Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 1, March 2024, pp. 285∼296
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i2.3352

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:

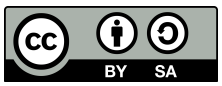
Received September 05, 2023
Revised November 13, 2023
Accepted January 05, 2024

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.



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.

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.

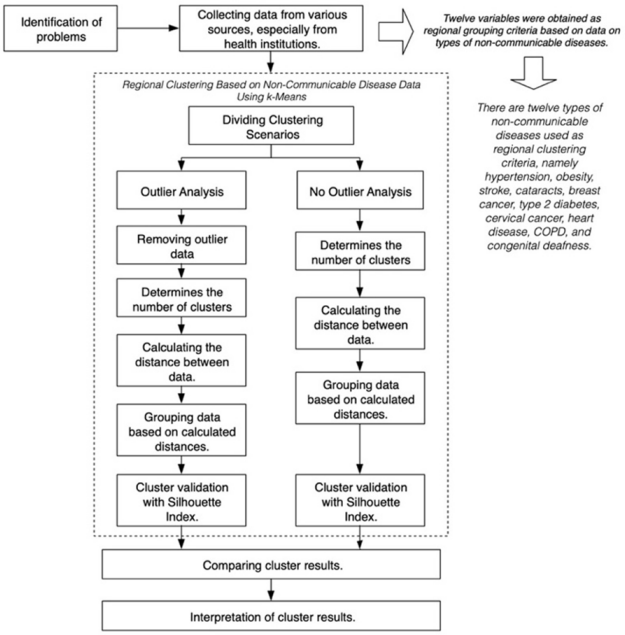


Figure 1. Research Stages

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).

$$d(x, y) = \sqrt{\Sigma_{i=1}^{n}(y_1 - x_1)^2} \quad (1)$$

- 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).

$$\mu_k = \frac{1}{N_k} \Sigma_{q=1}^{N_k} X_q \tag{2}$$

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).

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

Where the variable $a_{(i)}$ represents the average dissimilarity between the ith object and all other objects within the same cluster, the variable $b_{(i)}$ represents the average dissimilarity between the ith 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).

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

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

The variable $|C_I|$ represents the cardinality of cluster CI, 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.

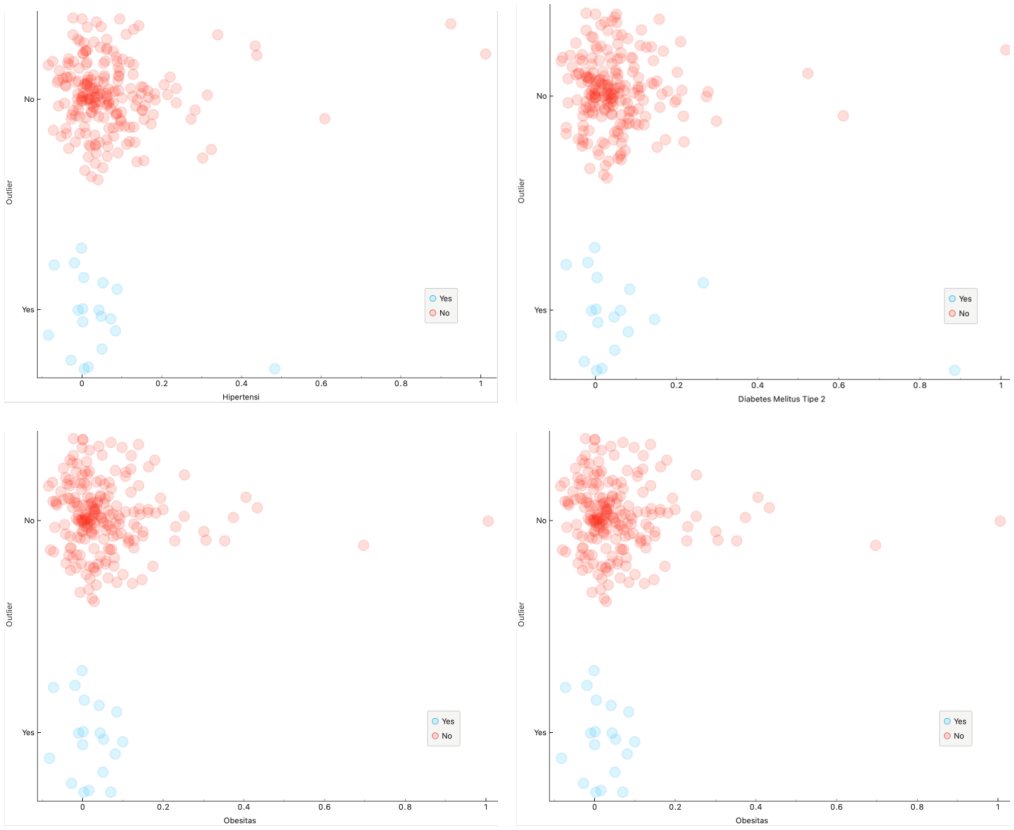


Figure 2. Results of Outlier Analysis for Some Features

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 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 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

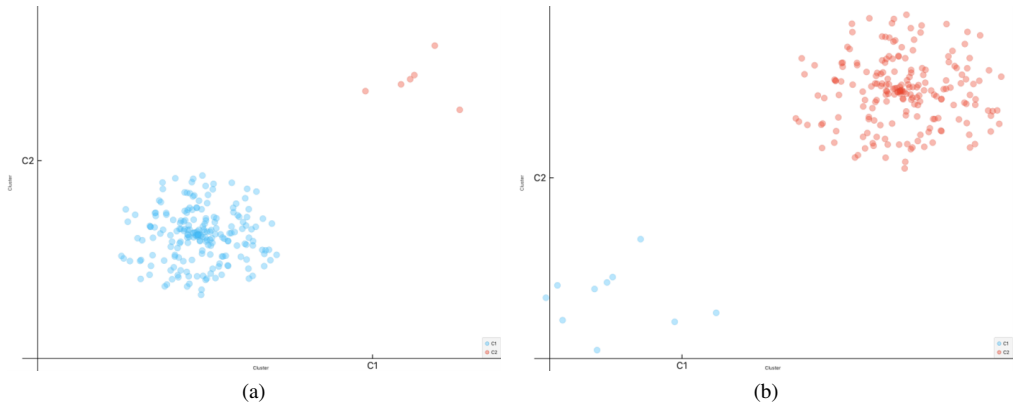
silhouette-NN	0,789	0,728	0,732	0,667	0,388	0,419	0,42
silhouette-Nr	0,812	0,701	0,664	0,649	0,456	0,494	0,514
Number of k	2	3	4	5	6	7	8
C1-Nr	202	193	14	6	157	12	13
C1-NN	10	191	193	185	1	134	134
C2-Nr	6	14	189	5	5	5	2
C2-NN	198	16	12	17	5	6	5
C3-Nr	0	1	4	1	7	20	153
C3-NN	0	1	1	1	122	1	1
C4-Nr	0	0	1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



(a) (b)

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 CONTIBUTION

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. 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, 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://bmcpublichealth.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 Thalangama, 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. WactawskiWende, 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.universitasbumigora.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.universitasbumigora.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