Ensemble Cluster Method for Clustering Cabbage Production in East Java

Maulidya Maghfiro[1*], Ni Wayan Surya Wardhani[2] , and Atiek Iriany[3]
[1,2,3] Statistics Department, Brawijaya University, Malang, Indonesia
*Corresponding author: maulidya56@student.ub.ac.id

Received: 31 May 2024
Revised: 24 June 2024
Accepted: 25 June 2024

ABSTRACT – Cluster analysis is a multivariate analysis method classified under interdependence methods, where explanatory variables are not differentiated from response variables. The methods used include hierarchical cluster analysis, such as agglomerative and divisive, and non-hierarchical methods such as Self Organizing Maps (SOM) based on Artificial Neural Networks (ANN). Various cluster analysis methods often yield diverse solutions, making it challenging to determine the optimal solution. Therefore, the ensemble cluster method is employed to combine various clustering solutions without considering the initial data characteristics providing better results. One case study of clustering is the grouping of cabbage production. East Java Province has become the third-highest cabbage-producing province in Indonesia with a production of 210,454 tons. Clustering of cabbage-producing regencies/cities was conducted to optimize production and identify areas that have not yet reached their maximum potential. This study compares five clustering methods which are hierarchical analysis (complete linkage, single linkage, average linkage), Self-Organizing Map (SOM), and Ensemble Cluster. The quality of clustering was evaluated using the Silhouette Coefficient (SC), Dunn Index (DI), and Connectivity Index (CI). The results indicate that the Ensemble Cluster method showed the best performance, with an SC value of 0.9124, a DI value of 1.3734, and a CI value of 2.9290, indicating excellent cluster separation. Therefore, the ensemble cluster method is recommended as the best clustering method in this study.

Keywords – Cluster, hierarchical analysis, SOM, Ensemble Cluster, Cabbage.

I. INTRODUCTION

Cluster analysis is a type of multivariate analysis that is included in the interdependence method, where independent or explanatory variables are not distinguished from dependent or response variables [1]. Cluster analysis aims to group objects that have similar characteristics so that the resulting group consists of objects that have the closest characteristics in common [2]. The success of a group is determined by the high level of homogeneity between objects in the group and the high level of heterogeneity with other group objects. Methods that can be used for cluster analysis include hierarchical and non-hierarchical cluster analysis.

Clusters with a hierarchical approach will group similar data in the same hierarchy and dissimilar data in different hierarchies [3]. There are two types of hierarchical cluster methods, namely agglomerative and divisive. The agglomerative method combines individuals into groups gradually, while the divisive method separates individuals into successively smaller groups [4]. In the hierarchical method, there are several techniques such as single linkage, complete linkage, and average linkage. Each has a different approach to cluster formation. There are several methods for measuring distance in cluster analysis, including Euclidean distance, Manhattan distance, and Mahalanobis distance. The hierarchical method has the advantage that the grouping of objects is presented in a dendrogram so that the grouping results are more informative [5].

One non-hierarchical cluster method is Self Organizing Maps (SOM). SOM is a cluster method based on artificial neural networks (ANN) or also called Artificial Neural Networks (ANN) [6]. Although SOM is ANN-based, this analysis does not use target class values and does not assign classes to each data, so SOM can be used for grouping purposes. SOM is used to group data based on data characteristics/features. Previous research conducted by [7] stated that the use of the SOM method is a method that has a fairly good level of accuracy for classifying a place, area, region, object, etc.

In Cluster analysis there are many methods that can be applied, namely conventional methods in the form of hierarchical and non-hierarchical cluster analysis, however, these methods produce diverse cluster solutions so that in this condition the cluster solution will be difficult to determine. A set of cluster solutions obtained from various methods are complementary so that an effective way to obtain good cluster results is to combine these solutions. Cluster ensemble is a method used to combine various clustering solutions without looking at the characteristics of the initial data [8]. This method was introduced by [9]. The basic idea of cluster ensemble is to combine a set of cluster results from other methods. According to Strehl's research, ensemble clusters can provide higher quality cluster results.

