Available online at https://ejournal.corespub.com/index.php/ijmsc/index

International Journal of Mathematics, Statistics, and Computing

e-ISSN 3025-0803

Vol. 1, No. 3, pp. 8-15, 2023

Analysis Of Unemployment Clusters In Indonesia Using The Self Organizing MAP Method

Chairani Gunawan[1*], Haposan Sirait[2]

[1]*Bachelor of Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Riau, Bina Widya Campus, Pekanbaru 28293, Indonesia*

[2]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Riau, Bina Widya Campus, Pekanbaru 28293, Indonesia*

**Corresponding author email: chairani.gunawan0776@student.unri.ac.id*

Abstract

Unemployment is a situation where someone is not working or is trying to find a job but unable to find work. The spread of unemployment in Indonesia has different characteristics in each region, so it is necessary to classify the unemployed so that each government policy program can be carried out in a more focused and directed manner. This study discusses cluster analysis of unemployment using the Self Organizing Map (SOM) method in classifying the unemployed in Indonesia in 2020. The SOM method is able to show dominant patterns and variables in clusters. The variables used in this study consisted of school enrollment rates, average length of schooling, labor force participation rates, and the percentage of the population using computers. The results of this study formed 3 unemployment rate clusters with cluster 1 being a low unemployment group consisting of 144 districts/cities, cluster 2 with a medium level consisting of 287 districts/cities, and cluster 3 with a high level consisting of 83 districts/cities. The grouping using the SOM method on district/city unemployment data in Indonesia is good because it has a minimum standard deviation ratio of 0.529.

Keywords: Cluster analysis, unemployment, standard deviation, SOM

1. Introduction

Machine learning is a branch of artificial intelligence that studies computer algorithms that automatically improve through experience (Mitchell, 1997). Methods in machine learning are divided into two types, namely supervised learning and unsupervised learning. The unsupervised learning method is a modeling, in which an algorithm automatically models a set of inputs without guidance to produce the desired output. Unlike supervised learning, which is an algorithm that creates a function that maps input to the desired output (Putra, 2020). The unsupervised learning technique that is commonly used is cluster analysis (Faizal & Nugrahadi, 2019).

Cluster analysis is a statistical analysis that aims to group data sets into several groups based on certain characteristics (Tan et al., 2006). The method in cluster analysis uses an unsupervised learning model, namely an artificial neural network (Kania et al., 2019). The artificial neural network method for cluster analysis is the Self Organizing Map (SOM). The SOM method is an artificial neural network that can group data based on the characteristics possessed by the data and can show the dominant variable in the clusters formed (Shieh & Liao, 2012). Cluster analysis can be applied in various fields including the economic and employment fields.

The increasing population of Indonesia can lead to the emergence of unemployment problems (Aurangzeb & Khola, 2013). The increase in the number of unemployed was also caused by a decrease in employment (Aqil et al., 2014). The spread of unemployment in Indonesia has different characteristics in each region. Things that can be done to assist the government in knowing the spread of unemployment in Indonesia are grouping the unemployed in each district/city so that the government can find out the spread of unemployment in Indonesia using cluster analysis with the SOM method.

Research using the SOM method was used to determine the relationship between several symptoms of the Covid-19 case in making medical decisions which resulted in 3 clusters of characteristics of the Covid-19 patient group, namely the patient's age was directly related to the length of stay in the hospital, inpatients were directly related to older age, but patient gender is not associated with other symptoms (Ilbeigipour et al., 2022). Research using the SOM method was also used in classifying seasonal and spatial hydrogeochemistry of groundwater in assessing groundwater quality

in the Red River Delta, Vietnam by Nguyen et al (2015) which resulted in three types of distribution including water with high salinity, low salinity, and fresh water.

This research discusses cluster analysis of unemployment using the Self Organizing Map (SOM) method in classifying the unemployed in Indonesia in 2020. It is hoped that this research can provide useful information in the formation of policy programs by the government in an effort to overcome unemployment in Indonesia.

2. Materials and Methods

2.1. Materials

This study uses secondary data obtained from the website of the Central Bureau of Statistics. The data obtained are in the form of school enrollment rates and labor force participation rates derived from the national labor force survey, as well as data on the average length of schooling and the percentage of the population using computers derived from the national socio-economic survey with a total of observations of 514 districts/cities in Indonesia.

