other k values. The distribution of cluster members shows inequality between $k = 2$ and $k = 8$. However, as k increases, there is a noticeable pattern where the distribution of cluster members becomes more spread out, although the overall pattern of members clustering together remains unchanged. In the second clustering scenario, Table 2 provides detailed information about the distribution of cluster members for each value of k . Furthermore, Figure 3 presents a graphical depiction of the clustering results, differentiating between normalised (a) and non-normalised data (b), respectively.

Table 2. Distribution of Cluster Member for Each k for the Second Scenario

silhouette-NN	0,769	0,682	0,595	0,613	0,566	0,57	0,467
silhouette-Nr	0,805	0,724	0,666	0,696	0,404	0,466	0,463
Number of k	2	3	4	5	6	7	8
C1-Nr	218	19	12	17	16	6	41
C1-NN	18	36	58	13	7	22	21
C2-Nr	9	206	205	1	1	166	2
C2-NN	209	188	151	2	50	7	7
C3-Nr	0	2	2	2	2	7	167
C3-NN	0	3	2	151	6	139	101
C4-Nr	0	0	8	204	147	1	1
C4-NN	0	0	16	55	139	49	38
C5-Nr	0	0	0	3	3	2	3
C5-NN	0	0	0	6	2	6	6
C6-Nr	0	0	0	0	58	44	5
C6-NN	0	0	0	0	23	2	2
C7-Nr	0	0	0	0	0	1	1
C7-NN	0	0	0	0	0	2	2
C8-Nr	0	0	0	0	0	0	7
C8-NN	0	0	0	0	0	0	50

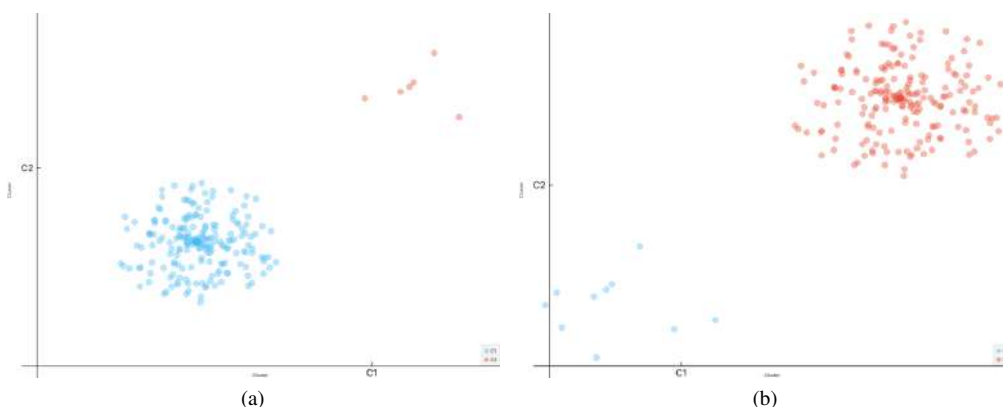


Figure 3. The difference between cluster findings with normalisation (a) and without normalisation (b)

3.2. Discussion

The cluster analysis of the two scenarios produces significant results, especially when evaluating the average silhouette index values. The initial analysis reveals that the highest average silhouette index value is observed when k is set to 2. This finding serves as the fundamental criterion for determining the value of k in the following analyses conducted in this inquiry. Each cluster member at $k = 2$ demonstrates a silhouette value near 1, suggesting strong cohesiveness and isolation within the clusters. Aligning with

previous studies conducted by [32] and [36], the Silhouette Index determines the most suitable number of clusters. According to their statement, this index can be utilised to ascertain the most suitable number of clusters before clustering. The results, especially after data normalisation, confirm that the clustering results with $k = 2$ have higher validity than other configurations. Previous studies have found that normalising data before applying machine learning approaches can enhance the effectiveness of both clustering [37, 38], and classification [39, 40], according to research findings. Hence, it is unsurprising that in this study, the application of normalisation can enhance cluster validity, thereby enabling it to ascertain the most suitable number of clusters.

Confirming the cluster results with a value of $k = 2$ shows that 202 regions are grouped under cluster C1, while the rest are assigned to cluster C2. By lining up these cluster results with infographics, it's clear that cluster C2 includes areas with almost all 10 types of noncommunicable diseases that were looked at in this study. Cluster C1 exhibits the largest frequency of individuals with non-communicable diseases, particularly for the five most common conditions: diabetes, hypertension, obesity, stroke, and cataracts. In contrast, the 202 locations in cluster C1 have a remarkably low occurrence of diseases such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

The results of this study emphasise the urgent requirement for focused attention in up to 202 regions, especially those vulnerable to the occurrence of the five illnesses: diabetes, hypertension, obesity, stroke, and cataracts. At the same time, the six places that makeup cluster C2 need immediate attention because they have high rates of the five diseases listed above, as well as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This investigation confirms the crucial significance of customised healthcare interventions based on specific regional health profiles. The focus on clusters C1 and C2 enables the implementation of specific initiatives that recognise the distinct patterns of non-communicable illnesses in each cluster. Public health programmes must prioritise addressing the distinct healthcare requirements of these clusters by customising interventions to tackle prevalent disorders and alleviate the burden of diseases in the designated locations.

4. CONCLUSION

To summarise, the research results emphasise the urgent requirement for specific healthcare interventions in the province of Banten, specifically about non-communicable diseases (NCDs). Using the k-means algorithm to do clustering analysis on NCD markers for 208 regions shows how important each part of the province has its unique health profile. The clustering results, especially when k is set to 2, indicate that 202 regions require urgent attention because of the high occurrence of diabetes, hypertension, obesity, stroke, and cataracts. Besides the main diseases listed above, the province's other 28 regions deal with a wider range of noncommunicable diseases (NCDs), such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This extensive analysis of regional health profiles establishes a basis for focused public health initiatives, highlighting the imperative need for government action in specified locations.

Nevertheless, it is imperative to recognise the constraints of the clustering methodology utilised in this research. Although the k-means algorithm efficiently classifies regions into two main clusters based on NCD markers, it does not provide information about the spatial proximity of NCD sufferers across regions. Given this constraint, it is clear that further research using other methods is necessary to thoroughly investigate the pattern of closeness among individuals with non-communicable diseases in different areas. This necessitates an intricate method of comprehending the frequency of particular illnesses in each group and the interaction of geographical elements that impact healthcare dynamics.

Future research needs to concentrate on improving techniques for capturing the spatial linkages and proximity patterns among individuals with non-communicable diseases (NCDs). This deeper comprehension will enable the implementation of more accurate and customised healthcare plans, guaranteeing that interventions are customised to the individual requirements of each location. Integrating spatial analyses into future research will enhance the effectiveness of tackling the intricate terrain of non-communicable diseases in the province of Banten.

5. DECLARATIONS

AUTHOR CONTRIBUTION

All authors contributed to the writing of this article.

FUNDING STATEMENT

This research was self-funded, and the authors did not receive any external financial support for the design, data collection, analysis, or interpretation of the study. All expenses related to this research were borne by the authors personally.

COMPETING INTEREST

The authors declare no conflict of interest in this article.

REFERENCES

- [1] S. Bhattacharya, P. Heidler, and S. Varshney, "Incorporating neglected non-communicable diseases into the national health program: A review," *Frontiers in Public Health*, vol. 10, no. January, pp. 1–9, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.1093170/full>
- [2] M. F. Owusu, J. Adu, B. A. Dorte, S. Gyamfi, and E. Martin-Yeboah, "Exploring health promotion efforts for non-communicable disease prevention and control in Ghana," *PLOS Global Public Health*, vol. 3, no. 9, pp. 1–14, 2023. [Online]. Available: <https://dx.plos.org/10.1371/journal.pgph.0002408>
- [3] A. Odunyemi, T. Rahman, and K. Alam, "Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19," *BMC Public Health*, vol. 23, no. 1, pp. 1–12, 2023. [Online]. Available: <https://bmcpublihealth.biomedcentral.com/articles/10.1186/s12889-023-16498-7>
- [4] H. G. A. S. Samarasinghe, D. A. T. D. S. Ranasinghe, W. R. Jayasekara, S. A. A. D. Senarathna, J. D. P. M. Jayakody, P. M. Kalubovila, M. D. Edirisuriya, and N. S. A. S. N. Senarath, "Barriers to Accessing Medical Services and Adherence to Recommended Drug Regimens among Patients with Non-Communicable Diseases: A Study at Divisional Hospital Thalagama, Sri Lanka," in *IECN 2023*. MDPI, 2023, pp. 1–6. [Online]. Available: <https://www.mdpi.com/2673-9976/29/1/14>
- [5] K. S. Maliangkay, U. Rahma, S. Putri, and N. D. Istanti, "Analisis Peran Promosi Kesehatan Dalam Mendukung Keberhasilan Program Pencegahan Penyakit Tidak Menular Di Indonesia," *Jurnal Medika Nusantara*, vol. 1, no. 2, pp. 108–122, 2023.
- [6] H. B. H. Akbar, and S. Sarman, "Pencegahan Penyakit Tidak Menular Melalui Edukasi Cerdik Pada Masyarakat Desa Moyag Kotamobagu," *Abdimas Universal*, vol. 3, no. 1, pp. 83–87, 2021. [Online]. Available: <http://abdimasuniversal.uniba-bpn.ac.id/index.php/abdimasuniversal/article/view/94>
- [7] R. Gupta, K. Gaur, and C. V. S. Ram, "Emerging trends in hypertension epidemiology in India," *Journal of Human Hypertension*, vol. 33, no. 8, pp. 575–587, 2019. [Online]. Available: <https://www.nature.com/articles/s41371-018-0117-3>
- [8] C. Antza, G. Kostopoulos, S. Mostafa, K. Nirantharakumar, and A. Tahrani, "The links between sleep duration, obesity and type 2 diabetes mellitus," *Journal of Endocrinology*, vol. 252, no. 2, pp. 125–141, 2022. [Online]. Available: <https://joe.bioscientifica.com/view/journals/joe/252/2/JOE-21-0155.xml>
- [9] L. Wang, S. Wang, Q. Zhang, C. He, C. Fu, and Q. Wei, "The role of the gut microbiota in health and cardiovascular diseases," *Molecular Biomedicine*, vol. 3, no. 1, pp. 1–50, 2022. [Online]. Available: <https://link.springer.com/10.1186/s43556-022-00091-2>
- [10] A. A. Samarraie, M. Pichette, and G. Rousseau, "Role of the Gut Microbiome in the Development of Atherosclerotic Cardiovascular Disease," *International Journal of Molecular Sciences*, vol. 24, no. 6, pp. 1–17, 2023. [Online]. Available: <https://www.mdpi.com/1422-0067/24/6/5420>
- [11] R. T. Chlebowski, J. Luo, G. L. Anderson, W. Barrington, K. Reding, M. S. Simon, J. E. Manson, T. E. Rohan, J. Wactawski-Wende, D. Lane, H. Strickler, Y. Mosaver-Rahmani, J. L. Freudenheim, N. Saquib, and M. L. Stefanick, "Weight loss and breast cancer incidence in postmenopausal women," *Cancer*, vol. 125, no. 2, pp. 205–212, 2019. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.31687>
- [12] M. Ellingjord-Dale, S. Christakoudi, E. Weiderpass, S. Panico, L. Dossus, A. Olsen, A. Tjønneland, R. Kaaks, M. B. Schulze, G. Masala, I. T. Gram, G. Skeie, A. H. Rosendahl, M. Sund, T. Key, P. Ferrari, M. Gunter, A. K. Heath, K. K. Tsilidis, and E. Riboli, "Long-term weight change and risk of breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) study," *International Journal of Epidemiology*, vol. 50, no. 6, pp. 1914–1926, 2022. [Online]. Available: <https://academic.oup.com/ije/article/50/6/1914/6182058>
- [13] E. Kričković, T. Lukić, and D. Jovanović-Popović, "Geographic Medical Overview of Noncommunicable Diseases (Cardiovascular Diseases and Diabetes) in the Territory of the AP Vojvodina (Northern Serbia)," *Healthcare*, vol. 11, no. 1, pp. 1–33, 2022. [Online]. Available: <https://www.mdpi.com/2227-9032/11/1/48>

- [14] T. B. Darikwa and S. O. Manda, "Spatial Co-Clustering of Cardiovascular Diseases and Select Risk Factors among Adults in South Africa," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, pp. 1–16, 2020. [Online]. Available: <https://www.mdpi.com/1660-4601/17/10/3583>
- [15] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning," *Applied Sciences*, vol. 13, no. 3, pp. 1–15, 2023. [Online]. Available: <https://www.mdpi.com/2076-3417/13/3/1509>
- [16] L. Zhang, G. Yang, and X. Li, "Mining sequential patterns of PM2.5 pollution between 338 cities in China," *Journal of Environmental Management*, vol. 262, no. March, pp. 1–8, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0301479720302760>
- [17] D. Majcherek, M. A. Weresa, and C. Ciecierski, "A Cluster Analysis of Risk Factors for Cancer across EU Countries: Health Policy Recommendations for Prevention," *International Journal of Environmental Research and Public Health*, vol. 18, no. 15, pp. 1–14, 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/15/8142>
- [18] M. A. Emon, A. Heinson, P. Wu, D. Domingo-Fernández, M. Sood, H. Vrooman, J.-C. Corvol, P. Scordis, M. Hofmann-Apitius, and H. Fröhlich, "Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020. [Online]. Available: <https://www.nature.com/articles/s41598-020-76200-4>
- [19] J. Prakash, V. Wang, R. E. Quinn, and C. S. Mitchell, "Unsupervised Machine Learning to Identify Separable Clinical Alzheimer's Disease Sub-Populations," *Brain Sciences*, vol. 11, no. 8, pp. 1–21, 2021. [Online]. Available: <https://www.mdpi.com/2076-3425/11/8/977>
- [20] S. Bhattacharjee, Y.-B. Hwang, R. I. Sumon, H. Rahman, D.-W. Hyeon, D. Moon, K. S. Carole, H.-C. Kim, and H.-K. Choi, "Cluster Analysis: Unsupervised Classification for Identifying Benign and Malignant Tumors on Whole Slide Image of Prostate Cancer," in *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*. IEEE, 2022, pp. 1–5. [Online]. Available: <https://ieeexplore.ieee.org/document/10052952/>
- [21] Y. Jiang, Z.-G. Yang, J. Wang, R. Shi, P.-L. Han, W.-L. Qian, W.-F. Yan, and Y. Li, "Unsupervised machine learning based on clinical factors for the detection of coronary artery atherosclerosis in type 2 diabetes mellitus," *Cardiovascular Diabetology*, vol. 21, no. 1, pp. 1–10, 2022. [Online]. Available: <https://cardiab.biomedcentral.com/articles/10.1186/s12933-022-01700-8>
- [22] G. Sarveswaran, V. Kulothungan, and P. Mathur, "Clustering of noncommunicable disease risk factors among adults (1869 years) in rural population, South-India," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1005–1014, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1871402120301624>
- [23] S. V. Rocha, S. C. de Oliveira, H. L. R. Munaro, C. F. R. Squarcini, B. M. P. Ferreira, F. de Oliveira Mendonça, and C. A. dos Santos, "Cluster analysis of risk factors for chronic non-communicable diseases in elderly Brazilians: population-based cross-sectional studies in a rural town," *Research, Society and Development*, vol. 10, no. 17, pp. 1–10, 2021. [Online]. Available: <https://rsdjournal.org/index.php/rsd/article/view/24202>
- [24] R. Uddin, E.-Y. Lee, S. R. Khan, M. S. Tremblay, and A. Khan, "Clustering of lifestyle risk factors for non-communicable diseases in 304,779 adolescents from 89 countries: A global perspective," *Preventive Medicine*, vol. 131, no. December, pp. 1–8, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0091743519304384>
- [25] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 337–350, 2022. [Online]. Available: <https://journal.universitatumigora.ac.id/index.php/matrik/article/view/1411>
- [26] H. Hairani, D. Susilowati, I. P. Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 275–282, 2022. [Online]. Available: <https://journal.universitatumigora.ac.id/index.php/matrik/article/view/1542>