Previous research on hierarchical cluster analysis was conducted by [10] with the title "Comparison of Single Linkage, Complete Linkage, and Average Linkage Methods on Community Welfare Analysis in Cities and Regencies in East Java". his study compares different types of linkages in hierarchical analysis, such as single linkage, complete linkage, and average linkage, in the context of welfare analysis in cities and districts in East Java. Research on the Self Organizing Maps (SOM) clustering algorithm was conducted by [11] with the title "Analysis of Spatial Spread Relationships of the Coronavirus (COVID-19) Pandemic in the World using Self Organizing Maps." The objective of this study is to map the spatial spread of the COVID-19 pandemic worldwide. Research on ensemble clustering was carried out by [12] with the

DOI: 10.12962/j27213862.v7i2.20378

title "Grouping Districts/Cities on the Island of Java Based on Socio-Economic Conditions Before and After Entering the COVID-19 Pandemic." This study aims to analyze the socio-economic conditions on Java Island before and after the onset of the COVID-19 pandemic. An ensemble method using the Cluster-based Similarity Partitioning Algorithm (CSPA) approach was employed. The results indicate that ensemble clusters perform better than single clusters

One case of grouping is the grouping of cabbage production. Cabbage is one of the horticultural commodities that is widely cultivated. Cabbage is a type of vegetable that has been known since ancient times, especially in the period 2500-2000 BC [13]. Cabbage is included in the category of vegetable plants that only live for one season or have a short lifespan, only producing once before finally dying. Cabbage harvesting is usually done after the plants are 60-80 days old from planting [13]. In 2020, East Java Province ranked fourth in cabbage production in Indonesia with the amount reaching 203,708 tons. However, cabbage production in East Java province decreased in 2021 to 193,026 tons. Then, in 2022, East Java Province succeeded in increasing its production and rose to third place as the province with the highest cabbage production, reaching 210,454 tons. To optimize cabbage production in East Java Province, you can group districts/cities that produce cabbage. This aims to determine the potential for cabbage production in each region as well as identify areas that have not achieved maximum cabbage production.

This study extends previous research by developing a clustering method that combines hierarchical clusters and SOM through an ensemble clustering approach. This research applies the Cluster-based Similarity Partitioning Algorithm (CSPA) for an ensemble clustering approach. It is hoped that the results of this research will provide better and optimal cluster analysis. The selection of the best cluster method was carried out based on the Silhouette Score, Dunn Index and Connectivity Index criteria.

II. LITERATURE REVIEW

A. Data Standardization

Data standardization is used to standardize data variables to avoid problems caused by the use of scale values that are not balanced between object grouping variables. The data standardization that is usually used is z-score, which is calculating the middle value and dividing the results. The advantage of the z-score method is that it can compare the quality of information to the average data in a group based on the standard deviation value [14]. The following is the z-score formula [15]:

$$Z_i = \frac{(x_i - \bar{x})}{s} \tag{1}$$

Where:
Z_i: variable standardization
x_i: data to i
$\bar{x}$: the average of the overall data for each variable

B. Assumption Testing

1. Assumptions of Representation in Samples

Another important assumption that must be met is sampling adequacy, which determines whether the data used is sufficient and appropriate for clustering. This is assessed using the KMO (Kaiser-Meyer-Olkin) test. The KMO is an index that compares the observed correlation coefficients with the partial correlation coefficients [16], The formula for calculating the KMO is as follows [17]:

$$KMO = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} r_{ij}^2}{\sum_{i=1}^{n}\sum_{j\neq i}^{n} r_{ij}^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n} r_{x_i x_k - x_j}^2} \tag{2}$$

with
i : $1,2,\ldots,p$
j : $1,2,\ldots,p$
k : $1,2,\ldots,p$
$r_{x_i x_k - x_j}$: Partial correlation coefficient between the i-th variable and the k-th variable, controlling for the j-th variable
$r_{x_i x_k}$: Partial correlation coefficient of the i-th variable and the k-th variable
$r_{x_j x_k}$: Partial correlation coefficient of the j-th variable and the k-th variable
$r_{x_i x_j}$: Partial correlation coefficient of the i-th variable and the j-th variable