2.2. Methods

According to the Central Bureau of Statistics (2020) in the employment indicator, unemployment is residents who are not working but are looking for work and are preparing for a new business or residents who are not looking for work because they have been hired but have not started working. Factors that affect unemployment include not accommodating the entire workforce due to limited available jobs, education, and technological developments (Aqil et al., 2014).

Cluster analysis is a multivariate statistical analysis that is used to group data sets into several groups based on certain characteristics and data with similar characteristics will be grouped into one cluster (Tan et al., 2006). Cluster analysis has characteristics, namely the similarity between one member and another in one cluster which is called internal homogeneity (within cluster) and the difference between one cluster and another which is called external heterogeneity (between cluster). Cluster analysis is useful for discovering previously unknown groups in the data and as a tool for obtaining information about the distribution of data and observing the characteristics of each cluster. In addition, the clusters formed can be further analyzed (Han et al., 2012).

Data standardization is the process of making the standard size and shape of data uniform. The most commonly used data standardization method is the z-score using location size as the average value and standard deviation as the scale measure. Data standardization is done by changing the average value of the data to 0 and the standard deviation value to 1. The standardized z-score value between the i-th object and the j-variable denoted by $z_{i,j}$ can be formulated by the following equation (Gan et al., 2007):

$$z_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_j} \tag{1}$$

where x_i represents the value of the i-th object to the j-variable, $\bar{x}$ represents the average value of the j-variable, and σ_j represents the value of the standard deviation of the j-variable.

The SOM method is an artificial neural network method introduced by Teuvo Kohonen as a topological form of Unsupervised Artificial Neural Network, called "Self Organizing" because it does not require supervision or output targets (unsupervised learning) and is called "Map" because SOM tries to map the weights to match with the given input data. The SOM method can be used to group data based on data characteristics (Shieh & Liao, 2012).

The SOM network consists of two layers, namely the input layer and the output layer. Every neuron in the input layer is connected to every neuron in the output layer. Each neuron in the output layer represents a class (cluster) of the given input. Each output neuron has a weight for each input neuron. During the cluster analysis process, the neuron in the output layer in the cluster that has the closest distance to the input pattern will be selected as the winner and along with its neighboring neurons will improve its weight (Siang, 2009). The SOM method uses the winner takes all basis, that is, only neurons that become winners will update their weights (Ilbeigipour et al., 2022).

According to Larose (2004), the SOM method has three characteristics, namely competition (each weight vector competes with each other to become the winning neuron), cooperation (each winning neuron works with its environment), and adaptation (changes in the winning neuron and its environment). The SOM method requires iteration to get the best grouping results until it gets convergent results. The more iterations are carried out, the smaller the mean distance to closest unit (the average distance of each cluster unit) so that the resulting clusters are also getting better.

Determining the optimal number of clusters in this study uses internal validation. Internal validation is validation that involves utilizing internal information to evaluate the quality of the cluster structure without relying on external information and can be used in determining the optimal number of clusters. Criteria for assessing goodness in determining the optimal number of clusters can be reviewed using internal validation based on the smallest connectivity index value, the largest Dunn index value, and a silhouette width value close to 1 (Brock et al., 2008).

The connectivity index is a method that aims to show the level of cluster relationship based on the number of nearest neighbors (Handl et al., 2005). The value of connectivity is between zero to ∞ and is declared good if the

value is lower in the cluster formed. Suppose $nn_{i(j)}$ is the j-th nearest neighbor of object i where $x_{i,nn_{i(j)}}$ will be 0 if object i and $nn_{i(j)}$ is in the same cluster and has a value of $1/j$ otherwise, so the value of the connectivity index denoted by CC can be formulated by the following equation (Brock et al., 2008):

$$CC = \sum_{i=1}^{N}\sum_{j=1}^{L} x_{i,nn_{i(j)}} \tag{2}$$

where N represents the number of observations, L represents the parameter for the number of nearest neighbors.

Dunn's index is the ratio between the smallest distance formed between two clusters to the largest distance formed within a cluster. The higher the Dunn index value, the better (Brock et al., 2008). Suppose there is a data set with k clusters in which there are clusters p, q, and r with x_i being the i-point in cluster p and y_i being the i-point in cluster q, and z_i and z_j being the i-point and the j-th point in cluster r, so Dunn's index denoted by DI can be formulated by the following equation (Dunn, 1974):

$$DI = min_{p=1,\dots,k}\left\{min_{q=i+1,\dots,k}\left(\frac{d(c_p,c_q)}{max_{r=1,\dots,k}diam(c_r)}\right)\right\} \tag{3}$$

where $d(c_p, c_q)$ represents the distance between cluster p and cluster q where $(c_p, c_q) = min_{x_i \in c_p, y_i \in c_q} d(x_i, y_i)$, $diam(c_r)$ represents the cluster diameter r, $diam(c_r) = max_{z_i, z_j \in c_r} d(z_i, z_j)$.