- [27] S. Annas, B. Poerwanto, S. Sapriani, and M. F. S, "Implementation of K-Means Clustering on Poverty Indicators in Indonesia," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 257–266, 2022. [Online]. Available: <https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/1289>
- [28] Y. Zhao and X. Zhou, "K-means Clustering Algorithm and Its Improvement Research," *Journal of Physics: Conference Series*, vol. 1873, no. 1, pp. 1–5, 2021. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1873/1/012074>
- [29] M. Darwis, L. H. Hasibuan, M. Firmansyah, N. Ahady, and R. Tiaharyadini, "Implementation of K-Means clustering algorithm in mapping the groups of graduated or dropped-out students in the Management Department of the National University," *JISA(Jurnal Informatika dan Sains)*, vol. 4, no. 1, pp. 1–9, 2021. [Online]. Available: <http://trilogi.ac.id/journal/ks/index.php/JISA/article/view/848>
- [30] A. R. Danurisa and J. Heikal, "Customer Clustering Using the K-Means Clustering Algorithm in the Top 5 Online Marketplaces in Indonesia," *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, vol. 5, no. 3, 2022.
- [31] A. Chaerudin, D. T. Murdiansyah, and M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Indonesia Journal on Computing (Indo-JC)*, vol. 6, no. 1, pp. 53–60, 2021.
- [32] A. Dudek, *Silhouette Index as Clustering Evaluation Tool*, 2020, pp. 19–33. [Online]. Available: http://link.springer.com/10.1007/978-3-030-52348-0_{_}2
- [33] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/6/759>
- [34] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Techno.Com*, vol. 20, no. 2, pp. 186–197, 2021. [Online]. Available: <http://publikasi.dinus.ac.id/index.php/technoc/article/view/4556>
- [35] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 6, no. 2, pp. 48–54, 2021. [Online]. Available: <https://ojs.unanda.ac.id/index.php/jiit/article/view/659>
- [36] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 174–182, 2023. [Online]. Available: <https://online-journals.org/index.php/i-joe/article/view/37059>
- [37] T. Li, Y. Ma, and T. Endoh, "Normalization-Based Validity Index of Adaptive K-Means Clustering for Multi-Solution Application," *IEEE Access*, vol. 8, pp. 9403–9419, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/8952702/>
- [38] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *Journal of Physics: Conference Series*, vol. 1566, no. 1, pp. 1–7, 2020. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1566/1/012112>
- [39] H. A. Ahmed, P. J. M. Ali, A. K. Faeq, and S. M. Abdullah, "An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method," *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 2, pp. 29–37, 2022. [Online]. Available: <https://aro.koyauniversity.org/index.php/aro/article/view/970>
- [40] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, vol. 10, no. 11, p. 1942, 2022. [Online]. Available: <https://www.mdpi.com/2227-7390/10/11/1942>



Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Submissions

- See page 2

2

Add the research objectives explicitly and the contribution of this research to the development of science or its benefits.

We've already add the research objective and the benefit (contribution).

- Mark as yellow on page 3

3

Move Figure 1 to the methods section

- We've moved Figure 1 as your sugesstion.

- See Page 3

4

1. findings of this research explicitly, starting with the words "The findings of this research are..."

2. There needs to be justification from previous research that strengthens or contradicts the results of this research which begins with the words "The results of this research are in line with or supported by..."

- We've adjusted the paragraph based on your sugesstion and adding some reference to justified the finding.

- Mark as yellow on page 7-8

5

References marked yellow should be replaced with journals from the last 5 years, from 2019 - 2023

- We already replace the marked yellow reference and add 4 references (ref 1-4) and also reference 6. All references already replaced with journal from the last 5

information about the contents of the declaration.



Participants

Hairani Hairani (hairani10)

Tb Ai Munandar (tbaimunandar)

Messages

Note


From

dear Author,

khairanmarzuki

Feb 09

Please fill in the declaration completely, wait until February 12, 2024
regards

 [khairanmarzuki, Production editor, 3352-Article Text-19674 \(Munandar\)_Copyediting.docx](#)

▶ Dear Editor Team,


tbaimunandar

Mar 27

I am really sorry for the late response. I did not receive any notification in my email about this discussion, but I declare that this research was funded by myself and my team. Also, I declare that there is no conflict of interest in this paper.

And, I already checked my galley at the production section for acknowledgement; it is already fixed.

Best regards,

 [tbaimunandar, Author, 3352-Article Text-20486-2-9-20240201.docx](#)

Add Message



Submission Library

View Metadata

Submissions

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Revision_1



Participants

Hairani Hairani (hairani10)

Tb Ai Munandar (tbaimunandar)

Messages

Note

From

Dear Author,

The manuscript has been reviewed, but there are some suggestions from our editors to improve the quality of the manuscript. Some suggestions that the author should do, can be seen below. Revised articles can be uploaded here no later than November 25, 2023.

Reviewer 1 :

1. What is the purpose of this study? The abstract's final results do not match with the conclusions.
2. Many typos were found
3. Clearly state the research purpose in the introduction
4. The purpose of this study is to compile data on noncommunicable diseases. The author should explain the criteria for grouping. The author must ensure that the process in question is clustering, not classification, because the final results of the two processes are different.
5. The research steps necessary to reach the aim are not yet visible. It is preferable to create research stages in chart design.
6. In result section :
 - Why is the conversation focusing on outlier data? According to the title or abstract, outliers are not the main cause of the

hairani10
Nov 13

problem. The discussion here is not acceptable.

-What are the criteria for grouping non-communicable diseases?

-The author must explain the occurrence of outliers

-The author must explain how outliers and the silhouette index are related.

7. Use references from the last 5 years (2019-2023)



Submission Library

View Metadata

Submissions

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Recommendation



Participants

Hairani Hairani (hairani10)

Tb Ai Munandar (tbaimunandar)

Messages

Note

From

:

haira
Jan 0

The recommendation regarding the submission to MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, "Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm" is: Accept Submission

So that your manuscript can be published in the Matric Journal, please make a payment of IDR. 1,000,000 To No. Rec. 0002101500563504, Bank: BTN, Name : LPPM Bumigora University until Januari 25 2024. Proof of payment can be sent to this email.

This payment includes proofreading services, plagiarism checking, and the excess number of article pages.

In addition, the author is expected to send a statement letter via email no later than Januari 25 2024. The statement can be downloaded at the following link.

https://docs.google.com/document/d/1_wMqxYhq0dkedEw1q28QUUM6IU7t3YNm/edit

Hairani Hairani
Universitas Bumigora
hairani@universitasbumigora.ac.id

[MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer](#)

▶ Dear MATRIK Journal Editorial Team.

tbain

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Tb Ai Munandar and Ajif Yunizar Yusuf Pratama

Department of Informatics, Universitas Bhayangkara Jakarta Raya, Indonesia

Article Info

Article history:

Received mm dd, yyyy

Revised mm dd, yyyy

Accepted mm dd, yyyy

Keywords:

clustering

k-means

non-communicable diseases

regional clustering

silhouette index

ABSTRACT (10 PT)

Noncommunicable diseases (NCDs) have become a global threat to public health, necessitating a comprehensive understanding of their geographic and epidemiological distribution in order to devise appropriate interventions. The objective of this study is to clustering areas of Banten Province based on NCDS profiles using the unsupervised learning technique. The processing and normalisation of NCDS prevalence data from various health sources preceded cluster analysis using the K-Means clustering algorithm. This research is categorised into two scenarios: the first involves the clustering of data obtained from outlier analysis, while the second scenario excludes any outliers. The objective is to observe disparities in regional clustering outcomes by categorising non-communicable diseases according to these two scenarios. The silhouette index is used to determine the validity of cluster results. These findings are analysed in depth to determine the geographic and socioeconomic patterns associated with each cluster's NCDS profile. Based on the mean silhouette index value of 0.812, the results indicate that the sum of $k = 2$ in the k-means algorithm is the optimal cluster result in this case. Five non-communicable diseases, namely diabetes, hypertension, obesity, stroke, and cataracts, necessitate significant focus in the first cluster (C1), where 202 regions were grouped. Six regions belong to the second cluster (C2), which includes areas that are not only susceptible to the five non-communicable diseases in cluster C1 but also to breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

Copyright ©2022 The Authors.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Tb Ai Munandar, 081384512710

Faculty of Computer Science, Informatics Department

Universitas Bhayangkara Jakarta Raya, Jawa Barat, Indonesia

Email: tbaimunandar@gmail.com

How to Cite:

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

1 INTRODUCTION

Non-communicable diseases (NCDs) comprise a collection of chronic health conditions that do not spread via direct human-to-human contact and are the leading cause of cancer-related deaths worldwide. This grouping contains a variety of illnesses, including chronic respiratory conditions, diabetes, cardiovascular disease, and cancer, which collectively impose a significant burden on global health. The gravity of non-communicable diseases (NCDs) is emphasised by the World Health Organisation (WHO), which estimates that these conditions are responsible for around 41 million deaths per year, or 71% of all casualties worldwide [1]–[4]. The development of non-communicable diseases (NCDs) is strongly associated with detrimental lifestyle decisions, which include unequal dietary patterns, a lack of physical activity, and the use of tobacco and alcohol. Due to factors such as urbanisation, evolving behaviours, and rapid population growth, the Indonesian province of Banten faces substantial challenges associated with non-communicable diseases (NCDs). In 2021, Banten Provincial Health Service released data that

indicates Banten Province has experienced an ongoing and consistent increase in the prevalence of diabetes, hypertension, and obesity [5]. The substantial increase in the prevalence of non-communicable diseases presents a tough obstacle for the regional health infrastructure, significantly affecting the community's overall standard of living [6], [7]. An in-depth examination of the patterns and distribution of non-communicable diseases (NCDs) in Banten is necessary in order to address their complexity. When it comes to developing targeted interventions, the current regional grouping system predicated on NCD categories may not be ideal.

Applying unsupervised learning techniques, specifically the cluster method, to divide Banten Province into groups based on different types of non-communicable diseases (NCDs) is the point of this research. Clustering presents a more sophisticated methodology in comparison to conventional grouping techniques, as it enables the detection of inherent groupings within the data that do not require predetermined labelling. This study seeks to obtain a more precise comprehension of the distribution of non-communicable diseases (NCDs) throughout Banten Province by delineating regional clusters according to different NCD categories. For informing health policies and directing medical interventions, precise and comprehensive data on the prevalence and distribution of noncommunicable diseases (NCDs) in various regions of Banten Province are indispensable. By providing health policymakers with the ability to customise more efficacious prevention and intervention strategies, the regional clustering inferred from this study has the potential to yield significant insights regarding the spatial distribution of non-communicable disease prevalence. The resulting regional groups will provide a solid foundation for the development of targeted non-communicable disease (NCD) prevention programmes, letting policymakers and medical professionals pay close attention to the specific needs of each cluster. Using unsupervised learning techniques, particularly clustering, to illustrate the intricate picture of NCD prevalence, this study's primary objective is to aid in the improvement of public health strategies in Banten Province. By employing this novel methodology, the research endeavours to establish a strong groundwork for decision-making grounded in evidence, thereby promoting a more sophisticated and focused approach to addressing the complex issues presented by non-communicable diseases in the area.

To face the challenges of NCDS in the province of Banten, a comprehensive and integrated approach to grouping areas based on the nature of NCDS is required. Currently, regional groupings are typically determined by administration or geographic location, without taking into account the unique health profiles of each region. Using unsupervised learning techniques, such as the clustering method, we will be able to identify patterns and relationships in Banten Province NCDS data. This phase is essential for designing more targeted interventions and enhancing health services in regions affected by NCDS. Moreover, research conducted by [8] on the epidemiology of hypertension in India suggests that accelerated urbanisation may contribute to the rising prevalence of hypertension in urban areas. The prevalence of obesity tends to be higher in urban areas than in rural areas, highlighting the significance of taking regional differences into account when managing NCDs. These results indicate that the social and economic changes that result from urbanisation can have an effect on disease patterns in the region, and it is not inconceivable that this could occur in Banten Province. Noncommunicable diseases (NCDs) are a major public health concern because they contribute to the world's elevated mortality rate and disease burden. Numerous studies have been conducted in recent years to understand the risk factors and transmission patterns of NCDS and to identify more effective prevention strategies. A study by Chen provide the most recent data regarding the prevalence and risk factors of type 2 diabetes in various populations. The findings of this study shed light on the association between diet and physical activity and the risk of type 2 diabetes. Several studies have investigated the impact of diet in reducing the risk of cardiovascular disease as part of efforts to prevent NCDS. For instance, research conducted by found a correlation between adult fast food consumption and increased insulin resistance. Additionally, recent research has emphasised the significance of optimal sleep patterns in reducing the risk of NCDS. Study by [9] discovered that adult sleep deprivation increases the risk of adiposity. In order to comprehend the effect of physical activity on the prevention of NCDs, Daskalopoulou (2021) conducted a meta-analysis which revealed that higher levels of physical activity are associated with lower risks for certain NCDs. Recent research has also focused on the role that the environment plays in NCDS. The composition of the intestinal microbiota is associated with the development of cardiovascular disease, according to [10], [11]. In terms of NCDS prevention, understanding the function of the environment is crucial. Zimmet in 2021 investigate the global repercussions of the diabetes epidemic and other NCDs. In addition to focusing on NCDS in the adult population, researchers have also studied the disorder in children and at the same time, Samavat in 2021 examined the association between a child's to adult diet and their future risk of breast cancer [12], [13]. However, a portion of the aforementioned research places greater emphasis on discerning various methods of intervention and enhancing healthcare provisions by utilising assessments of individuals afflicted with non-communicable diseases. The topic of regional-specific approaches to healthcare intervention and management has not yet been addressed. The issue of non-communicable illness spreading often remains unresolved to some extent due to a lack of information regarding treatment priorities specific to different regions. In order to enhance health services and interventions in the future, it is imperative to adopt a regional cluster-based approach.

The distinction between the present study and prior studies is in the categorization of non-communicable diseases based on geographical regions. By implementing interventions and making enhancements, the quality of health care can be significantly improved. Collectively, the most recent research provides valuable insights into the effective management of NCDS. Through a greater understanding of risk factors and prevention strategies, it is anticipated that targeted preventive measures can be implemented to lessen the burden of noncommunicable diseases (NCDs) and improve public health as a whole.