2. Assumption of Correlation between Variables

One of the assumptions that must be met in cluster analysis is the absence of a high correlation between independent variables. According to [1], data is considered to have a strong correlation if the correlation coefficient is large ($r > 0.5$). The correlation test is used to measure the strength of the relationship between variables without determining whether one variable depends on another. The correlation coefficient is used to indicate the degree of relationship between variables. This study employs the Pearson product-moment correlation test (r) to determine the closeness of the relationship between variables, expressed by the correlation coefficient (r) [18]. The Pearson correlation coefficient can be calculated using the following formula [19]:

$$r_{x_i,x_j} = \frac{n\sum_{i=1}^{n} x_i x_j - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_j}{\sqrt{[n\ \sum_{i=1}^{n}\ x_i^2 - (\sum_{i=1}^{n} x_i)\hat{}2\ ]\ [n\ \sum_{i=1}^{n}\ x_j^2 - (\sum_{i=1}^{n} x_j)\hat{}2\ ]}} \tag{3}$$

with,

DOI: 10.12962/j27213862.v7i2.20378

r_{x_i,x_j} : Correlation between the *i*-th variable and the *j*-th variable, with $i = 1,2,\dots,n$ and $j = 1,2,\dots,p$
x_i, x_j : The *i*-th variable and the *j*-th variable with $i = 1,2,\dots,n$ and $j = 1,2,\dots,p$

C. Principal Component Analysis

PCA is a multivariate data selection technique that transforms the original data matrix into a set of homogeneous combinations that are fewer in number but capture a large portion of the variance from the initial data [20]. The main objective is to represent as much diversity as possible in the original data using as few principal components as possible [21]. PCA creates a linear combination of the original variables, which geometrically represents a new coordinate system obtained by rotating the original system [22]. The PCA method is particularly useful when the data contains a large number of variables that are correlated. PCA calculations are based on computing eigenvalues and eigenvectors, which represent the distribution of data within a dataset [2].

There is a matrix dataset X of size ($n\ x\ p$) which consists of n observations $x_i,\ i = 1, 2, \dots, n)$with p dimensions. The algorithm for PC as described [2] is as follows:

1. Calculate the characteristic root $\boldsymbol{\lambda}$ that satisfies the equation:
$$|\lambda I - X| = 0 \tag{4}$$
2. Calculate the value of X that satisfies the equation:
$$AX = \lambda X \tag{5}$$
3. Write down the linear combination equation:
$$Y = \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_p X_p \tag{6}$$
4. The variance that can be explained by the new variable i depends on the percentage contribution
$$P_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} x\ 100\% \tag{7}$$
5. Determining the number of new variables used depends on the percentage of cumulative contribution from the cumulative characteristic roots that have been derived from the largest value. The percentage value of cumulative contribution to the rth component is calculated using the formula:
$$p_i = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i} x\ 100\% \tag{8}$$
with,$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$

D. Euclidean Distance

Euclidean distance is the most commonly used type of distance measurement because it is one of the simplest methods to understand and model. It is used to measure the distance from data objects to the cluster center, making it suitable for determining the shortest distance between two data points. Euclidean distance represents the geometric distance between two data objects [2]. The Euclidean distance between two points can be calculated using the following equation [23].

$$d(x_i, y_j) = \sqrt{\sum_{i=p}^{p} (x_{ik} - y_{ik})^2} \tag{9}$$

Where :
$d(x_i, y_j)$: Distance between object *i* and object *j*
x_{ik} : value of object *i* in the *k*-th variable
y_{jk} : value of object *j* in the *k*-th variable
p : many variables are observed

E. Hierarchical Cluster Analysis

In hierarchical cluster analysis, it is assumed that initially, each object forms its own cluster. Then, the two closest objects or clusters are merged to form a single, smaller cluster [2]. Hierarchical methods offer advantages over non-hierarchical methods. One advantage is that hierarchical methods make it easier to study all the formed clusters and provide more information, as the clustering stages are represented in a dendrogram or tree diagram. Hierarchical cluster analysis includes two methods: agglomerative (merging) and divisive (splitting). In the agglomerative method, each object starts as its own cluster, and then clusters that are close to each other are combined into one cluster. In the divisive method, all objects start in a single cluster, and then the most dissimilar properties are separated to form another cluster [2].

1. Single Linkage Method

Determining the distance between clusters using the single linkage method involves examining the distance between two existing clusters and selecting the shortest distance, also known as the nearest neighbor rule. According to [2], the single linkage is defined as the minimum distance between a point in cluster *A* and a point in cluster *B*, which is expressed by the following equation:

$$d_{(AB)C} = \min\,(\mathrm{d_{AC}}, d_{BC}) \tag{10}$$

Where :
$d_{(AB)C}$: Distance between cluster (AB) and cluster C
d_{AC} : Distance of object A in cluster C
d_{BC} : Distance of object B in cluster C