Silhouette width is the average of each observed silhouette value. The silhouette value can measure the level of confidence in the grouping placement of certain observations (Brock et al., 2008). The silhouette width value has an interval from -1 to 1. A silhouette width value that is close to 1 indicates that the number of clusters is optimal for cluster analysis, whereas if the silhouette width value is close to -1 then the number of clusters is not optimal for cluster analysis. The silhouette width value denoted by S(i) can be formulated by the following equation (Rousseeuw, 1987):

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}, i = 1,2,\dots,n \tag{4}$$

where a_i represents the average distance of the i-th observation to other observations in the same cluster where $a_i = \frac{1}{n(C(i))}\sum_{j \in C(i)} dist(i,j)$, b_i represents the average minimum distance of the i-th observation to all other observations in the nearest neighbor cluster where $b_i = min_{C_k \in C \setminus C(i)} \sum_{j \in C_k} \frac{dist(i,j)}{n(C_k)}$.

The grouping method is said to be good if it has a minimum within-group standard deviation (S_w) and a maximum inter-group standard deviation (S_b) (Barakbah & Arai, 2004). The standard deviation value in a cluster is denoted by (S_w) which can be formulated using the following equation (Bunkers & Miller, 1994):

$$S_w = \frac{1}{K}\sum_{k=1}^{K} S_k \tag{5}$$

where K represents the number of clusters formed, S_k represents the standard deviation of the k-th group, where $S_k = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_i - \bar{X}_k)^2}$

If there is an average variable in each k cluster denoted by $\bar{X}_k$, then the components of each cluster are different, and the standard deviation between groups is denoted by (S_b) can be formulated by the following equation (Barakbah & Arai, 2004):

$$S_b = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})^2} \tag{6}$$

where $\bar{X}$ represents the average of all clusters.

The criterion of a grouping is said to be good if it produces a small standard deviation ratio between S_w and S_b (Barakbah & Arai, 2004). The standard deviation ratio used between the values of S_w and S_b denoted by R can be formulated by the following equation:

$$R = \frac{S_w}{S_b} \times 100\% \tag{7}$$

If the value of $R < 1$, it can be said that the clustering method is good at grouping the data being analyzed.

3. Results and Discussion

Descriptive statistical analysis to present a general description of the data to be used in research. The data used is unemployment data in Indonesia for 2020 with a total of 514 districts/cities. The variables used are 4 variables, namely school enrollment rate, average length of schooling, labor force participation rate and percentage of population using computers. The descriptive statistics are presented in Table 1 below:

Table 1: Descriptive Statistics

Variable	Minimum	Maximum	Average	Standard Deviation
X_1	22.55	99.82	74.53	10.73
X_2	1.13	12.65	8.34	1.63
X_3	36.65	96.25	69.18	6.34
X_4	0.06	47.62	13.02	7.07

Based on Table 1, it is found that all variables have a standard deviation value that is smaller than the average value so that they have an even distribution of data and a low level of data variation. The highest score is 99.82 in the school enrollment rate variable (X_1) which is located in Central Maluku district. The lowest value is 0.06 for the variable percentage of the population using a computer (X_4) located in Nduga and Lanny Jaya districts.

This research data has variables with different measurement scales so it needs to be standardized. The results of the calculation of standardized data are presented in Table 2 below:

Table 2: Results of Data Standardization Using the Z-Score.

Regency/City	APS	RLS	TPAK	PK
Simeulue	1.32	0.61	0.19	-0.26
Aceh Singkil	0.97	0.11	-1.14	-0.51
Aceh Selatan	0.81	0.32	-1.23	-0.50
Aceh Tenggara	0.83	0.81	0.34	-0.34
Aceh Timur	-0.12	-0.12	-1.15	-1.21
Aceh Tengah	1.30	0.92	1.70	0.67
⋮	⋮	⋮	⋮	⋮
Deiyai	1.65	1.97	-0.96	1.83
Kota Jayapura	-1.33	-3.15	-5.13	-1.83