Noncommunicable diseases (NCDs) are a significant global health burden, and it is essential to understand the patterns and patterns of dissemination of these diseases in a specific region for prevention and appropriate management. The unsupervised learning method has become a valuable instrument in the analysis of health data, including the grouping of regions by NCD type. Several recent studies have utilised unsupervised learning techniques, such as clustering and cluster analysis, to identify geographic and epidemiological patterns of NCDs in different regions. A study by [14] grouped regions based on the type of NCDs in a country using cluster analysis. This study reveals how to identify the most and least burdened counties. In a related study by [15], [16] utilised spatial clustering and unsupervised learning to categorise regions based on the pattern of cardiovascular disease distribution. The results of this study indicate the existence of certain health clusters that can be targeted by interventions to reduce the risk of cardiovascular disease. In addition, unsupervised learning techniques have been employed to determine the connection between air pollution patterns, the urban spread of respiratory diseases and health interventions. Research by [17] found that cluster patterns from air pollution data and the incidence of respiratory disease are frequently interconnected. Recent research has also investigated the use of machine learning, including unsupervised learning, for grouping regions based on other categories of diseases, such as cancer, benign and malignant tumors, diabetes, and neurodegenerative diseases [18]–[22]. In order to address the challenges posed by NCDs at the regional level, an unsupervised learning algorithm used to identify patterns of interrelationships between specific NCDs in specific regions of a country [23] and also at the same time some researchers grouped regions based on the level of exposure to certain NCDs risk factors, using an unsupervised learning method [24], [25]. Overall, the application of unsupervised learning has created new opportunities to segment regions based on the type of NCDs and for gaining a deeper understanding of the patterns of disease transmission, one of the most popular algorithms is k-means. Using this strategy, it is anticipated that prevention and intervention can be more effectively targeted based on the local community's health characteristics. The k-means algorithm is used for a variety of reasons. Apart from being frequently used in health research, it may also be used to segment promotional places in education, facilities, and teachers [26], [27] and group poverty indicators in a region [28]. The objective of this study is to categorise non-communicable diseases based on geographical regions by analysing patient data obtained from public health institutions in Banten Province. This research is anticipated to have a significant impact on local governments' ability to identify regional priorities for managing the development of non-communicable diseases. In addition, unsupervised learning methods, such as the k-means algorithm, can be utilised as an alternate strategy to analyse non-communicable disease data, thereby making a valuable contribution to the health sector.

This paper is divided into four sections. The first section presents the introduction, which includes the problem's background, a literature review, research gaps, and systematic information about the article. The second section explains the study methodologies used, and the third section includes the research results and commentary. Meanwhile, the fourth portion is the conclusion, which contains the findings of the investigation.

2 RESEARCH METHOD

This study use unsupervised learning as its research methodology to categorise regions within Banten Province according to the specific noncommunicable disease (NCD) type. Figure 1 provides a more detailed overview of the study stages.

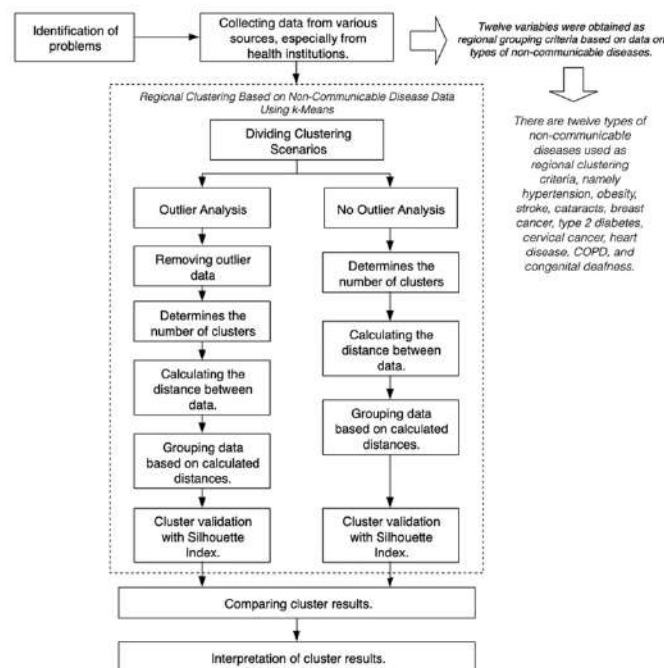


Figure 1. Research Stages

During the preliminary stage of the inquiry, data pertaining to the incidence of noncommunicable diseases (NCDs) were gathered from several sources, encompassing public health institutions and hospitals situated within Banten Province. A total of 227 instances of regional information representing subdistricts were gathered as a result of the data gathering process. The data is subsequently submitted to a preprocessing stage, wherein incomplete or fragmented data are repaired, and all variables are normalised to guarantee consistency in scale. Consequently, the strategy for categorising regions based on NCDS profiles involved the use of the K-Means clustering algorithm, which falls under the umbrella of unsupervised learning techniques. Cluster analysis is conducted by determining the distance between the data points and the centroid of each cluster. This process involves combining regions that exhibit similar NCDS profiles into a cohesive cluster. The quality of the clusters was evaluated by performing validation of the cluster results using the silhouette index. The clustering criteria are determined by twelve specific non-communicable diseases, which include hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, COPD, and congenital deafness.

2.1. K-means

The k-means algorithm is a commonly employed clustering technique that divides a dataset into separate groups according to their similarity. This is achieved through an iterative process of assigning data points to the nearest cluster centroid and adjusting the centroids to minimise the total sum of squared distances. The success of the method in numerous domains can be attributed to its efficiency and simplicity, despite the acknowledged drawbacks of being sensitive to initial centroid selection and prone to converging to local optima [29], [30]. The k-means algorithm is widely recognised for its prevalence in the field of unsupervised learning. It is commonly employed in several domains such as picture segmentation, customer segmentation, and anomaly detection. Although the utilisation of this method is extensive, it is crucial for practitioners to exercise caution when interpreting the findings. It is important to acknowledge that the outcomes can vary depending on the initialization and scale factors [31], [32]. The phases of the k-means algorithm are as follows:

1) Initialization:

- Choose the number of clusters, k.
- Initialize k centroids.

2) Assignment Step:

- For each data point in your dataset, calculate the distance (e.g., Euclidean distance) to each of the k centroids using (1).

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (1)$$

- Assign the data point to the cluster whose centroid is closest to it. This forms k clusters.

3) Update Step:

- Calculate the new centroids of the clusters based on the data points assigned to each cluster. This is done by computing the mean of all data points within each cluster using (2).

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad (2)$$

4) Convergence Check:

- Check if the centroids have changed significantly from the previous iteration.
- If the centroids have changed significantly, repeat steps 2 and 3 (Assignment and Update) until convergence. If not, the algorithm has converged, and the final centroids represent the cluster centers.

5) Termination:

- The algorithm terminates once the centroids no longer change significantly or after a predetermined number of iterations.

2.1. Silhouette Index

The silhouette index is a commonly employed statistic in the evaluation of clustering outcomes, since it quantifies the degree of separation and compactness exhibited by the clusters. The metric quantifies the degree of clustering for each data point in relation to its neighbouring cluster, hence offering valuable insights about the suitability of the selected number of clusters and the success of the clustering algorithm. The silhouette index is a metric that falls within the range of -1 to 1, with higher values indicating clusters that are more well-defined. A number in proximity to 1 indicates the presence of distinct and independent clusters, whilst values in proximity to 0 show the existence of overlapping clusters. Negative values, on the other hand, reflect the possibility of erroneous assignment of data points to clusters. The silhouette index has become widely recognised and valued in the field, primarily because of its intuitive interpretation and its effectiveness in accommodating diverse cluster shapes and densities [33]–[36]. The silhouette index can be derived by utilising (3).

$$S_{(i)} = \frac{(b_{(i)} - a_{(i)})}{(\max\{a_{(i)}, b_{(i)}\})} \quad (3)$$



Where, the variable $a_{(i)}$ represents the average dissimilarity between the i th object and all other objects within the same cluster. The variable $b_{(i)}$ represents the average dissimilarity between the i th object and all other objects in the nearest cluster. The values of $a_{(i)}$ and $b_{(i)}$ can be derived by utilising (4) and (5).

$$a_{(i)} = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i, j) \tag{4}$$

$$b_{(i)} = \min_{j \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \tag{5}$$

The variable $|C_I|$ represents the cardinality of cluster C_I , which denotes the number of data points belonging to that specific cluster. On the other hand, $d(i, j)$ represents the distance between two data points, i and j , within the cluster C_I .

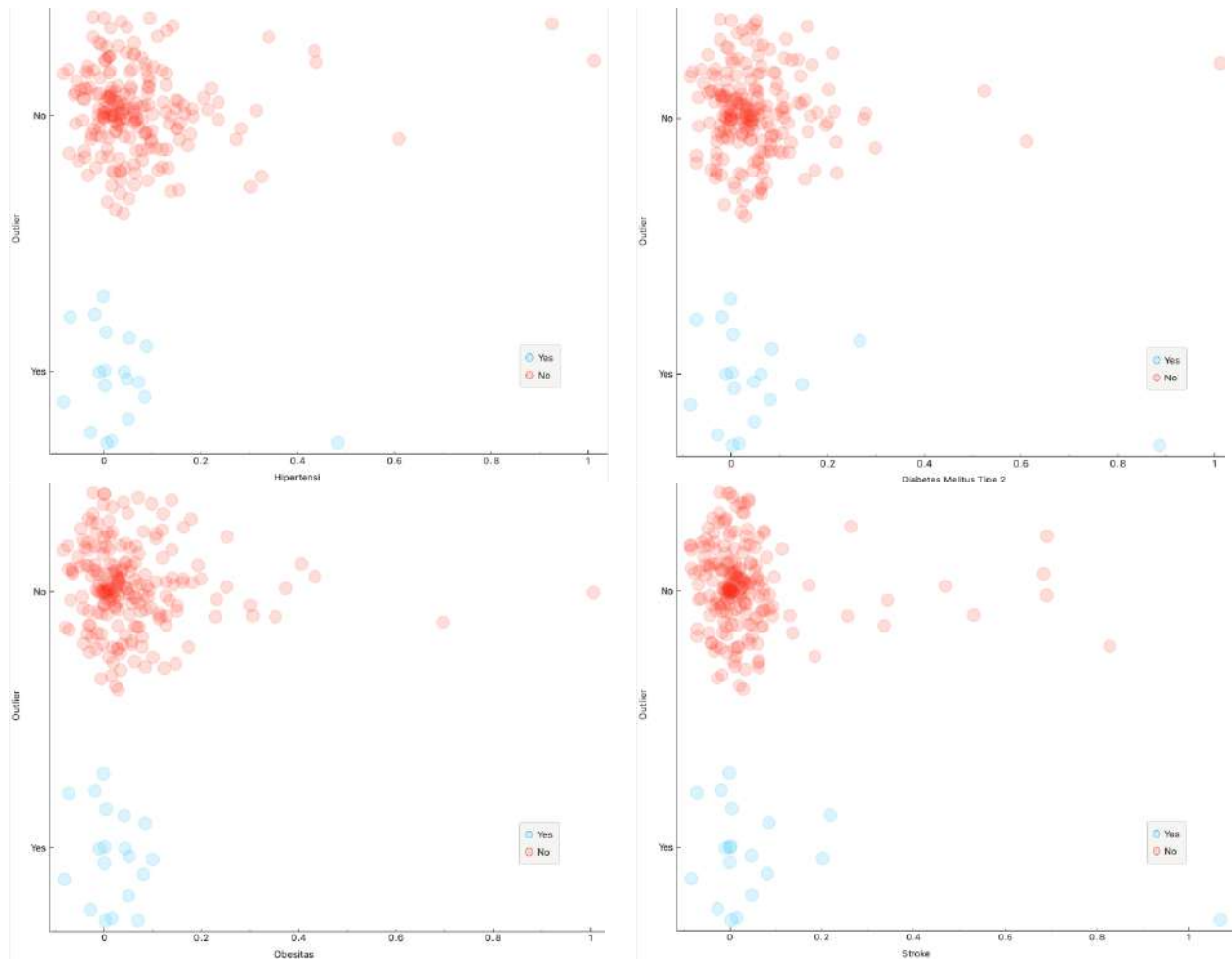


Figure 2. Results of Outlier Analysis for Some Features

3 RESULTS AND ANALYSIS

This section contains two subsections. The first subsection describes the findings of the conducted research, while the second section discusses the findings, including the interpretation of cluster outcomes.

3.1. Results

Outlier analysis is essential in the initial phases of data clustering, serving as a critical step to detect and assess occurrences that exhibit substantial deviation from the average. The main goal is to improve the precision and dependability of the following

categorization procedure. Within this particular context, the analysis fulfils a dual function, aiding in the division of data into two clearly defined scenarios.

In the first scenario, the clusters are carefully organised depending on the results of the outlier analysis. This strategy guarantees that the categorization procedure is carried out on a purified dataset, devoid of the impact of outliers. The purpose of this strategy is to identify and emphasise groups that demonstrate exceptional quality by carefully dealing with any exceptional data points. Through the process of isolating these clusters, researchers are able to extract more significant insights and provide more precise forecasts. In contrast, the second scenario entails the grouping of data without considering the outcomes of the outlier analysis. This methodology offers a method of comparison, enabling researchers to evaluate the influence of outliers on the clustering procedure. It assists in comprehending the degree to which outliers impact the overall grouping and aids in assessing the resilience of the clustering algorithm under various circumstances. In order to conduct a thorough analysis of outliers, various health-related attributes such as hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness were examined closely. The selection of these attributes encompasses a wide range of health-related criteria, guaranteeing a comprehensive evaluation of exceptional cases across multiple dimensions. All of the given health characteristics were found to be outliers, according to the results of the outlier analysis. Along with Pondok Ranji, Sepatan, Banjar, Sukadiri, Cisauk, Bhakti Jaya, Kresek, Mancak, Cihara, Ciledug, Rengas, Tunjung Teja, Cibitung, Angsana, Sindang Resmi, Cinangka, Anyar, Rajeg, and Carenang, the outliers were found in 19 different datasets or regions. Figure 2 displays a scatter plot that visualises the outliers for a particular feature.

Upon completion of the outlier analysis, the results are incorporated into the clustering procedure in the initial study scenario. This eliminates 19 data points from the entire regional dataset. The ideal number of clusters is determined using the average silhouette index value before starting the clustering process. The k-means widget in the Orange data mining framework allows for the emulation of an optimal number of clusters by specifying a defined range of cluster numbers. In the current study context, where dataset normalisation is lacking, the specified range extends from 2 to 8 clusters. The cluster distribution shows unevenness, as indicated by the incomplete attainment of the optimal average silhouette index for the comprehensive cluster set. The uneven distribution highlights the need for more improvement within the clustering architecture. Data normalisation plays a crucial role in improving the results of clustering. After the normalisation process, there is a noticeable increase in the average silhouette index value for each cluster. This rise demonstrates an enhanced level of unity and distinctiveness among the clusters. These findings demonstrate that normalising the data has a beneficial impact on enhancing the quality of the clustering outcomes. Despite attempts to normalise the data, the cluster arrangement that yields the highest silhouette index remains at $k = 2$. The current setup produces an average silhouette index of 0.81, indicating a strong level of distinction and unity among the clusters. The comparative silhouette indices for different cluster topologies, with values of k equal to 3, 4, 5, 6, 7, and 8, are 0.701, 0.664, 0.649, 0.456, 0.494, and 0.514, respectively. The values provide evidence for the exceptional quality and consistency of the $k = 2$ cluster setup. To summarise, the repeated incorporation of outlier analysis and subsequent clustering procedures is crucial in extracting detailed insights from complex datasets. The methodical process of refining the data, which includes removing outliers and normalising the data, enhances the optimisation of cluster quality. Using the silhouette index as a metric helps determine the optimal number of clusters, leading to a better understanding of the patterns in the dataset.