DOI: 10.12962/j27213862.v7i2.20378

2. Complete Linkage Method

In the complete linkage method, the distance between clusters is determined by the farthest distance between two objects in different clusters [2]. According to [2], complete linkage is defined as the maximum distance between a point in cluster A and a point in cluster B, which is expressed by the following equation:

$$d_{(AB)C} = max\ (d_{AC}, d_{BC}) \tag{11}$$

Where :

$d_{(AB)C}$: Distance between cluster (AB) and cluster C
d_{AC} : Distance of object A in cluster C
d_{BC} : Distance of object B in cluster C

3. Average Linkage Method

In the average linkage method, the distance between two clusters is calculated as the average distance between all members of one cluster and all members of the other cluster. According to [2], the distance between cluster AB and another cluster is calculated using the formula expressed by the following equation:

$$d_{(AB)C} = \frac{\sum_A \sum_B d_{AB}}{N_{AB} N_c} \tag{12}$$

Where :

$d_{(AB)C}$: Distance between cluster (AB) and cluster C
d_{AC} : Distance of object A in cluster C
d_{BC} : Distance of object B in cluster C
N_{AB} : The number of objects in the cluster (AB)
N_C : The number of objects in the cluster (C)

F. Self Organizing Maps (SOM)

Self Organizing Maps (SOM) was first introduced by Teuvo Kalevi Kohonen in 1982. SOM is an artificial neural network (ANN) technique used for clustering data into several groups. It groups data that share similarities with other objects into distinct clusters. SOM operates as an unsupervised learning algorithm, meaning it does not require predefined output targets. Instead, its goal is to determine a set of centroids (neurons) that represent clusters while adhering to topological constraints. Topology in this context refers to the arrangement of centroids on the output grid, with hexagonal and rectangular grids being the most commonly used [11]. One of the advantages of the SOM algorithm for visualization and cluster analysis is its ability to explore groupings and relationships in high-dimensional data by projecting the data into two dimensions that clearly show areas of similarity [24]:

The stages of the SOM network pattern grouping algorithm are as follows [25] :

1. Initialize the weights W_{ji}randomly with the weight matrix columns being the number of elements in a vector and the weight matrix rows being the maximum number of clusters that will be formed.
2. Calculate the distance between input values and weights $D_j$using Euclidean distance as follows [26] :

$$D_j = \sqrt{\sum_{i=1}^{m} \left(W_{ji} - x_i\right)^2} \tag{13}$$

where:

W_{ji} : The weight of the link between the input neuron and the output neuron
x_i : Neurons in the ith input layer
m : The many variables

3. Determine the minimum value by looking at the distance vector calculation results D_j, then changes will be made to the weights using the following formula [26] :

$$W_{ji}(baru) = W_{ji}(lama) + \ \alpha[x_i - W_{ji}(lama)] \tag{14}$$

At the stage of obtaining new weights, the learning rate value $\alpha$is $0 \ \leq \ \alpha \leq 1$. To make changes to the learning rate, you can use the following formula [27] :

$$\alpha(t) = \ \alpha_i(1 - \frac{t}{t_{max}}) \tag{15}$$

Where :

α_i : Initial learning rate value i
t_{max} : Maximum number of iterations

4. The state that the test has stopped can be determined by calculating the W_{ji}(new) weight with $W_{ji}$the (old) when the weight $W_{ji}$is known to have not changed or the change is not much, it can be said that the test has stopped and has reached convergence.

G. Cluster Ensembles

Strehl (2002) introduced a method used to combine a set of group solutions called Cluster Ensemble. Research conducted by [9] shows that the cluster ensemble method can improve the quality and robustness of cluster solutions. The implementation of this technique in cluster analysis is called Cluster Ensemble or consensus cluster [28]. Clustering in ensemble clusters is carried out by combining various solutions from various clustering methods until a better final cluster is obtained [29]. Ensemble clustering has advantages in 4 aspects, namely:

DOI: 10.12962/j27213862.v7i2.20378

1. Robustness, providing better performance.
2. Novelty, produces solutions that are not tied to one type of algorithm.
3. Stability and confidence estimation, have a low level of sensitivity to noise, outliers, and variance generated by sampling.
4. Parraleization and scalability, provide opportunities to integrate several cluster results through parallelization techniques.