The standardized data will then be used to determine the optimal number of clusters. Determination of the optimal number of clusters can be chosen based on the lowest connectivity index value, the dunn index value that is close to 1 and the highest silhouette coefficient value. The results of determining the optimal number of clusters are presented in Table 3 as follows:

Table 3: Determination of the Optimal Number of Clusters

Cluster Validation	Number of Clusters				
	2	**3**	**4**	**5**	**6**
Connectivity	95.806	**70.208**	133.769	160.302	196.771
Dunn	0.014	0.019	0.014	0.019	**0.030**
Silhouette	0.307	**0.368**	0.257	0.258	0.252

The results of determining the optimal number of clusters based on Table 3 show that the value of the lowest connectivity index is located in cluster 3, which is 70,208, the dunn index that is close to 1 is located in cluster 6, which is 0.030, and the highest silhouette value is located in cluster 3, which is 0.368. Therefore, the optimal number of clusters used based on the results of cluster validation is 3 clusters.

Before starting the cluster analysis stage with the SOM method, you must first know the number of neurons that represent the many clusters that are formed. The optimal cluster results are 3 with 4 input vectors representing each variable, while the amount of data used is 514 districts/cities. In the cluster analysis process, 4 input vectors will be used according to the number of variables in the study. Furthermore, the optimal number of clusters used is $p = 3$.

Iteration 1:

Initialize random weights as follows:

$$\begin{bmatrix} 0.10 & 0.60 & 0.50 \\ 0.40 & 0.30 & 0.60 \\ 0.20 & 0.70 & 0.10 \\ 0.10 & 0.20 & 0.30 \end{bmatrix}$$

The first vector is:

$$[1.32 \quad 0.61 \quad 0.19 \quad -0.26]$$

Calculating the vector distance $d_{(i,j)}$ based on equation (8) by adding up the difference between the weight vector (w_{ij}) and the input vector (x_i), the following results are obtained in Table 4:

Table 4: Vector Distance Results

Vector to-k	Vector Distance
$d_{1,1}$	3.16

$d_{2,1}$	4.98
$d_{3,1}$	5.09

Based on the calculation of the weight distance values in Table 4, the smallest $d_{(i,j)}$ value is 3.16, then the 1st weight vector $(d_{1,1})$ is selected to proceed to the weight change stage $(w_{ij\ new})$ based on equation (9) the results are obtained in table 5 below:

Table 5: New Vector Distance Results

Vector to-*k*	**New Vector Distance**
$w_{1,1}$	0.16
$w_{2,1}$	0.41
$w_{3,1}$	0.20
$w_{4,1}$	0.08

Then the results of the weight matrix are obtained by replacing random w_{ij} with $w_{ij\ new}$ values as follows:

$$\begin{bmatrix} \mathbf{0.16} & 0.60 & 0.50 \\ \mathbf{0.41} & 0.30 & 0.60 \\ \mathbf{0.20} & 0.70 & 0.10 \\ \mathbf{0.08} & 0.20 & 0.30 \end{bmatrix}$$

So on like that until the to 514 data. The next step is to update the learning rate (α) based on equation (10) with $\alpha = 0.01$, namely $\alpha = 0.5(0.01) = 0.005$ and the process is continued until to the n iteration until it converges as shown in the following training progress chart:

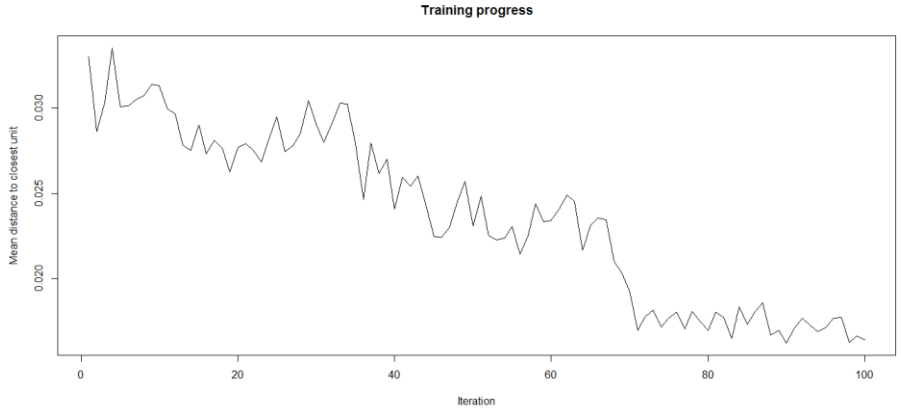

Figure 1: Graph of Training Progress

Based on the graph in Figure 1, it is found that the training progress has been carried out 100 iterations. Iterations show convergent conditions starting when iterations are close to 70 and stopping when they are at 100 iterations with the mean distance to closest value starting to stabilize when it is below 0.020.