Table 1. Distribution of Cluster Member for Each k in the First Scenario Two

silhouette-NN	0,789	0,728	0,732	0,667	0,388	0,419	0,42
silhouette-Nr	0,812	0,701	0,664	0,649	0,456	0,494	0,514
Number of k	2	3	4	5	6	7	8
C1-Nr	202	193	14	6	157	12	13
C1-NN	10	191	193	185	1	134	134
C2-Nr	6	14	189	5	5	5	2
C2-NN	198	16	12	17	5	6	5
C3-Nr	0	1	4	1	7	20	153
C3-NN	0	1	1	1	122	1	1
C4-Nr	0	0	1	184	32	6	2
C4-NN	0	0	2	2	12	2	2
C5-Nr	0	0	0	12	1	158	5
C5-NN	0	0	0	3	62	3	1
C6-Nr	0	0	0	0	6	6	9
C6-NN	0	0	0	0	6	45	45
C7-Nr	0	0	0	0	0	1	23
C7-NN	0	0	0	0	0	17	17
C8-Nr	0	0	0	0	0	0	1
C8-NN	0	0	0	0	0	0	3

Initially, the clustering method yields that 202 regions are assigned to the first cluster (C1), while the remaining 6 regions are allocated to the second cluster (C2). More precisely, there are a total of 193 regions grouped as members of category C1, 14 regions assigned to category C2, and one region allocated to category C3 in clusters where the value of k is equal to 3. Table 1 presents the detailed distribution of cluster members for each k , categorised based on their normalisation state (normalised indicated as Nr and non-normalised denoted as NN). The data presented in Table 1 indicates that when the value of k decreases, there is a stronger inclination for individuals from one cluster to merge with individuals from other clusters that have a larger



number of members, resulting in a reduction in the number of members in the remaining clusters. This discovery highlights the correlation between the selected number of clusters and the distribution patterns of regional members. Furthermore, the arrangement of cluster participants, regardless of whether it is standardised or not, demonstrates the enduring presence of regional clustering. However, it is important to mention that the occurrence of regional clustering only undergoes a slight change in its location. Although there may be differences in data normalisation, the overall trend of regional clustering stays generally stable. Table 1 presents the outcomes of the clustering study, illustrating the variations in the number of clusters selected and the distribution patterns of regional members over time. The phenomenon of regional agglomeration is evident in both normalised and non-normalised datasets, and any disparities in their manifestation are relatively inconsequential.

In the second clustering scenario, the dataset is straightaway processed using the k-means method without any prior outlier data removal. The cluster outcomes show a pattern where, as the value of k lowers, the cluster member areas become more concentrated within one cluster or specific clusters. Significantly, the number of cluster members in C1 (with data normalisation indicated as Nr) exceeds that of C2 when k is equal to 2. In contrast, if the data is not normalised, a reverse correlation arises, where cluster C2 includes a greater number of members compared to cluster C1. This trend remains consistent for the majority of other k values. The distribution of cluster members shows inequality between k = 2 and k = 8. However, as k increases, there is a noticeable pattern where the distribution of cluster members becomes more spread out, although the overall pattern of members clustering together remains unchanged. In the second clustering scenario, Table 2 provides detailed information about the distribution of cluster members for each value of k. Furthermore, Figure 3 presents a graphical depiction of the clustering results, differentiating between normalised (a) and non-normalised data (b), respectively.

Table 2. Distribution of Cluster Member for Each k for the Second Scenario

silhouette-NN	0,769	0,682	0,595	0,613	0,566	0,57	0,467
silhouette-Nr	0,805	0,724	0,666	0,696	0,404	0,466	0,463
Number of k	2	3	4	5	6	7	8
C1-Nr	218	19	12	17	16	6	41
C1-NN	18	36	58	13	7	22	21
C2-Nr	9	206	205	1	1	166	2
C2-NN	209	188	151	2	50	7	7
C3-Nr	0	2	2	2	2	7	167
C3-NN	0	3	2	151	6	139	101
C4-Nr	0	0	8	204	147	1	1
C4-NN	0	0	16	55	139	49	38
C5-Nr	0	0	0	3	3	2	3
C5-NN	0	0	0	6	2	6	6
C6-Nr	0	0	0	0	58	44	5
C6-NN	0	0	0	0	23	2	2
C7-Nr	0	0	0	0	0	1	1
C7-NN	0	0	0	0	0	2	2
C8-Nr	0	0	0	0	0	0	7
C8-NN	0	0	0	0	0	0	50

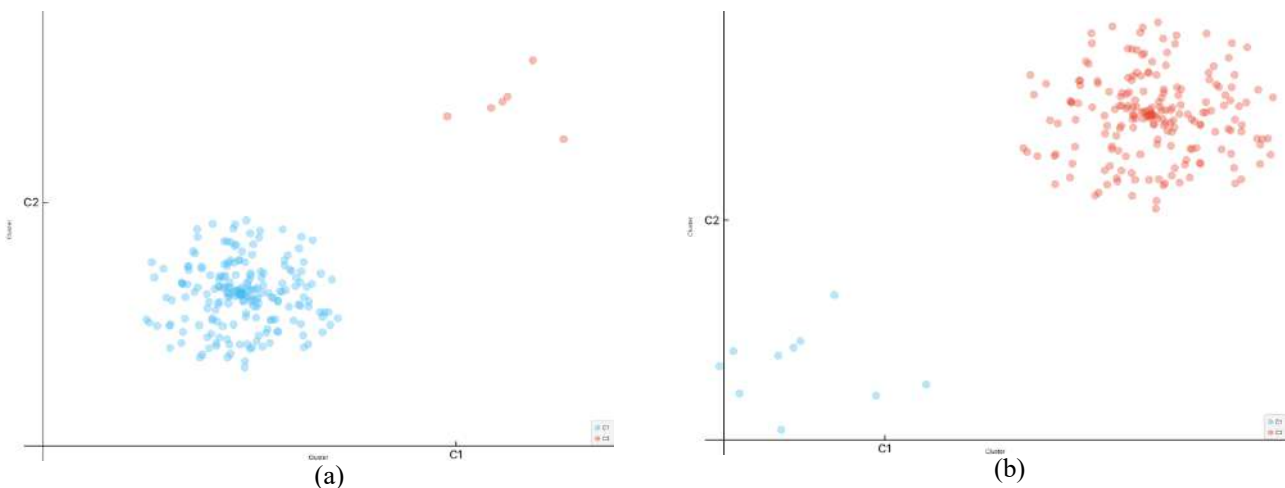


Figure 3. The difference between cluster findings with normalisation (a) and without normalisation (b).

3.2. Discussion

The cluster analysis of the two scenarios produces significant results, especially when evaluating the average silhouette index values. The initial analysis reveals that the highest average silhouette index value is observed when the value of k is set to 2. This finding serves as the fundamental criterion for determining the value of k in the following analyses conducted in this

inquiry. Each member of the cluster at $k = 2$ demonstrates a silhouette value near 1, suggesting a strong level of cohesiveness and isolation within the clusters. Aligning with previous studies conducted by [33] and [37], the Silhouette Index is employed to determine the most suitable number of clusters. According to their statement, this index can be utilised to ascertain the most suitable number of clusters prior to clustering. The results, especially after data normalisation, confirm that the clustering results with $k = 2$ have higher validity compared to other configurations. Previous studies have found that normalising data before applying machine learning approaches can enhance the effectiveness of both clustering [38], [39], and classification [40], [41], according to research findings. Hence, it is unsurprising that in this study, the application of normalisation has the capacity to enhance cluster validity, thereby enabling it to ascertain the most suitable number of clusters.

Confirming the cluster results with a value of $k = 2$ shows that 202 regions are grouped under cluster C1, while the rest are assigned to cluster C2. By lining up these cluster results with infographics, it's clear that cluster C2 includes areas with almost all 10 types of noncommunicable diseases that were looked at in this study. Cluster C1 exhibits the largest frequency of individuals with non-communicable diseases, particularly for the five most common conditions: diabetes, hypertension, obesity, stroke, and cataracts. In contrast, the 202 locations in cluster C1 have a remarkably low occurrence of diseases such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

The results of this study emphasise the urgent requirement for focused attention in up to 202 regions, especially those vulnerable to the occurrence of the five illnesses: diabetes, hypertension, obesity, stroke, and cataracts. At the same time, the six places that make up cluster C2 need immediate attention because they have high rates of the five diseases listed above, as well as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This investigation confirms the crucial significance of customised healthcare interventions that are based on specific regional health profiles. The focus on clusters C1 and C2 enables the implementation of specific initiatives that recognise the distinct patterns of non-communicable illnesses in each cluster. Public health programmes must prioritise addressing the distinct healthcare requirements of these clusters by customising interventions to tackle prevalent disorders and alleviate the burden of diseases in the designated locations.

4 CONCLUSION

To summarise, the research results emphasise the urgent requirement for specific healthcare interventions in the province of Banten, specifically in relation to non-communicable diseases (NCDs). Using the k-means algorithm to do clustering analysis on NCD markers for 208 regions shows how important it is for each part of the province to have its own unique health profile. The clustering results, especially when k is set to 2, indicate that there are 202 regions that require urgent attention because of the high occurrence of diabetes, hypertension, obesity, stroke, and cataracts. Besides the main diseases listed above, the province's other 28 regions deal with a wider range of noncommunicable diseases (NCDs), such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This extensive analysis of regional health profiles establishes a basis for focused public health initiatives, highlighting the imperative need for government action in specified locations.

Nevertheless, it is imperative to recognise the constraints of the clustering methodology utilised in this research. Although the k-means algorithm efficiently classifies regions into two main clusters based on NCD markers, it does not provide information about the spatial proximity of NCD sufferers across regions. Given this constraint, it is clear that further research using other methods is necessary to thoroughly investigate the pattern of closeness among individuals with non-communicable diseases in different areas. This necessitates an intricate method of comprehending not just the frequency of particular illnesses in each group but also the interaction of geographical elements that impact healthcare dynamics.

In the future, it is important for research to concentrate on improving techniques for capturing the spatial linkages and proximity patterns among individuals with non-communicable diseases (NCDs). This deeper comprehension will enable the implementation of more accurate and customised healthcare plans, guaranteeing that interventions are customised to the individual requirements of each location. Integrating spatial analyses into future research will enhance the effectiveness of tackling the intricate terrain of non-communicable diseases in the province of Banten.

REFERENCES

- [1] S. Bhattacharya, P. Heidler, and S. Varshney, "Incorporating neglected non-communicable diseases into the national health program—A review," *Front Public Health*, vol. 10, Jan. 2023, doi: 10.3389/fpubh.2022.1093170.
- [2] M. F. Owusu, J. Adu, B. A. Dortey, S. Gyamfi, and E. Martin-Yeboah, "Exploring health promotion efforts for non-communicable disease prevention and control in Ghana," *PLOS Global Public Health*, vol. 3, no. 9, p. e0002408, Sep. 2023, doi: 10.1371/journal.pgph.0002408.
- [3] A. Odunyemi, T. Rahman, and K. Alam, "Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19," *BMC Public Health*, vol. 23, no. 1, p. 1563, Aug. 2023, doi: 10.1186/s12889-023-16498-7.
- [4] H. G. A. S. Samarasinghe *et al.*, "Barriers to Accessing Medical Services and Adherence to Recommended Drug Regimens among Patients with Non-Communicable Diseases: A Study at Divisional Hospital Thalangama, Sri Lanka," in *IECN 2023*, Basel Switzerland: MDPI, Oct. 2023, p. 14. doi: 10.3390/IECN2023-15526.



- [5] Dinas Kesehatan Provinsi Banten, "Profil Kesehatan Provinsi Banten Tahun 2021," Serang, 2021. Accessed: Nov. 24, 2023. [Online]. Available: <https://dinkes.bantenprov.go.id/storage/dinkes/files/1109/Profil%20Kesehatan/Profil%20Kesehatan%20Banten%20Tahun%202021.pdf>
- [6] K. Stelin Maliangkay Fakultas Ilmu Kesehatan, K. Masyarakat, U. Rahma Fakultas Ilmu Kesehatan, S. Putri Fakultas Ilmu Kesehatan, and N. Dwi Istanti Fakultas Ilmu Kesehatan, "Analisis Peran Promosi Kesehatan Dalam Mendukung Keberhasilan Program Pencegahan Penyakit Tidak Menular Di Indonesia," *Jurnal Medika Nusantara*, vol. 1, no. 2, 2023.
- [7] H. B. H. Akbar, and S. Sarman, "Pencegahan Penyakit Tidak Menular Melalui Edukasi Cerdik Pada Masyarakat Desa Moyag Kotamobagu," *Abdimas Universal*, vol. 3, no. 1, pp. 83–87, Apr. 2021, doi: 10.36277/abdimasuniversal.v3i1.94.
- [8] R. Gupta, K. Gaur, and C. V. S. Ram, "Emerging trends in hypertension epidemiology in India," *J Hum Hypertens*, vol. 33, no. 8, pp. 575–587, 2019, doi: 10.1038/s41371-018-0117-3.
- [9] C. Antza, G. Kostopoulos, S. Mostafa, K. Nirantharakumar, and A. Tahrani, "The links between sleep duration, obesity and type 2 diabetes mellitus," *Journal of Endocrinology*, vol. 252, no. 2, pp. 125–141, 2022, doi: 10.1530/JOE-21-0155.
- [10] L. Wang, S. Wang, Q. Zhang, C. He, C. Fu, and Q. Wei, "The role of the gut microbiota in health and cardiovascular diseases," *Molecular Biomedicine*, vol. 3, no. 1, p. 30, 2022, doi: 10.1186/s43556-022-00091-2.
- [11] A. Al Samarraie, M. Pichette, and G. Rousseau, "Role of the Gut Microbiome in the Development of Atherosclerotic Cardiovascular Disease," *Int J Mol Sci*, vol. 24, no. 6, 2023, doi: 10.3390/ijms24065420.
- [12] R. T. Chlebowski *et al.*, "Weight loss and breast cancer incidence in postmenopausal women," *Cancer*, vol. 125, no. 2, pp. 205–212, 2019, doi: <https://doi.org/10.1002/cncr.31687>.
- [13] M. Ellingjord-Dale *et al.*, "Long-term weight change and risk of breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) study," *Int J Epidemiol*, vol. 50, no. 6, pp. 1914–1926, Dec. 2021, doi: 10.1093/ije/dyab032.
- [14] E. Kričković, T. Lukić, and D. Jovanović-Popović, "Geographic Medical Overview of Noncommunicable Diseases (Cardiovascular Diseases and Diabetes) in the Territory of the AP Vojvodina (Northern Serbia)," *Healthcare*, vol. 11, no. 1, p. 48, Dec. 2022, doi: 10.3390/healthcare11010048.
- [15] T. B. Darikwa and S. O. Manda, "Spatial Co-Clustering of Cardiovascular Diseases and Select Risk Factors among Adults in South Africa," *Int J Environ Res Public Health*, vol. 17, no. 10, p. 3583, May 2020, doi: 10.3390/ijerph17103583.
- [16] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning," *Applied Sciences*, vol. 13, no. 3, p. 1509, Jan. 2023, doi: 10.3390/app13031509.
- [17] L. Zhang, G. Yang, and X. Li, "Mining sequential patterns of PM2.5 pollution between 338 cities in China," *J Environ Manage*, vol. 262, p. 110341, May 2020, doi: 10.1016/j.jenvman.2020.110341.
- [18] D. Majcherek, M. A. Weresa, and C. Ciecierski, "A Cluster Analysis of Risk Factors for Cancer across EU Countries: Health Policy Recommendations for Prevention," *Int J Environ Res Public Health*, vol. 18, no. 15, p. 8142, Jul. 2021, doi: 10.3390/ijerph18158142.
- [19] M. A. Emon *et al.*, "Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms," *Sci Rep*, vol. 10, no. 1, p. 19097, Nov. 2020, doi: 10.1038/s41598-020-76200-4.
- [20] J. Prakash, V. Wang, R. E. Quinn, and C. S. Mitchell, "Unsupervised Machine Learning to Identify Separable Clinical Alzheimer's Disease Sub-Populations," *Brain Sci*, vol. 11, no. 8, p. 977, Jul. 2021, doi: 10.3390/brainsci11080977.
- [21] S. Bhattacharjee *et al.*, "Cluster Analysis: Unsupervised Classification for Identifying Benign and Malignant Tumors on Whole Slide Image of Prostate Cancer," in *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*, IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/IPAS55744.2022.10052952.
- [22] Y. Jiang *et al.*, "Unsupervised machine learning based on clinical factors for the detection of coronary artery atherosclerosis in type 2 diabetes mellitus," *Cardiovasc Diabetol*, vol. 21, no. 1, p. 259, Nov. 2022, doi: 10.1186/s12933-022-01700-8.
- [23] G. Sarveswaran, V. Kulothungan, and P. Mathur, "Clustering of noncommunicable disease risk factors among adults (18–69 years) in rural population, South-India," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1005–1014, Sep. 2020, doi: 10.1016/j.dsx.2020.05.042.
- [24] S. V. Rocha *et al.*, "Cluster analysis of risk factors for chronic non-communicable diseases in elderly Brazilians: population-based cross-sectional studies in a rural town," *Research, Society and Development*, vol. 10, no. 17, p. e18101724202, Dec. 2021, doi: 10.33448/rsd-v10i17.24202.
- [25] R. Uddin, E.-Y. Lee, S. R. Khan, M. S. Tremblay, and A. Khan, "Clustering of lifestyle risk factors for non-communicable diseases in 304,779 adolescents from 89 countries: A global perspective," *Prev Med (Baltim)*, vol. 131, p. 105955, Feb. 2020, doi: 10.1016/j.ypmed.2019.105955.
- [26] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 337–350, Mar. 2022, doi: 10.30812/matrik.v21i2.1411.