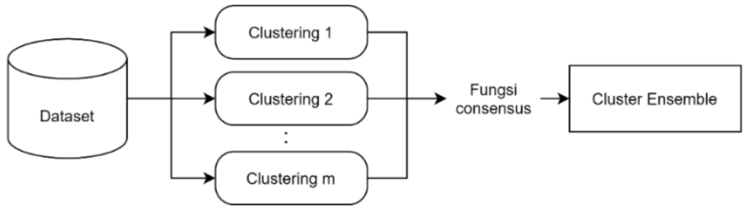

Figure 1 Grouping steps with Cluster Ensemble
Source: [12]

Grouping objects using the Cluster Ensemble method is carried out in two stages [12], namely:

1. Forming ensemble members whose members are solutions from various grouping methods.
2. Combining all ensemble members to obtain one final solution called the Consensus function.

After obtaining the final ensemble clustering results, a validity test is carried out to assess the quality of the cluster results formed.

In the Cluster Ensemble, there is a function called the consensus function. The consensus function is defined as a function that maps a set of cluster solutions into a combined solution until a stable and robust grouping result is obtained [12]. The idea of forming a consensus is to combine existing data perspectives into a stable partition. The consensus function is divided into two stages [30] :

a. The results of each cluster method are combined by partitionp = $\{\pi^1, \pi^2, \dots, \pi^n\}$.
b. Existing partitions are combined to produce the final partition with$p^1 = \{c_1^1, c_2^1, \dots, c_{k_n}^n\}$

Research by [9] suggests that there are three effective approaches to the consensus function: (Cluster-based Similarity Partitioning Algorithm) CSPA ($O(k.n^2.m)$), (Hyper-Graph Partitioning Algorithm) HGPA ($O(k.n.m)$), and (Meta-Clustering Algorithm) MCLA ($O(k^2.n.m^2)$), where n is the number of observations, m is the partition number, and k is the cluster number in the consensus partition. In this study, the CSPA algorithm is employed for the consensus function.

The Cluster-based Similarity Partitioning Algorithm (CSPA), proposed by [9] is a graph-based consensus clustering algorithm that efficiently generates a consensus partition from an ensemble of partitions created during the clustering process [12]. Generally, CSPA constructs a similarity matrix based on the ensembles, which measures how frequently pairs of data points co-occur in the ensemble partitions [31]. This similarity matrix provides a measure of pairwise similarity between data points and is used to re-group or re-cluster observations, thereby producing a combined clustering result following is the algorithm and illustration of the Cluster-based Similarity Partitioning Algorithm calculation [32]:

1. Building ensemble members by forming relabeling of datasets $x = \{x_1, x_2, \dots, x_p\}$ and clustering methods $\prod = \{\pi_1, \pi_2, \dots, \pi_m\}$.
2. The resulting relabeling is transformed into an order matrix $n\ x\ p$for each ensemble cluster member, denoted by S_m Each $m = 1, 2, \dots, n$. entry in this matrix represents the relationship between two data points
3. Form a weighting matrix with the following steps:
   a. Form a matrix (w) with the equation

$$W_{ij} = \frac{\left|X_{c_i} \cap X_{c_j}\right|}{\left|X_{c_i} \cap X_{c_j}\right|} \tag{16}$$

   b. Form a WCT (Weight Connected Triple) matrix with the equation

$$WCT_{IJ} = \sum_{k=1}^{q} min(W_{ik}, W_{jk}) \tag{17}$$

$$Sim^{WCT}(i,j) = \frac{WCT_{IJ}}{WCT_{\max}} \tag{18}$$

with
q : Number of labels
$WCT_{\max}$: the highest value in the WCT matrix.

c. Create a similarity matrix
If this cluster entry is connected to the same cluster, then the entry will be 1 and 0 for the others. The similarity between two data points $x_i, x_j \in X$from m ensemble cluster members can be calculated as follows:

$$S_m(X_i, X_j) = \begin{cases} 1\ , jika\ C(X_i) = C(X_j) \\ 0,\ lainnya \end{cases} \tag{19}$$

with:
$S_m(X_i, X_j)$: The similarity value between the th object $i$and the th object $j$in the th grouping method algorithm m
$C(X_i) = C(X_j)$: Similarity value between the ith object label $i$and the th object label j

So, m matrices are combined to form a matrix CO. Each element in the matrix CO represents the degree of similarity between two data points assigned to the same cluster by the total number of ensemble cluster members. Formally, the similarity between $x_i, x_j \in X$is defined as:

$$CO(x_i, x_j) = \frac{1}{m}\sum_{m=1}^{m} S_m(x_i, x_j) \tag{20}$$

where, $m$is the number of cluster members formed.