Visualization using the SOM method will produce a fan diagram presented in Figure 2 below:

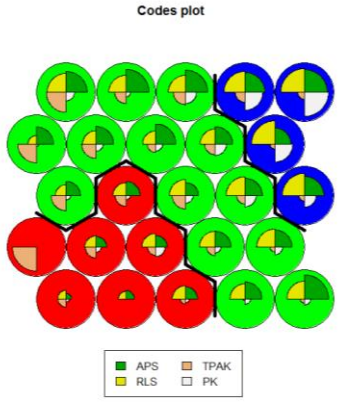

Figure 2: Visualization Results Using the SOM Method

Based on the visualization results of the SOM method using the hexagonal topology in Figure 2, a fan diagram is obtained which shows the grouping of districts/cities in Indonesia. This study uses 4 variables with a 5x5 grid. The fan diagram has a border line to distinguish the colors in each cluster. The colors in cluster 1 are marked with a red fan diagram, cluster 2 is shown in green, cluster 3 is shown in blue

Circles represent 4 variables with different colors, the color differences in the visualization results indicate clusters that are formed based on the variables used in the study. Cluster 2 has a labor force participation rate variable marked in brown indicating that brown in cluster 2 is more dominant with a higher average value than brown in clusters 1 and 3. The cluster results are presented in Table 6 as follows:

Table 6: SOM Cluster Results

Cluster to-k	Number of Members
Cluster 1	144
Cluster 2	287
Cluster 3	83

Based on Table 6, the cluster results obtained using the SOM method show that unemployment in Indonesia is divided into 3 clusters with 144 districts/cities, 287 districts/cities, and 83 districts/cities each member. The distribution map of district/city cluster results in Indonesia is presented in Figure 3 below:

Figure 3: Results of SOM Method Cluster Mapping

Based on Figure 3, it can be seen that the majority of cluster 1 is spread in the western part of Indonesia, the majority of cluster 2 is spread on the islands of Sumatra and Kalimantan, and the majority of cluster 3 is spread on the island of Java.

The interpretation of the cluster results from each cluster can be seen from the average variables in each cluster which are presented in Table 7 as follows:

Table 7: Average Cluster Variable Value

Variable	Cluster 1	Cluster 2	Cluster 3
APS	-1.044	0.311	**0.735**
RLS	-0.831	-0.014	**1.490**
TPAK	-0.006	**0.156**	-0.527
PK	-0.622	-0.198	**1.765**

Based on Table 7, it is found that cluster 1 is declared as a group of regencies/cities with a low level of unemployment, cluster 2 is declared as a group of regencies/cities with a moderate or medium level of unemployment, and cluster 3 is declared as a group of regencies/cities with a high level of unemployment. Cluster 1 is a region that has the characteristics of a relatively low spread of unemployment compared to other clusters. Cluster 2 is a region that has a spread of unemployment with the characteristics of a high labor force participation rate. Cluster 3 has the characteristics of high enrollment rates, average length of schooling, and high computer users.

Next, we will calculate the standard deviation within the cluster (S_w) and the standard deviation between clusters (S_b) to assess the cluster's performance in classifying unemployment by district/city in Indonesia. Calculation of the value of (S_k) is done before calculating (S_w). The overall results of the S_k values of each cluster are presented in Table 8 as follows:

Table 8. Cluster Standard Deviation

k	Baku Devi to -k
1	0.433
2	0.356

3	0.397

Based on Table 6, the standard deviation value in the (S_w) cluster is obtained which is calculated based on equation (5) as follows:

$$S_w = \frac{1}{K}\sum_{k=1}^{K} S_k$$
$$= \frac{(0.433 + 0.356 + 0.397)}{3}$$

$S_w = 0.395$

Calculation of the average of the entire group is carried out before calculating the standard deviation between clusters (S_b). Calculation of the average of the entire group as follows:

$$\bar{X} = \frac{-0.626 + 0.064 + 0.866}{3}$$
$$\bar{X} = 0.101$$

After obtaining the average of the entire group, the standard deviation between clusters (S_b) can be calculated based on equation (6) as follows:

$$S_b = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\bar{X} - \bar{X}_k)^2},$$
$$= \frac{(-0.626 - 0.101)^2}{2} + \frac{(0.064 - 0.101)^2}{2} + \frac{(0.866 - 0.101)^2}{2}$$