- [27] H. Hairani, D. Susilowati, I. Puji Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 275–282, Mar. 2022, doi: 10.30812/matrik.v21i2.1542.
- [28] S. Annas, B. Poerwanto, S. Sapriani, and M. F. S., "Implementation of K-Means Clustering on Poverty Indicators in Indonesia," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 257–266, Mar. 2022, doi: 10.30812/matrik.v21i2.1289.
- [29] Y. Zhao and X. Zhou, "K-means Clustering Algorithm and Its Improvement Research," *J Phys Conf Ser*, vol. 1873, no. 1, p. 012074, Apr. 2021, doi: 10.1088/1742-6596/1873/1/012074.
- [30] M. Darwis, L. H. Hasibuan, M. Firmansyah, N. Ahady, and R. Tiaharyadini, "Implementation of K-Means Clustering Algorithm in Mapping the Groups of Graduated or Dropped-out Students in the Management Department of the National University," *JISA (Jurnal Informatika dan Sains)*, vol. 4, no. 1, pp. 1–9, Jun. 2021.
- [31] A. R. Danurisa and J. Heikal, "Customer Clustering Using the K-Means Clustering Algorithm in the Top 5 Online Marketplaces in Indonesia," *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, vol. 5, no. 3, pp. 24287–24301, Aug. 2022, doi: 10.33258/birci.v5i3.6450.
- [32] A. Chaerudin, D. T. Murdiansyah, and M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Ind. Journal on Computing*, vol. 6, no. 1, pp. 53–60, 2020, doi: 10.34818/indojc.2021.6.1.547.
- [33] A. Dudek, "Silhouette Index as Clustering Evaluation Tool," 2020, pp. 19–33. doi: 10.1007/978-3-030-52348-0_2.
- [34] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/e23060759.
- [35] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Techno.Com*, vol. 20, no. 2, pp. 186–197, May 2021, doi: 10.33633/tc.v20i2.4556.
- [36] S. Paembonan, H. Abduh, and K. Kunci, "Penerapan Metode Silhouette Coefficient Untuk Evaluasi Clustering Obat Clustering; K-means; Silhouette coefficient," *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 6, no. 2, pp. 48–54, 2021, [Online]. Available: <https://ojs.unanda.ac.id/index.php/jiit/index>
- [37] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 174–182, Apr. 2023, doi: 10.3991/ijoe.v19i04.37059.
- [38] T. Li, Y. Ma, and T. Endoh, "Normalization-Based Validity Index of Adaptive K-Means Clustering for Multi-Solution Application," *IEEE Access*, vol. 8, pp. 9403–9419, 2020, doi: 10.1109/ACCESS.2020.2964763.
- [39] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *J Phys Conf Ser*, vol. 1566, no. 1, p. 012112, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012112.
- [40] H. A. Ahmed, P. J. Muhammad Ali, A. K. Faeq, and S. M. Abdullah, "An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method," *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 2, pp. 29–37, Sep. 2022, doi: 10.14500/aro.10970.
- [41] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, vol. 10, no. 11, p. 1942, Jun. 2022, doi: 10.3390/math10111942.

Revised Article



Participants [Edit](#)

Tb Ai Munandar (tbaimunandar)

Messages

Note

From

Dear Editor,

tbaimunandar

We have corrected the article according to feedback. The revised article is as attached.

Sep 19

Best regards,

 [tbaimunandar, Author, Rev - 3352-Article Text-17764-1-18-20230914.docx](#)

Add Message

Initial Screening



Participants

Tb Ai Munandar (tbaimunandar)

Messages

Note

From

Dear authors, after initial proofreading, this article still does not meet the template and style guideline. Here we attach a word file that has comments available to fix the manuscript according to template.
Thank you.

dindalesta
Sep 14

 [dindalesta, Journal editor, 3352-Tb Ai Munandar--.docx](#)

▶ Dear Editor,

We have corrected the article according to feedback. The revised article is as attached.

tbaimunandar
Sep 19

Best regards,

 [tbaimunandar, Author, Rev - 3352-Article Text-17764-1-18-20230914.docx](#)

Add Message



STATEMENT OF MANUSCRIPT AUTHENTICITY

The undersigned declares that:

Manuscript title:

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Authors:

1. Name : Tb Ai Munandar

e-mail : tbaimunandar@gmail.com

2. Name : Ajif Yunizar Yusuf Pratama

e-mail : ajifpratama86@gmail.com

Name and address of 1st author (representative) for correspondence:

Name : Tb Ai Munandar

Address : Universitas Bhayangkara Jakarta Raya
Jl. Raya Perjuangan Bekasi Utara, Kota Bekasi, Jawa Barat 17121,
Indonesia

Phone Number : 081384512710

E-mail : tbaimunandar@gmail.com

- The article above is an original manuscript, the author's work, and is not a plagiarism of other people's articles or scientific works.
- The article has not been published or submitted for publication in any other journal or media.
- If this statement is untrue in the future, the author is willing to accept any sanctions following applicable laws and regulations.

Jakarta, 23 Januari 2024

1st Author,

Tb. Ai Munandar

2nd Author,

Ajif Yunizar Yusuf Pratama

Editorial Address

Universitas Bumigora, Jl. Ismail Marzuki Mataram, NTB

E-mail: matrik@universitasbumigora.ac.id

Phone: 0859 3308 3240



STATEMENT OF PUBLICATION CONSENT

I, the undersigned:

Name (1st Author) : Tb. Ai Munandar

Affiliation : Universitas Bhayangkara Jakarta Raya

Phone Number : 081384512710

Email : tbaimunandar@gmail.com

With this, I declare my **WILLINGNESS/UNWILLINGNESS*** for the publication of the paper titled:

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

In the Jurnal Manajemenn, Teknik Informatika dan Rekayasa Komputer / MATRIK: Vol 23 No. 2 2024

I affirm that the academic work mentioned above is free from plagiarism and duplicate publication (it has not been previously published in any media). It will not be republished in other journals, books, or magazines.

I am also willing to cover the publication fees for the Journal of Management, Information Technology, and Computer Engineering / MATRIK, following the applicable regulations and rules**).

This statement is made voluntarily and in good health, without coercion from any party, and may be used as necessary.

Jakarta, 23 January 2024

Statement Maker
1st Author,



Tb. Ai munandar

2nd Author,

Ajif Yunizar Yusuf Pratama

Note:

*: Cross out those that do not fit (adjust accordingly)

.: Paid after the article is declared **READY TO PUBLISH. Publication fees can be seen on the MATRIK Journal website.

Editorial Address

Universitas Bumigora, Jl. Ismail Marzuki Mataram, NTB

E-mail: matrik@universitasbumigora.ac.id

Phone: 0859 3308 3240

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

By Matrik Journal

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

Tb Ai Munandar, Ajif Yunizar Yusuf Pratama
Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia

Article Info

Article history:

Received September 05, 2023

Revised November 13, 2023

Accepted January 05, 2024

Keywords:

Clustering

k-Means

Non-Communicable Diseases

Regional Clustering

Silhouette Index

ABSTRACT

Noncommunicable diseases (NCDs) have become a global threat to public health, necessitating a comprehensive understanding of their geographic and epidemiological distribution to devise appropriate interventions. This study aims to cluster Banten Province areas based on NCDs profiles using the unsupervised learning technique. The method used in this study is the k-means algorithm for grouping types of non-communicable diseases based on region. The processing and formalisation of NCDs prevalence data from various health sources preceded cluster analysis using the k-means clustering algorithm. This research is categorised into two scenarios: the first involves clustering data obtained from outlier analysis, while the second excludes any outliers. The objective is to observe disparities in regional clustering outcomes by categorising non-communicable diseases according to these two scenarios. The silhouette index is used to determine the validity of cluster results. These findings are analysed to determine the geographic and socioeconomic patterns associated with each cluster's NCDs profile. Based on the mean silhouette index value of 0.812, the results indicate that the sum of $k = 2$ in the k-means algorithm is the optimal cluster result. Five non-communicable diseases, namely diabetes, hypertension, obesity, stroke, and cataracts, necessitate significant focus in the first cluster (C1), where 202 regions were grouped. Six regions belong to the second cluster (C2), which includes areas that are not only susceptible to the five non-communicable diseases in cluster C1 but also to breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

Copyright ©2022 The Authors.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Tb Ai Munandar, +6281384512710

Faculty of Computer Science, Informatics Department,
Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia.

Email: tbaimunandar@gmail.com

How to Cite:

T. A. Munandar and A. Y. Y. Pratama, "Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm", *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 2, pp. 285-296, March, 2024.

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

Journal homepage: <https://journal.universitasbumigora.ac.id/index.php/matrik>

1. INTRODUCTION

Non-communicable diseases (NCDs) comprise a collection of chronic health conditions that do not spread via direct human-to-human contact and are the leading cause of cancer-related deaths worldwide. This grouping contains a variety of illnesses, including chronic respiratory conditions, diabetes, cardiovascular disease, and cancer, which collectively impose a significant burden on global health. The gravity of non-communicable diseases (NCDs) is emphasised by the World Health Organisation (WHO), which estimates that these conditions are responsible for around 41 million deaths per year, or 71% of all casualties worldwide [1–4]. The development of non-communicable diseases (NCDs) is strongly associated with detrimental lifestyle decisions, which include unequal dietary patterns, a lack of physical activity, and the use of tobacco and alcohol. Due to urbanisation, evolving behaviours, and rapid population growth, the Indonesian province of Banten faces substantial challenges associated with non-communicable diseases (NCDs). In 2021, Banten Provincial Health Service released data that indicates Banten Province has experienced an ongoing and consistent increase in the prevalence of diabetes, hypertension, and obesity. The substantial increase in the prevalence of non-communicable diseases presents a tough obstacle for the regional health infrastructure, significantly affecting the community's overall standard of living [5, 6]. To address their complexity, an in-depth examination of the patterns and distribution of non-communicable diseases (NCDs) in Banten is necessary. The current regional grouping system predicated on NCD categories may not be ideal when developing targeted interventions.

Applying unsupervised learning techniques, specifically the cluster method, to divide Banten Province into groups based on different types of non-communicable diseases (NCDs) is the point of this research. Clustering presents a more sophisticated methodology than conventional grouping techniques, enabling the detection of inherent groupings within the data that do not require predetermined labelling. This study seeks a more precise comprehension of the distribution of non-communicable diseases (NCDs) throughout Banten Province by delineating regional clusters according to different NCD categories. For informing health policies and directing medical interventions, precise and comprehensive data on the prevalence and distribution of non-communicable diseases (NCDs) in various regions of Banten Province are indispensable. By allowing health policymakers to customise more efficacious prevention and intervention strategies, the regional clustering inferred from this study has the potential to yield significant insights regarding the spatial distribution of non-communicable disease prevalence. The resulting regional groups will provide a solid foundation for developing targeted non-communicable disease (NCD) prevention programmes, letting policymakers and medical professionals pay close attention to the specific needs of each cluster. Using unsupervised learning techniques, particularly clustering, to illustrate the intricate picture of NCD prevalence, this study's primary objective is to aid in the improvement of public health strategies in Banten Province; by employing this novel methodology, the research endeavours to establish a strong groundwork for decision-making grounded in evidence, thereby promoting a more sophisticated and focused approach to addressing the complex issues presented by non-communicable diseases in the area.