H. Cluster Validation

Clusters that have been formed need to be validated, to find out whether the data is suitable to be divided into how many clusters. There are two categories of cluster validation, namely internal validation and external validation. External validation is validation that evaluates clusters based on the class labels given. External validation uses information that is not in the data, while internal validation only uses information that is in the data [33]. Several cluster validation methods that can be used are the Silhouette Score (SC), Dunn Index (DI), and Connectivity Index (CI)

1. Silhouette Score (SC)

This method is a cluster evaluation technique that integrates cohesion and separation methods [34]. Cohesion is quantified by counting all objects within a cluster, while separation is assessed by computing the average distance of each object in a cluster to its nearest neighboring cluster [35]. The Euclidean distance formula is used to calculate distances between data points. The Silhouette Score evaluates the appropriateness of each object's placement within its respective cluster. The Silhouette Score can be formulated as follows:

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{21}$$

with :

$$S_i = \frac{b_i - a_i}{\max\limits_{a_i, b_i}} \tag{22}$$

$$b_i = \min d_{i,c} \tag{23}$$

$$a_i = \frac{1}{|A|-1}\sum_{j \in A, j \neq i} d_{i,j} \tag{24}$$

With:

b_i : the minimum value of the average distance of the ith object to all objects in other clusters.

a_i : the average distance of the ith object to all objects in one cluster.

Table 1 Silhouette Score Classification

Coefficient Value	Classification
$-1 \leq SC < 0{,}2$	Bad
$0{,}2 \leq SC < 0{,}5$	Enough
$0{,}5 \leq SC < 1$	Good

Source: [35]

2. Dunn Index (DI)

Dunn Index (DI) is a validation function that is able to provide effective assessment results for applications that use several different clustering methods. The Dunn Index produces the best cluster if the Dunn value obtained is greater [36]. A large or high Dunn Index value indicates that the clusters formed have been divided between one cluster and another in an orderly and full or dense manner [37][37]. Dunn's index is defined as follows:

$$DI(c) = \min_{1 \leq i \leq c}\left\{\min_{1 \leq j \leq c}\left\{\frac{\delta(X_i, X_j)}{\max\limits_{1 \leq k \leq c}[\Delta(X_k)\}}\right\}\right\} \tag{25}$$

information:

$\delta(X_i, X_j)$	:	Distance between clusters
$\Delta(X_k)$	:	Intra-cluster distance
c	:	The number of clusters

3. Connectivity Index (CI)

This validation technique aims to assess how well the clusters formed comply with the concept of connectedness, namely the extent to which clusters observe local density and group data members together with their nearest neighbors in one cluster. The Connectivity Index forms the best number of clusters if the resulting value is smaller compared to the values from other clusters [38]. The density value is measured by the connectivity coefficient. The connectivity value is between zero and ∞. The C-index value is formulated as follows [39]:

$$C = \sum_{i-1}^{N}\sum_{j=1}^{L} x_i, nn_{i(j)} \tag{26}$$

information:

$nn_{i(j)}$	:	Distance between clusters
$x_i, nn_{i(j)}$	:	A measure of the closeness of i and j in one cluster
L	:	Parameters measure the number of neighbors

III. METHODOLOGY

A. Data source

This research uses secondary data regarding the quantity of cabbage production in 2023 in East Java Province. Research data comes from the East Java Province Agriculture and Food Security Service. The data used is cabbage production data in 18 districts/cities in East Java Province. This research uses the following variables.

Table 2 Variable Explanation

Variable	Variable Name	Unit
X_1	Cabbage Production	Tons
X_2	Harvest area	Hectare
X_3	Productivity	Production Ratio and Land Area

B. Data Analysis Steps

Data analysis was carried out using R Studio Software. The analysis steps carried out in this research are as follows:

1. Carry out descriptive analysis by finding the average, minimum value, maximum value and standard deviation of each variable.
2. Standardize data
3. Carrying out assumption testing, namely representative assumptions on the sample using the KMO test (Keiser Meyer Olkin) and assuming correlation between cluster variables using the Pearson correlation coefficient.
4. If the correlation assumption between cluster variables is not met, then further analysis is carried out, namely Principal Component Analysis (PCA).
5. Perform hierarchical cluster analysis using single linkage, complete linkage, and average linkage methods.
6. Conduct Self Organizing Maps (SOM) cluster analysis
7. Forming a consensus function
8. Grouping with ensemble clustering
9. Test the validity of Cgan using the Silhouette Score (SC), Dunn Index (DI), and Connectivity Index (CI).
10. Obtain final clustering results and cluster profiling for each cluster formed.