$S_b = 0.746$

Based on the calculations (S_w) and (S_b), it is found that the standard deviation within the cluster (S_w) produces the minimum value while the standard deviation between clusters (S_b) produces the maximum value. Furthermore, the value of the standard deviation ratio is calculated to determine whether the resulting grouping can be said to be good or not. The standard deviation ratio of the results of the cluster analysis which is calculated based on equation (7) is as follows:

$$R = \frac{S_w}{S_b} \times 100\%$$
$$= \frac{0.395}{0.746} \times 100\%,$$
$$R = 0.529.$$

The standard deviation ratio yields a value of $R < 1$, so it can be said that the SOM method is good at classifying the unemployed based on districts/cities in Indonesia.

4. Conclusion

Based on the results that have been obtained, it can be concluded that the results of the analysis using the SOM method obtain 3 optimal clusters, namely cluster 1 which is a group of regencies/cities with a low unemployment rate consisting of 144 members, cluster 2 with a moderate unemployment rate consisting of 287 members, and cluster 3 with a high level of unemployment which has 83 members. Based on the analysis that has been done, it can be concluded that the results of grouping using the SOM method on district/city unemployment data in Indonesia are good. This can be explained from the small standard deviation ratio value of 0.529. The government as a policy maker is expected to be able to take advantage of the grouping results as material for consideration in making policies regarding the problem of unemployment in Indonesia based on regions that have the same characteristics.

References

Aqil, M., Qureshi, M. A., Ahmed, R. R., & Qadeer, S. (2014). Determinants of unemployment in Pakistan. *International Journal of Physical and Social Sciences*, *4*(4), 676-682.

Aurangzeb & Khola, A. (2013). Factor effecting unemployment: A cross country analysis. *International Journal of Academic Research in Business and Social Science*. *3*(1), 219-230.

Badan Pusat Statistik. (2020). *Indikator Pasar Tenaga Kerja Indonesia*. Jakarta: Badan Pusat Statistik.

Barakbah, A. R., & Arai, K. (2004). Determining constraints of moving variance to find global optimum and make automatic clustering. *Industrial Electronics Seminar (IES)*, *1*(1), 409–413.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: an R package for cluster validation. *Journal of Statistical Software*, *25*(4), 1–22.

Bunkers, M. J., & Miller, J. R. (1994. Definition of climate regions in the northern plains using an objective cluster modification technique. *Journal of Climate*, *9*(1), 130-146.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104.

Faizal, M. R., & Nugrahadi, D. T. (2019). *Belajar data science: Klasifikasi dengan bahasa pemrograman R*. Banjarbaru: Scripta Cendekia.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering theory, algorithms, and applications.* Pennsylvania: SIAM.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (2nd ed.). Massachusetts: Morgan Kaufmann Publishers.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201–3212.

Ilbeigipour, S., Albadvi, A., Noughabi, E. A. (2022). Cluster-based analysis of covid-19 cases using self-organizing map neural network and k-means methods to improve medical decision-making. *Informatics in Medicine Unlocked*, *32*(1), 1-12.

Kania, S., Rachmatin, D., & Dahlan, J. A. (2019). Program aplikasi pengelompokan objek dengan metode self organizing map menggunakan bahasa R. *Jurnal EurekaMatika*, *7*(2), 17-29.

Larose, D. T. (2004). *Discovering knowledge in data: An introduction to data mining*. New Jersey: John Wiley & Sons, Inc.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill, Inc.

Nguyen, T. T., Kawamura, A., Tong, T. N., Nakagawa, N., Amaguchi, H., & Jr. R. G. (2015). Clustering spatio–seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the red river delta, Vietnam. *Journal of Hydrology*, *522*(10), 661-673.

Putra, J. W. G. (2019). *Pengenalan konsep pembelajaran mesin dan deep learning*. Tokyo: Inst. Technol.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(1), 53–65.

Shieh, S. L., & Liao, I. E., S. D. (2012). A new approach for data clustering and visualization using self-organizing maps. *International Journal of Expert Systems with Applications*, *39*(15), 1924–11933.

Siang, J. J. (2009). *Jaringan syaraf tiruan & pemrogramannya menggunakan matlab*. Yogyakarta: Andi.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Education.