To face the challenges of NCDs in the province of Banten, a comprehensive and integrated approach to grouping areas based on the nature of NCDs is required. Currently, regional groupings are typically determined by administration or geographic location without considering each region's unique health profile. We can identify patterns and relationships in Banten Province NCDs data using unsupervised learning techniques, such as the clustering method. This phase is essential for designing more targeted interventions and enhancing health services in regions affected by NCDs. Moreover, research conducted by [7] on the epidemiology of hypertension in India suggests that accelerated urbanisation may contribute to the rising prevalence of hypertension in urban areas. The prevalence of obesity tends to be higher in urban areas than rural areas, highlighting the significance of taking regional differences into account when managing NCDs. These results indicate that the social and economic changes that result from urbanisation can affect disease patterns in the region, and it is not inconceivable that this could occur in Banten Province. Noncommunicable diseases (NCDs) are a major public health concern because they contribute to the world's elevated mortality rate and disease burden. Numerous studies have been conducted in recent years to understand the risk factors and transmission patterns of NCDs and to identify more effective prevention strategies. A study by Chen provides the most recent data regarding the prevalence and risk factors of type 2 diabetes in various populations. The findings of this study shed light on the association between diet and physical activity and the risk of type 2 diabetes. Several studies have investigated the impact of diet in reducing the risk of cardiovascular disease as part of efforts to prevent NCDs. For instance, research found a correlation between adult fast-food consumption and increased insulin resistance. Additionally, recent research has emphasised the significance of optimal sleep patterns in reducing the risk of NCDs. A study by [8] discovered that adult sleep deprivation increases the risk of adiposity. To complement the effect of physical activity on the prevention of NCDs, Daskalopoulou (2021) conducted a meta-analysis which revealed that higher levels of physical activity are associated with lower risks of certain NCDs. Recent research has also focused on the environment's role in NCDs. The composition of the intestinal microbiota is associated with the development of cardiovascular disease, according to [9, 10]. In terms of NCDs prevention, understanding the function of the environment is crucial. Zimmet, in 2021, investigated the global repercussions of the diabetes epidemic and other NCDs. In addition to focusing on NCDs in the adult population, researchers have also studied the

disorder in children. At the same time, Samavat 2021 examined the association between a child's adult diet and their future risk of breast cancer [11, 12]. However, as mentioned earlier, a portion of the research places greater emphasis on discerning various intervention methods and enhancing healthcare provisions by utilising assessments of individuals afflicted with non-communicable diseases. The topic of regional-specific approaches to healthcare intervention and management has not yet been addressed. The issue of non-communicable illness spreading [39] remains unresolved to some extent due to a lack of information regarding treatment priorities specific to different regions. It is imperative to adopt a regional cluster-based approach to enhance health services and interventions in the future.

The distinction between the present and prior studies categorises non-communicable diseases based on geographical regions. By implementing interventions and making enhancements, healthcare quality can be significantly improved. Collectively, the most recent research provides valuable insights into the effective management of NCDS. Through a greater understanding of risk factors and prevention strategies, it is anticipated that targeted preventive measures can be implemented to lessen the burden of non-communicable diseases (NCDs) and improve public health as a whole. Noncommunicable diseases (NCDs) are a significant global health burden, and it is essential to understand the patterns and patterns of dissemination of these diseases in a specific region for prevention and appropriate management. The unsupervised learning method has become a valuable instrument in analysing health data, including grouping regions by NCD type. Several recent studies have utilised unsupervised learning techniques, such as clustering and cluster analysis, to identify geographic and epidemiological patterns of NCDs in different regions. A study by [13] grouped regions based on the type of NCDS in a country using cluster analysis. This study reveals how to identify the most and least burdened counties. A related study by [14] and [15] used spatial clustering and unsupervised learning to categorise regions based on the pattern of cardiovascular disease distribution. The results of this study indicate certain health clusters that can be targeted by interventions to reduce the risk of cardiovascular disease. In addition, unsupervised learning techniques have been employed to determine the connection between air pollution patterns, the urban spread of respiratory diseases and health interventions. Research by [16] found that cluster patterns from air pollution data and the incidence of respiratory disease are frequently interconnected. Recent research has also investigated the use of machine learning, including unsupervised learning, for grouping regions based on other categories of diseases, such as cancer, benign and malignant tumours, diabetes, and neurodegenerative diseases [17–21]. To address the challenges posed by NCDS at the regional level, an unsupervised learning algorithm was used to identify patterns of interrelationships between specific NCDS in specific regions of a country [22]. Also, at the same time, some researchers grouped regions based on the level of exposure to certain NCDS risk factors using an unsupervised learning method [23, 24]. Overall, the application of unsupervised learning has created opportunities to segment regions based on the type of NCDs and to gain a deeper understanding of disease transmission patterns; one of the most popular algorithms is k-means. Using this strategy, it is anticipated that prevention and intervention can be more effectively targeted based on the local community's health characteristics. The k-means algorithm is used for a variety of reasons. Apart from being frequently used in health research, it may also be used to segment promotional places in education, facilities, and teachers [25, 26] and group poverty indicators in a region [27]. This study aims to categorise non-communicable diseases based on geographical regions by analysing patient data obtained from public health institutions in Banten Province. This research is anticipated to significantly impact local government's ability to identify regional priorities for managing the development of non-communicable diseases. In addition, unsupervised learning methods, such as the k-means algorithm, can be utilised as an alternate strategy to analyse non-communicable disease data, making a valuable contribution to the health sector.

This paper is divided into four sections. The first section presents the introduction, which includes the problem's background, a literature review, research gaps, and systematic information about the article. The second section explains the study methodologies used, and the third section includes the research results and commentary. Meanwhile, the fourth portion is the conclusion, which contains the investigation findings.

2. RESEARCH METHOD

This study uses unsupervised learning as its research methodology to categorise regions within Banten Province according to the specific noncommunicable disease (NCD) type. Figure 1 provides a more detailed overview of the study stages.

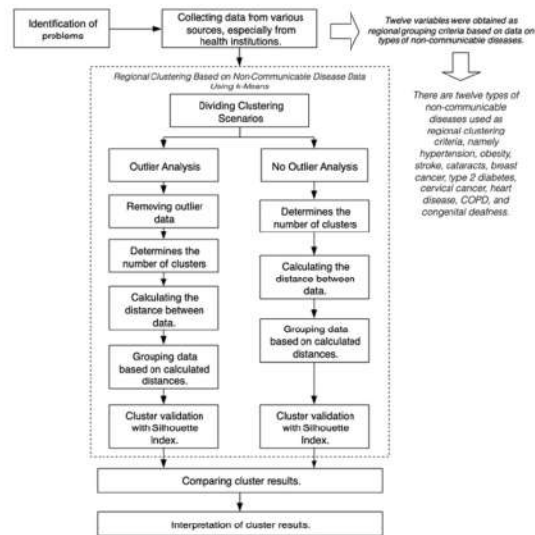


Figure 1. Research Stages

During the preliminary stage of the inquiry, data pertaining to the incidence of non-communicable diseases (NCDs) were gathered from several sources, encompassing public health institutions and hospitals situated within Banten Province. A total of 227 instances of regional information representing subdistricts were gathered due to the data-gathering process. The data is subsequently submitted to a preprocessing stage, wherein incomplete or fragmented data are repaired, and all variables are normalised to guarantee consistency in scale. Consequently, the strategy for categorising regions based on NCDs profiles involved using the k-means clustering algorithm, which falls under the umbrella of unsupervised learning techniques. Cluster analysis is conducted by determining the distance between the data points and the centroid of each cluster. This process involves combining regions that exhibit similar NCDs profiles into a cohesive cluster. The quality of the clusters was evaluated by performing validation of the cluster results using the silhouette index. The clustering criteria are determined by twelve specific non-communicable diseases, which include hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, COPD, and congenital deafness.

19

2.1. K-Means

The k-means algorithm is a commonly employed clustering technique that divides a dataset into separate groups according to their similarity. This is achieved through an iterative process of assigning data points to the nearest cluster centroid and adjusting the centroids to minimise the total sum of squared distances. The method's success in numerous domains can be attributed to its efficiency and simplicity despite the acknowledged drawbacks of being sensitive to initial centroid selection and prone to converging to local optima [28, 29]. The k-means algorithm is widely recognised for its prevalence in unsupervised learning. It is commonly employed in several domains, such as picture segmentation, customer segmentation, and anomaly detection. Although this method is extensive, practitioners must exercise caution when interpreting the findings. It is important to acknowledge that the outcomes can vary depending on the initialisation and scale factors [30, 31]. The phases of the k-means algorithm are as follows:

1) Initialization

- Choose the number of clusters, k .
- Initialize k centroids.

2) Assignment Step:

- For each data point in your dataset, calculate the distance (e.g., Euclidean distance) to each k centroid using Equation (1).

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (1)$$

1

- Assign the data point to the cluster whose centroid is closest to it. This forms k clusters.
- 3) Update Step:
- Calculate the new centroids of the clusters based on the data points assigned to each cluster. This is done by computing the mean of all data points within each cluster using Equation (2).

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} X_q \quad (2)$$

- 4) Convergence Check:
- Check if the centroids have changed significantly from the previous iteration.
 - If the centroids have changed significantly, repeat steps 2 and 3 (Assignment and Update) until convergence. If not, the algorithm has converged, and the final centroids represent the cluster centres.
- 5) Termination:
- The algorithm terminates once the centroids no longer change significantly or after a predetermined number of iterations.

2.2. Silhouette Index

The silhouette index is a commonly employed statistic in evaluating clustering outcomes since it quantifies the degree of separation and compactness exhibited by the clusters. The metric quantifies the degree of clustering for each data point in relation to its neighbouring cluster, hence offering valuable insights about the suitability of the selected number of clusters and the success of the clustering algorithm. The silhouette index is a metric that falls within the range of -1 to 1, with higher values indicating more well-defined clusters. A number in proximity to 1 indicates the presence of distinct and independent clusters, whilst values in proximity to 0 show the existence of overlapping clusters. Negative values, on the other hand, reflect the possibility of erroneous assignment of data points to clusters. The silhouette index has become widely recognised and valued in the field primarily because of its intuitive interpretation and its effectiveness in accommodating diverse cluster shapes and densities [32–35]. The silhouette index can be derived by utilising Equation (3).

$$S^{(i)} = \frac{b_{(i)} - a_{(i)}}{\max\{a_{(i)}, b_{(i)}\}} \quad (3)$$

Where the variable $a_{(i)}$ represents the average dissimilarity between the i th object and all other objects within the same cluster, the variable $b_{(i)}$ represents the average dissimilarity between the i th object and all other objects in the nearest cluster. The values of $a_{(i)}$ and $b_{(i)}$ can be derived by utilising Equations (4) and (5).

$$a_{(i)} = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \quad (4)$$

$$b_{(i)} = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \quad (5)$$

The variable $|C_I|$ represents the cardinality of cluster C_I , which denotes the number of data points belonging to that specific cluster. On the other hand, $d(i, j)$ represents the distance between two data points, i and j , within the cluster C_I .

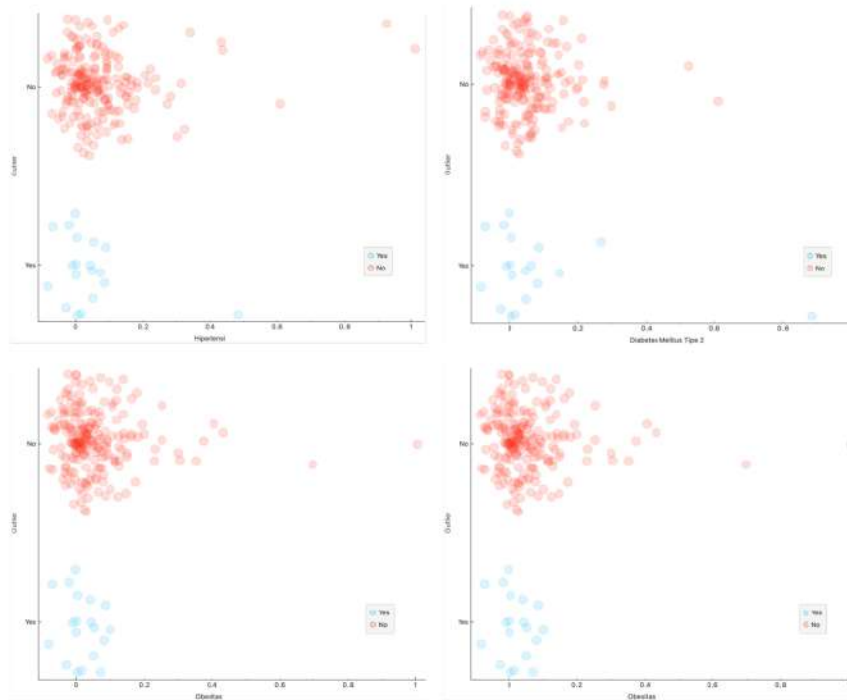


Figure 2. Results of Outlier Analysis for Some Features

3. RESULT AND ANALYSIS

This section contains two subsections. The first subsection describes the research findings, while the second section discusses the findings, including the interpretation of cluster outcomes.

3.1. Results

Outlier analysis is essential in the initial phases of data clustering, serving as a critical step to detect and assess occurrences that exhibit substantial deviation from the average. The main goal is to improve the precision and dependability of the following categorisation procedure. Within this particular context, the analysis fulfils a dual function, aiding in dividing data into two clearly defined scenarios.

In the first scenario, the clusters are carefully organised depending on the results of the outlier analysis. This strategy guarantees that the categorisation procedure is carried out on a purified dataset devoid of the impact of outliers. This strategy aims to identify and emphasise groups that demonstrate exceptional quality by carefully dealing with any exceptional data points. By isolating these clusters, researchers can extract more significant insights and provide more precise forecasts. In contrast, the second scenario entails data grouping without considering the outcomes of the outlier analysis. This methodology offers a comparison method, enabling researchers to evaluate the influence of outliers on the clustering procedure. It assists in comprehending the degree to which outliers impact the overall grouping and aids in assessing the resilience of the clustering algorithm under various circumstances. To conduct a thorough analysis [4] outliers, various health-related attributes such as hypertension, obesity, stroke, cataracts, breast cancer, type 2 diabetes, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness were examined closely. The selection of these attributes encompasses a wide range of health-related criteria, guaranteeing a comprehensive evaluation of exceptional cases across multiple dimensions. All of the given health characteristics were found to be outliers, according to the results of the outlier analysis. Along with Pondok Ranji, Sepatan, Banjar, Sukadiri, Cisauk, Bhakti Jaya, Kresek, Mancak, Cihara,

Ciledug, Rengas, Tunjung Teja, Cibitung, Angsana, Sindang Resmi, Cinangka, Anyar, Rajeg, and Carengang, the outliers were found in 19 different datasets or regions. Figure 2 displays a scatter plot that visualises the outliers for a particular feature.

Upon completion of the outlier analysis, the results are incorporated into the clustering procedure in the initial study scenario. This eliminates 19 data points from the entire regional dataset. The ideal number of clusters is determined using the average silhouette index value before starting the clustering process. The k-means widget in the Orange data mining framework allows for emulating an optimal number of clusters by specifying a defined range of cluster numbers. In the current study context, where dataset normalisation is lacking, the specified range extends from 2 to 8 clusters. The cluster distribution shows unevenness, as indicated by the incomplete attainment of the optimal average silhouette index for the comprehensive cluster set. The uneven distribution highlights the need for more improvement within the clustering architecture. Data normalisation plays a crucial role in improving the results of clustering. After the normalisation process, there is a noticeable increase in the average silhouette index value for each cluster. This rise demonstrates an enhanced level of unity and distinctiveness among the clusters. These findings demonstrate that normalising the data has a beneficial impact on enhancing the quality of the clustering outcomes. Despite attempts to normalise the data, the cluster arrangement that yields the highest silhouette index remains at $k = 2$. The current setup produces an average silhouette index of 0.81, indicating strong distinction and unity among the clusters. The comparative silhouette indices for different cluster topologies, with values of k equal to 3, 4, 5, 6, 7, and 8, are 0.701, 0.664, 0.649, 0.456, 0.494, and 0.514, respectively. The values provide evidence for the exceptional quality and consistency of the $k = 2$ cluster setup. To summarise, the repeated incorporation of outlier analysis and subsequent clustering procedures is crucial in extracting detailed insights from complex datasets. The methodical process of refining the data, which includes moving outliers and normalising the data, enhances the optimisation of cluster quality. Using the silhouette index as a metric helps determine the optimal number of clusters, leading to a better understanding of the patterns in the dataset.