IV. RESULTS AND DISCUSSIONS

A. Data Exploration

This research uses secondary data regarding the quantity of cabbage production in 2023 in East Java Province. Research data comes from the East Java Province Agriculture and Food Security Service. Based on the data obtained, regional characteristic conditions will be grouped in 18 districts and cities in East Java Province according to cabbage production. A summary of data descriptions for each research variable can be reviewed in Table 3.

Table 3 Descriptive Analysis of Data

Variable	Mean	Minimum	Maximum	Standard Deviation	Range
X_1	488.56	1	4334	1064.85	4333
X_2	110653.56	32	1095340	262137.18	1095308
X_3	195.49	58.5	355	75.93	296.5

B. Data Standardization

The first step to forming appropriate clusters is to standardize the data. The result of data standardization in this research is an effort to change the original variable (X) into a standard number (Z). Data standardization is carried out if the variables in the research data have different units. The data standardization process can use the z-score shown in Equation (1).

C. Test Assumptions

Before grouping, two assumptions must be met, namely the assumption of representativeness in the sample and the a ssumption of correlation between cluster variables.

1. Assumption of sample representativeness

The assumption that the sample can reflect the population can be determined using the Kaiser Mayer Olkin (KMO) test with the equation in equation (2). The hypothesis of this KMO test is as follows:

H_0: The sample does not represent the population

H_1: The sample represents the population

Decision criteria:

H_0 rejected if KMO ≥ 0.5 means the sample represents the population

H_1 accepted if KMO < 0.5 means the sample does not represent the population

Table 4 Sample Representation test values with the KMO test

KMO Test		
Overall MSA	0.48	
$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
0.49	0.49	0.36

The KMO value = 0.48 < 0.5 is H_0 accepted, which means the sample taken does not represent the population. So, the m ethod used to overcome this is to delete variables with low MSA values, therefore they are deleted for the productivity variable. So, it is tested again for the KMO value

Table 5 KMO test value after deducting one variable

KMO Test	
Overall MSA	0.5
$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
0.5	0.5

The KMO value = 0.5 ≥ 0.5 is H_0rejected, which means the sample taken already represents the population. So it can be continued to the next stage. So it can be concluded that the sample used in this research is sufficient to represent the population in East Java Province.

2. Assumption of correlation between variables

Apart from the sample adequacy test, the assumption that must be met in conducting cluster analysis is the assumption of correlation between variables. Determining the correlation assumption between variables can use the Pearson correlation coefficient.

The Pearson correlation value can be calculated using the formula in Equation (3) with the following hypothesis:

$H_0: \rho \ = \ 0$ (There is no correlation between independent variables)

$H_1: \rho \ \neq \ 0$ (There is a correlation between independent variables)

Test Criteria:

H_0 rejected if $p - value < 0.05$

H_1 accepted if $p - value \geq 0.05$

By using R studio software, the correlation between variables is as follows:

Table 6 Pearson Correlation Coefficient Value

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
$\boldsymbol{X_1}$	1.0000000	0.9932971
$\boldsymbol{X_2}$	0.9932971	1.0000000

The p-value obtained is close to 1, so the test criteria are negative, H_0 meaning the correlation between the independent variables is classified as strong. The way to overcome a strong correlation between independent variables is to use PCA analysis.

D. Principal Component Analysis (PCA)

PCA is carried out to help overcome strong correlations between independent variables so that the assumption in cluster analysis can be met and can proceed to the next stage. The following are the explanatory components of data diversity resulting from PCA analysis.

Table 7 Explanatory Components of Data Diversity Result of PCA Analysis

Components (PC)	Eigenvalues	Total Variance	Cumulative Total Variance
First Component (PC1)	1.4118417	0.9966485	0.9966485
First Component (PC2)	0.081871330	0.003351457	1.000000000

In Table 7 it can be seen that the first principal component which has a value can explain 99.66% of the diversity of the data. The second component can explain 0.3% of the data variability. Cumulatively, the first and second components can explain 100% of the total diversity.