Table 1. Distribution of Cluster Member for Each k in the First Scenario Two

silhouette-NN	0,789	0,728	0,732	0,667	0,388	0,419	0,42
silhouette-Nr	0,812	0,701	0,664	0,649	0,456	0,494	0,514
Number of k	2	3	4	5	6	7	8
C1-Nr	202	193	14	6	157	12	13
C1-NN	10	191	193	185	1	134	134
C2-Nr	6	14	189	5	5	5	2
C2-NN	198	16	12	17	5	6	5
C3-Nr	0	1	4	1	7	20	153
C3-NN	0	1	1	1	122	1	1
C4-Nr	0	0	1	184	32	6	2
C4-NN	0	0	2	2	12	2	2
C5-Nr	0	0	0	12	1	158	5
C5-NN	0	0	0	3	62	3	1
C6-Nr	0	0	0	0	6	6	9
C6-NN	0	0	0	0	6	45	45
C7-Nr	0	0	0	0	0	1	23
C7-NN	0	0	0	0	0	17	17
C8-Nr	0	0	0	0	0	0	1
C8-NN	0	0	0	0	0	0	3

Initially, the clustering method yields that 202 regions are assigned to the first cluster (C1), while the remaining 6 regions are allocated to the second cluster (C2). More precisely, there are 193 regions grouped as members of category C1, 14 regions assigned to category C2, and one region allocated to category C3 in clusters where the value of k equals 3. Table 1 presents the detailed distribution of cluster members for each k , categorised based on their normalisation state (normalised is indicated as Nr and non-normalised is denoted as NN). The data presented in Table 1 indicates that when the value of k decreases, there is a stronger inclination for individuals from one cluster to merge with individuals from other clusters with a larger number of members, reducing the number of members in the remaining clusters. This discovery highlights the correlation between the selected number of clusters and the distribution patterns of regional members. Furthermore, the arrangement of cluster participants, whether standardised or not, demonstrates the enduring presence of regional clustering. However, it is important to mention that the occurrence of regional clustering only undergoes a slight change in its location. Although there may be differences in data normalisation, the overall trend of regional clustering stays generally stable. Table 1 presents the outcomes of the clustering study, illustrating the variations in the number of clusters selected and the distribution patterns of regional members over time. Regional agglomeration is evident in both normalised and non-normalised datasets, and any disparities in their manifestation are relatively inconsequential.

In the second clustering scenario, the dataset is straightaway processed using the k-means method without any prior outlier data removal. The cluster outcomes show a pattern where, as the value of k lowers, the cluster member areas become more concentrated within one cluster or specific clusters. Significantly, the number of cluster members in C1 (with data normalisation indicated as Nr) exceeds that of C2 when k equals 2. In contrast, if the data is not normalised, a reverse correlation arises, where cluster C2 includes a greater number of members compared to cluster C1. This trend remains consistent for the majority of other k values. The distribution of cluster members shows inequality between k = 2 and k = 8. However, as k increases, there is a noticeable pattern where the distribution of cluster members becomes more spread out, although the overall pattern of members clustering together remains unchanged. In the second clustering scenario, Table 2 provides detailed information about the distribution of cluster members for each value of k. Furthermore, Figure 3 presents a graphical depiction of the clustering results, differentiating between normalised (a) and non-normalised data (b), respectively.

Table 2. Distribution of Cluster Member for Each k for the Second Scenario

silhouette-NN	0,769	0,682	0,595	0,613	0,566	0,57	0,467
silhouette-Nr	0,805	0,724	0,666	0,696	0,404	0,466	0,463
Number of k	2	3	4	5	6	7	8
C1-Nr	218	19	12	17	16	6	41
C1-NN	18	36	58	13	7	22	21
C2-Nr	9	206	205	1	1	166	2
C2-NN	209	188	151	2	50	7	7
C3-Nr	0	2	2	2	2	7	167
C3-NN	0	3	2	151	6	139	101
C4-Nr	0	0	8	204	147	1	1
C4-NN	0	0	16	55	139	49	38
C5-Nr	0	0	0	3	3	2	3
C5-NN	0	0	0	6	2	6	6
C6-Nr	0	0	0	0	58	44	5
C6-NN	0	0	0	0	23	2	2
C7-Nr	0	0	0	0	0	1	1
C7-NN	0	0	0	0	0	2	2
C8-Nr	0	0	0	0	0	0	7
C8-NN	0	0	0	0	0	0	50

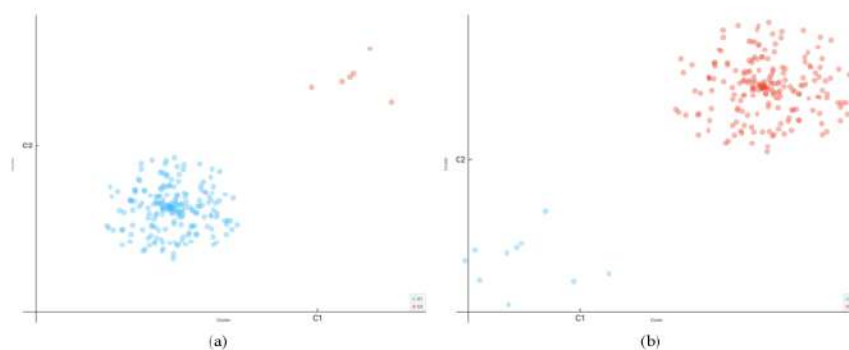


Figure 3. The difference between cluster findings with normalisation (a) and without normalisation (b)

3.2. Discussion

The cluster analysis of the two scenarios produces significant results, especially when evaluating the average silhouette index values. The initial analysis reveals that the highest average silhouette index value is observed when k is set to 2. This finding serves as the fundamental criterion for determining the value of k in the following analyses conducted in this inquiry. Each cluster member at k = 2 demonstrates a silhouette value near 1, suggesting strong cohesiveness and isolation within the clusters. Aligning with

previous studies conducted by [32] and [36], the silhouette Index determines the most suitable number of clusters. According to their statement, this index can be utilised to ascertain the most suitable number of clusters before clustering. The results, especially after data normalisation, confirm that the clustering results with $k = 2$ have higher validity than other configurations. Previous studies have found that normalising data before applying machine learning approaches can enhance the effectiveness of both clustering [37, 38], and classification [39, 40], according to research findings. Hence, it is unsurprising that in this study, the application of normalisation can enhance cluster validity, thereby enabling it to ascertain the most suitable number of clusters.

Confirming the cluster results with a value of $k = 2$ shows that 202 regions are grouped under cluster C1, while the rest are assigned to cluster C2. By lining up these cluster results with infographics, it's clear that cluster C2 includes areas with almost all 10 types of noncommunicable diseases that were looked at in this study. Cluster C1 exhibits the largest frequency of individuals with non-communicable diseases, particularly for the five most common conditions: diabetes, hypertension, obesity, stroke, and cataracts. In contrast, the 202 locations in cluster C1 have a remarkably low occurrence of diseases such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness.

The results of this study emphasise the urgent requirement for focused attention in up to 202 regions, especially those vulnerable to the occurrence of the five illnesses: diabetes, hypertension, obesity, stroke, and cataracts. At the same time, the six places that makeup cluster C2 need immediate attention because they have high rates of the five diseases listed above, as well as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This investigation confirms the crucial significance of customised healthcare interventions based on specific regional health profiles. The focus on clusters C1 and C2 enables the implementation of specific initiatives that recognise the distinct patterns of non-communicable illnesses in each cluster. Public health programmes must prioritise addressing the distinct healthcare requirements of these clusters by customising interventions to tackle prevalent disorders and alleviate the burden of diseases in the designated locations.

4. CONCLUSION

To summarise, the research results emphasise the urgent requirement for specific healthcare interventions in the province of Banten, specifically about non-communicable diseases (NCDs). Using the k-means algorithm to do clustering analysis on NCD markers for 208 regions shows how important each part of the province has its unique health profile. The clustering results, especially when k is set to 2, indicate that 202 regions require urgent attention because of the high occurrence of diabetes, hypertension, obesity, stroke, and cataracts. Besides the main diseases listed above, the province's other 28 regions deal with a wider range of noncommunicable diseases (NCDs), such as breast cancer, cervical cancer, heart disease, chronic obstructive pulmonary disease (COPD), and congenital deafness. This extensive analysis of regional health profiles establishes a basis for focused public health initiatives, highlighting the imperative need for government action in specified locations.

Nevertheless, it is imperative to recognise the constraints of the clustering methodology utilised in this research. Although the k-means algorithm efficiently classifies regions into two main clusters based on NCD markers, it does not provide information about the spatial proximity of NCD sufferers across regions. Given this constraint, it is clear that further research using other methods is necessary to thoroughly investigate the pattern of closeness among individuals with non-communicable diseases in different areas. This necessitates an intricate method of comprehending the frequency of particular illnesses in each group and the interaction of geographical elements that impact healthcare dynamics.

Future research needs to concentrate on improving techniques for capturing the spatial linkages and proximity patterns among individuals with non-communicable diseases (NCDs). This deeper comprehension will enable the implementation of more accurate and customised healthcare plans, guaranteeing that interventions are customised to the individual requirements of each location. Integrating spatial analyses into future research will enhance the effectiveness of tackling the intricate terrain of non-communicable diseases in the province of Banten.

5. DECLARATIONS

AUTHOR CONTRIBUTION

All authors contributed to the writing of this article.

FUNDING STATEMENT

This research was self-funded, and the authors did not receive any external financial support for the design, data collection, analysis, or interpretation of the study. All expenses related to this research were borne by the authors personally.

COMPETING INTEREST

The authors declare no conflict of interest in this article.

REFERENCES

- [1] S. Bhattacharya, P. Heidler, and S. Varshney, "Incorporating neglected non-communicable diseases into the national health programA review," *Frontiers in Public Health*, vol. 10, no. January, pp. 1–9, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.1093170/full>
- [2] M. F. Owusu, J. Adu, B. A. Dorte, S. Gyamfi, and E. Martin-Yeboah, "Exploring health promotion efforts for non-communicable disease prevention and control in Ghana." *PLOS Global Public Health*, vol. 3, no. 9, pp. 1–14, 2023. [Online]. Available: <https://dx.plos.org/10.1371/journal.pgph.0002408>
- [3] A. Odunyemi, T. Rahman, and K. Alam, "Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19," *BMC Public Health*, vol. 23, no. 1, pp. 1–12, 2023. [Online]. Available: <https://bmcpublihealth.biomedcentral.com/articles/10.1186/s12889-023-16498-7>
- [4] H. G. A. S. Samarasinghe, D. A. T. D. S. Ranasinghe, W. R. Jayasekara, S. A. A. D. Senarathna, J. D. P. M. Jayakody, P. M. Kalubovila, M. D. Edirisuriya, and N. S. A. S. N. Senarath, "Barriers to Accessing Medical Services and Adherence to Recommended Drug Regimens among Patients with Non-Communicable Diseases: A Study at Divisional Hospital Thalagama, Sri Lanka," in *IECN 2023*. MDPI, 2023, pp. 1–6. [Online]. Available: <https://www.mdpi.com/2673-9976/29/1/14>
- [5] K. S. Maliangkay, U. Rahma, S. Putri, and N. D. Istanti, "Analisis Peran Promosi Kesehatan Dalam Mendukung Keberhasilan Program Pencegahan Penyakit Tidak Menular Di Indonesia," *Jurnal Medika Nusantara*, vol. 1, no. 2, pp. 108–122, 2023.
- [6] H. B. H. Akbar, and S. Sarman, "Pencegahan Penyakit Tidak Menular Melalui Edukasi Cerdik Pada Masyarakat Desa Moyag Kotamobagu," *Abdimas Universal*, vol. 3, no. 1, pp. 83–87, 2021. [Online]. Available: <http://abdimasuniversal.uniba-bpn.ac.id/index.php/abdimasuniversal/article/view/94>
- [7] R. Gupta, K. Gaur, and C. V. S. Ram, "Emerging trends in hypertension epidemiology in India," *Journal of Human Hypertension*, vol. 33, no. 8, pp. 575–587, 2019. [Online]. Available: <https://www.nature.com/articles/s41371-018-0117-3>
- [8] C. Antza, G. Kostopoulos, S. Mostafa, K. Nirantharakumar, and A. Tahrani, "The links between sleep duration, obesity and type 2 diabetes mellitus," *Journal of Endocrinology*, vol. 252, no. 2, pp. 125–141, 2022. [Online]. Available: <https://joe.bioscientifica.com/view/journals/joe/252/2/JOE-21-0155.xml>
- [9] L. Wang, S. Wang, Q. Zhang, C. He, C. Fu, and Q. Wei, "The role of the gut microbiota in health and cardiovascular diseases," *Molecular Biomedicine*, vol. 3, no. 1, pp. 1–50, 2022. [Online]. Available: <https://link.springer.com/10.1186/s43556-022-00091-2>
- [10] A. A. Samarraie, M. Pichette, and G. Rousseau, "Role of the Gut Microbiome in the Development of Atherosclerotic Cardiovascular Disease," *International Journal of Molecular Sciences*, vol. 24, no. 6, pp. 1–17, 2023. [Online]. Available: <https://www.mdpi.com/1422-0067/24/6/5420>
- [11] R. T. Chlebowski, J. Luo, G. L. Anderson, W. Barrington, K. Reding, M. S. Simon, J. E. Manson, T. E. Rohan, J. Wactawski-Wende, D. Lane, H. Strickler, Y. MosaverRahmani, J. L. Freudenheim, N. Saquib, and M. L. Stefanick, "Weight loss and breast cancer incidence in postmenopausal women," *Cancer*, vol. 125, no. 2, pp. 205–212, 2019. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.31687>
- [12] M. Ellingjord-Dale, S. Christakoudi, E. Weiderpass, S. Panico, L. Dossus, A. Olsen, A. Tjønneland, R. Kaaks, M. B. Schulze, G. Masala, I. T. Gram, G. Skeie, A. H. Rosendahl, M. Sund, T. Key, P. Ferrari, M. Gunter, A. K. Heath, K. K. Tsilidis, and E. Riboli, "Long-term weight change and risk of breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) study," *International Journal of Epidemiology*, vol. 50, no. 6, pp. 1914–1926, 2022. [Online]. Available: <https://academic.oup.com/ije/article/50/6/1914/6182058>
- [13] E. Kričković, T. Lukić, and D. Jovanović-Popović, "Geographic Medical Overview of Noncommunicable Diseases (Cardiovascular Diseases and Diabetes) in the Territory of the AP Vojvodina (Northern Serbia)," *Healthcare*, vol. 11, no. 1, pp. 1–33, 2022. [Online]. Available: <https://www.mdpi.com/2227-9032/11/1/48>

- [14] T. B. Darikwa and S. O. Manda, "Spatial Co-Clustering of Cardiovascular Diseases and Select Risk Factors among Adults in South Africa," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, pp. 1–16, 2020. [Online]. Available: <https://www.mdpi.com/1660-4601/17/10/3583>
- [15] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning," *Applied Sciences*, vol. 13, no. 3, pp. 1–15, 2023. [Online]. Available: <https://www.mdpi.com/2076-3417/13/3/1509>
- [16] L. Zhang, G. Yang, and X. Li, "Mining sequential patterns of PM2.5 pollution between 338 cities in China," *Journal of Environmental Management*, vol. 262, no. March, pp. 1–8, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0301479720302760>
- [17] D. Majcherek, M. A. Weresa, and C. Ciecierski, "A Cluster Analysis of Risk Factors for Cancer across EU Countries: Health Policy Recommendations for Prevention," *International Journal of Environmental Research and Public Health*, vol. 18, no. 15, pp. 1–14, 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/15/8142>
- [18] M. A. Emon, A. Heinson, P. Wu, D. Domingo-Fernández, M. Sood, H. Vrooman, J.-C. Corvol, P. Scordis, M. Hofmann-Apitius, and H. Fröhlich, "Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020. [Online]. Available: <https://www.nature.com/articles/s41598-020-76200-4>
- [19] J. Prakash, V. Wang, R. E. Quinn, and C. S. Mitchell, "Unsupervised Machine Learning to Identify Separable Clinical Alzheimer's Disease Sub-Populations," *Brain Sciences*, vol. 11, no. 8, pp. 1–21, 2021. [Online]. Available: <https://www.mdpi.com/2076-3425/11/8/977>
- [20] S. Bhattacharjee, Y.-B. Hwang, R. I. Sumon, H. Rahman, D.-W. Hyeon, D. Moon, K. S. Carole, H.-C. Kim, and H.-K. Choi, "Cluster Analysis: Unsupervised Classification for Identifying Benign and Malignant Tumors on Whole Slide Image of Prostate Cancer," in *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*. IEEE, 2022, pp. 1–5. [Online]. Available: <https://ieeexplore.ieee.org/document/10052952/>
- [21] Y. Jiang, Z.-G. Yang, J. Wang, R. Shi, P.-L. Han, W.-L. Qian, W.-F. Yan, and Y. Li, "Unsupervised machine learning based on clinical factors for the detection of coronary artery atherosclerosis in type 2 diabetes mellitus," *Cardiovascular Diabetology*, vol. 21, no. 1, pp. 1–10, 2022. [Online]. Available: <https://cardiab.biomedcentral.com/articles/10.1186/s12933-022-01700-8>
- [22] G. Sarveswaran, V. Kulothungan, and P. Mathur, "Clustering of noncommunicable disease risk factors among adults (1869 years) in rural population, South-India," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1005–1014, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1871402120301624>
- [23] S. V. Rocha, S. C. de Oliveira, H. L. R. Munaro, C. F. R. Squarcini, B. M. P. Ferreira, F. de Oliveira Mendonça, and C. A. dos Santos, "Cluster analysis of risk factors for chronic non-communicable diseases in elderly Brazilians: population-based cross-sectional studies in a rural town," *Research, Society and Development*, vol. 10, no. 17, pp. 1–10, 2021. [Online]. Available: <https://rsdjournal.org/index.php/rsd/article/view/24202>
- [24] R. Uddin, E.-Y. Lee, S. R. Khan, M. S. Tremblay, and A. Khan, "Clustering of lifestyle risk factors for non-communicable diseases in 304,779 adolescents from 89 countries: A global perspective," *Preventive Medicine*, vol. 131, no. December, pp. 1–8, 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0091743519304384>
- [25] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 337–350, 2022. [Online]. Available: <https://journal.universitatumigora.ac.id/index.php/matrik/article/view/1411>
- [26] H. Hairani, D. Susilowati, I. P. Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 275–282, 2022. [Online]. Available: <https://journal.universitatumigora.ac.id/index.php/matrik/article/view/1542>

- [27] S. Annas, B. Poerwanto, S. Sapriani, and M. F. S., "Implementation of K-Means Clustering on Poverty Indicators in Indonesia," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 257–266, 2022. [Online]. Available: <https://journal.universitاسbumigora.ac.id/index.php/matrik/article/view/1289>
- [28] Y. Zhao and X. Zhou, "K-means Clustering Algorithm and Its Improvement Research," *Journal of Physics: Conference Series*, vol. 1873, no. 1, pp. 1–5, 2021. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1873/1/012074>
- [29] M. Darwis, L. H. Hasibuan, M. Firmansyah, N. Ahady, and R. Tiaharyadini, "Implementation of K-Means clustering algorithm in mapping the groups of graduated or dropped-out students in the Management Department of the National University," *JISA(Jurnal Informatika dan Sains)*, vol. 4, no. 1, pp. 1–9, 2021. [Online]. Available: <http://trilogi.ac.id/journal/ks/index.php/JISA/article/view/848>
- [30] A. R. Danurisa and J. Heikal, "Customer Clustering Using the K-Means Clustering Algorithm in the Top 5 Online Marketplaces in Indonesia," *Budapest International Research and Critics Insitute-Journal (BIRCI-Journal)*, vol. 5, no. 3, 2022.
- [31] A. Chacrudin, D. T. Murdiansyah, and M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Indonesia Journal on Computing (Indo-JC)*, vol. 6, no. 1, pp. 53–60, 2021.
- [32] A. Dudek, *Silhouette Index as Clustering Evaluation Tool*, 2020, pp. 19–33. [Online]. Available: http://link.springer.com/10.1007/978-3-030-52348-0_{_}2
- [33] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/6/759>
- [34] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Techno.Com*, vol. 20, no. 2, pp. 186–197, 2021. [Online]. Available: <http://publikasi.dinus.ac.id/index.php/technoc/article/view/4556>
- [35] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 6, no. 2, pp. 48–54, 2021. [Online]. Available: <https://ojs.unanda.ac.id/index.php/jiit/article/view/659>
- [36] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 174–182, 2023. [Online]. Available: <https://online-journals.org/index.php/i-joe/article/view/37059>
- [37] T. Li, Y. Ma, and T. Endoh, "Normalization-Based Validity Index of Adaptive K-Means Clustering for Multi-Solution Application," *IEEE Access*, vol. 8, pp. 9403–9419, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/8952702/>
- [38] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *Journal of Physics: Conference Series*, vol. 1566, no. 1, pp. 1–7, 2020. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1566/1/012112>
- [39] H. A. Ahmed, P. J. M. Ali, A. K. Faq, and S. M. Abdullah, "An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method," *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 2, pp. 29–37, 2022. [Online]. Available: <https://aro.koyauniversity.org/index.php/aro/article/view/970>
- [40] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, vol. 10, no. 11, p. 1942, 2022. [Online]. Available: <https://www.mdpi.com/2227-7390/10/11/1942>

Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm

ORIGINALITY REPORT

13%

SIMILARITY INDEX

PRIMARY SOURCES

- 1** garuda.kemdikbud.go.id 48 words — 1%
Internet
 - 2** Jaka Tirta Samudra, Rika Rosnelly, Zakarias Situmorang. "Comparative Analysis of SVM and Perceptron Algorithms in Classification of Work Programs", MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 2023 43 words — 1%
Crossref
 - 3** Relita Buaton, Solikhun Solikhun. "The Application of Numerical Measure Variations in K-Means Clustering for Grouping Data", MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 2023 43 words — 1%
Crossref
 - 4** www.coursehero.com 36 words — 1%
Internet
 - 5** eprints.uad.ac.id 32 words — 1%
Internet
 - 6** Tuli De, Tanuka Chattopadhyay, Asis Kumar Chattopadhyay. "Comparison Among Clustering and Classification Techniques on the Basis of Galaxy Data", Calcutta Statistical Association Bulletin, 2013 28 words — 1%
Crossref
-

-
- 7 Grandianus Seda Mada, Maria Julieta Esperanca Naibili, Siprianus Septian Manek, Estevania Daonce Mau, Wasim Raza. "Application of Mamdani's Fuzzy Inference System in the Diagnosis of Pre-eclampsia", Jurnal Varian, 2023
Crossref 26 words — < 1%
-
- 8 Adalakun Odunyemi, Taslima Rahman, Khurshid Alam. "Economic burden of non-communicable diseases on households in Nigeria: evidence from the Nigeria living standard survey 2018-19", BMC Public Health, 2023
Crossref 25 words — < 1%
-
- 9 mdpi-res.com
Internet 18 words — < 1%
-
- 10 export.arxiv.org
Internet 17 words — < 1%
-
- 11 www.mdpi.com
Internet 17 words — < 1%
-
- 12 Sachi Khemka, Aananya Reddy, Ricardo Isaiah Garcia, Micheal Jacobs et al. "Role of Diet and Exercise in Aging, Alzheimer's Disease, and other Chronic Diseases", Ageing Research Reviews, 2023
Crossref 15 words — < 1%
-
- 13 jutif.if.unsoed.ac.id
Internet 15 words — < 1%
-
- 14 xuanqi-net.com
Internet 15 words — < 1%
-
- 15 "Encyclopedia of Optimization", Springer Science and Business Media LLC, 2009
Crossref 12 words — < 1%
-

- 16 Kevin Oduor, Stephen Ogweno, Naila Chebet Koech, Harrison Ayallo, Ongola Otieno. "Assessing literacy on the interconnection between non-communicable diseases and climate change among youth in Nairobi, Kenya: An interventional study", *MOJ Public Health*, 2024
Crossref 12 words — < 1%
-
- 17 bmcpublichealth.biomedcentral.com
Internet 12 words — < 1%
-
- 18 5dok.net
Internet 11 words — < 1%
-
- 19 fenix.tecnico.ulisboa.pt
Internet 11 words — < 1%
-
- 20 hitconsultant.net
Internet 11 words — < 1%
-
- 21 Susana Limanto, Vincentius Riandaru Prasetyo, Mirella Mercifia. "Optimizing the Amount of Production Using Hybrid Fuzzy Logic and Census II", *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 2023
Crossref 10 words — < 1%
-
- 22 www.researchgate.net
Internet 10 words — < 1%
-
- 23 Fitra Ahya Mubarak, Mohammad Reza Faisal, Dwi Kartini, Dodon Turianto Nugrahadi, Triando Hamonangan Saragih. "Gender Classification of Twitter Users Using Convolutional Neural Network", *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 2023
Crossref 9 words — < 1%

24 Gabjo Kim, Jinwoo Bae. "A novel approach to forecast promising technology through patent analysis", *Technological Forecasting and Social Change*, 2017
Crossref 9 words — < 1%

25 eprints.stmik-banjarbaru.ac.id
Internet 9 words — < 1%

26 tip.ppj.unp.ac.id
Internet 9 words — < 1%

27 Ahmed Youssef, Belaid Moa, Yasser H. El-Sharkawy. "A Novel Visible and Near-Infrared Hyperspectral Imaging Platform for Automated Breast-Cancer Detection", *Photodiagnosis and Photodynamic Therapy*, 2024
Crossref 8 words — < 1%

28 Bobby Ranjan, Wenjie Sun, Jinyu Park, Ronald Xie, Fatemeh Alipour, Kunal Mishra, Vipul Singhal, Shyam Prabhakar. "DUBStepR: correlation-based feature selection for clustering single-cell RNA sequencing data", *Cold Spring Harbor Laboratory*, 2020
Crossref Posted Content 8 words — < 1%

29 Diego Perdigão Sousa, Rong Du, José Mairton Barros da Silva Jr, Charles Casimiro Cavalcante, Carlo Fischione. "Leakage detection in water distribution networks using machine-learning strategies", *Water Supply*, 2023
Crossref 8 words — < 1%

30 Maha Ben-Fares, Nistor Grozavu, Parisa Rastin, Pierre Holat. "High Dimensional Data Stream Clustering using Topological Representation Learning", 2022 *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022
Crossref 8 words — < 1%

31 Marzenna Anna Weresa, Christina Ciecierski, Lidia Filus. "Economics and Mathematical Modeling in Health-Related Research", Brill, 2022 8 words — < 1%
Crossref

32 Miftahuddin Fahmi, Anton Yudhana, Sunardi Sunardi. "Image Processing Using Morphology on Support Vector Machine Classification Model for Waste Image", MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 2023 8 words — < 1%
Crossref

33 Mobeen Shahroz, Muhammad Faheem Mushtaq, Rizwan Majeed, Ali Samad, Zaigham Mushtaq, Urooj Akram. "Feature Discrimination of News based on Canopy and KMGC-Search Clustering", IEEE Access, 2022 8 words — < 1%
Crossref

34 Vasilica Dandea, Gheorghe Grigoras, Bogdan-Constantin Neagu, Florina Scarlatache. "A Clustering-based Knowledge Extraction Methodology for Prosumers' Classification and Injected Power Profiles Grouping", 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2021 8 words — < 1%
Crossref

35 Wen-Chieh Yang, Jung-Pin Lai, Yu-Hui Liu, Ying-Lei Lin, Hung-Pin Hou, Ping-Feng Pai. "Using Medical Data and Clustering Techniques for a Smart Healthcare System", Electronics, 2023 8 words — < 1%
Crossref

36 Yongzheng Yang, Sajad Karampoor, Rasoul Mirzaei, Leonid Borozdkin, Ping Zhu. "The interplay between microbial metabolites and macrophages in cardiovascular diseases: A comprehensive review", International Immunopharmacology, 2023 8 words — < 1%

-
- 37 Yuli Sri Afrianti, Udjianna Sekteria Pasaribu, Fadhil Hanif Sulaiman, Grace Angelia, Henry Junus Wattimanela. "K-MEANS AND AGGLOMERATIVE HIERARCHY CLUSTERING ANALYSIS ON THE STAINLESS STEEL CORROSION PROBLEM", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2024
8 words — < 1%
Crossref
-
- 38 ejournal.uin-suska.ac.id
Internet
8 words — < 1%
-
- 39 www-emerald-com-443.webvpn.sxu.edu.cn
Internet
8 words — < 1%
-
- 40 www-thelancet-com.utorpa.ntunhs.edu.tw
Internet
8 words — < 1%
-
- 41 www.bartshealth.nhs.uk
Internet
8 words — < 1%
-
- 42 www.orfonline.org
Internet
8 words — < 1%
-
- 43 www.saujs.sakarya.edu.tr
Internet
8 words — < 1%
-
- 44 www.techscience.com
Internet
8 words — < 1%
-
- 45 www2.mdpi.com
Internet
8 words — < 1%
-
- 46 Ilkka Vuori. "Physical activity and health: Metabolic and cardiovascular issues", Advances in Physiotherapy, 2009
7 words — < 1%
Crossref

47 Nevena Rankovic, Dragica Rankovic, Igor Lukic, Nikola Savic, Verica Jovanovic. "Ensemble Model for Predicting Chronic Non-Communicable Diseases using Latin Square Extraction and Fuzzy-Artificial Neural Networks from 2013 to 2019", Heliyon, 2023

Crossref

7 words — < 1%

48 Ylber Januzaj, Edmond Beqiri, Artan Luma. "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique", International Journal of Online and Biomedical Engineering (ijOE), 2023

Crossref

7 words — < 1%

49 doaj.org

Internet

7 words — < 1%

50 ojs3.unpatti.ac.id

Internet

7 words — < 1%

51 Mark Fordjour Owusu, Joseph Adu, Sebastian Gyamfi, Ebenezer Martin-Yeboah, Benjamin Ansah Dorte. "Tackling the non-communicable disease epidemic: a framework for policy action in low- and middle-income countries", Pan African Medical Journal, 2024

Crossref

6 words — < 1%

EXCLUDE QUOTES OFF

EXCLUDE SOURCES OFF

EXCLUDE BIBLIOGRAPHY ON

EXCLUDE MATCHES OFF