Table 8 PCA loadings values

	Comp 1	Comp 2
X_1	0.707	-0.707
X_2	0.707	0.707

After obtaining the two main components from the PCA analysis, the coefficients of the main components will be determined which will form the main components equation. The equation formed is:

$$pc1 \ = \ 0.707\,X_1 + \ 0.707\,X_2$$
$$pc2 \ = \ -0.707\,X_1 + \ 0{,}707\,X_2$$

The pc1 and pc2 equations above are used to find the values of the main components. This main component score will be used to group districts/cities in East Java Province based on cabbage production levels using the hierarchical analysis method, Self Organizing Method, and cluster ensemble because the two components formed are free from strong correlation.

E. Hierarchical Cluster

Hierarchical clustering was carried out using RStudio software with the 'NbClust' package. The linkages used in this hierarchical analysis are single linkage, average linkage, and complete linkage. The range of the number of clusters to be formed is determined as 3 to 5 clusters. The distance matrix used is Euclidean distance. Meanwhile, the metrics chosen to measure the goodness of clusters are Silhouette Coefficient (SC), Dunn Index (DI), and Connectivity Index (CI). The Silhouette Coefficient and Dunn Index are optimal at the largest value, while the Connectivity index is optimal at the smallest value. **Table 9** shows the results of the cluster goodness metric index values for each type of hierarchical cluster linkage.

Table 9 Cluster goodness of fit metric index in hierarchical clusters

Clusters	Complete			Single			Average		
	SC	DI	CI	SC	DI	CI	SC	DI	CI
2	**0.8427**	1.4938	**2.9290**	**0.8492**	1.5178	**2.9290**	**0.8492**	1.5178	**2.9290**
3	0.7707	2.9178	5.8579	0.7699	2.6747	5.8579	0.7699	2.6747	5.8579
4	0.6737	0.9137	10.4905	0.6865	0.8130	10.7405	0.6624	0.3700	12.2528
5	0.6615	0.9141	13.0738	0.6382	0.7466	13.3238	0.6394	0.5450	15.5861

In clustering analysis using three different linking methods, namely complete linkage, single linkage, and average linkage, an evaluation of the optimal number of clusters is carried out by paying attention to the optimal values of the three metrics measured, namely Silhouette Coefficient (SC), Dunn Index (DI), and Connectivity Index (CI). The results show that for complete linkage, the optimal number of clusters is 2, with a SC value of 0.8427, DI of 1.4938, and CI of 2.9290. For single linkage, the optimal number of clusters is also 2, with a SC value of 0.8492, DI of 1.5178, and CI of 2.9290. Meanwhile, for average linkage, the optimal number of clusters remains 2, with a SC value of 0.8492, DI of 1.5178, and CI of 2.9290.

By considering this optimal value and using the majority vote from the three metrics measured, it can be concluded that the optimal number of clusters for this dataset, whether using the complete linkage, single linkage, or average linkage methods, is 2 clusters. This shows that dividing the data into two groups provides the most consistent and optimal results according to these three metrics.

F. Self Organizing Method (SOM) cluster

Before carrying out SOM clustering analysis, a cluster validity test is first carried out to determine the best number of clusters to use. The cluster validity test used is an internal validation test using the Dunn, Sillhoutte, and Connectivity indices. The validity test results can be seen in Table 7.

Table 10 Cluster goodness of fit metric index in SOM

Clusters	SC	DI	CI
2	**0.6095**	**0.4818**	**4.8579**
3	0.4356	0.1663	11.1401
4	0.3512	0.1451	17.6091
5	0.3534	0.3370	18.6091

Based on Table 10 Overall, these results show that clustering with 2 clusters in SOM is the most optimal, with higher Silhouette Coefficient (SC) and Dunn Index (DI) values and lower Connectivity Index (CI) compared to the number of clusters. the greater one. This indicates that 2 clusters provide the best cohesion and separation between data, with minimal fragmentation and connectivity between clusters. In the process of the SOM algorithm producing a SOM Model using the R program, it will produce a fan diagram as in Figure 2.

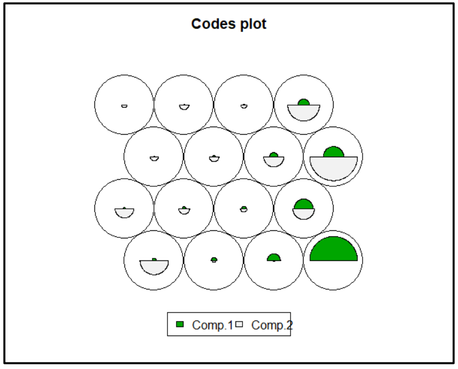

Figure 2. Fan diagram of SOM

The image above is a Codes Plot of the Self-Organizing Map (SOM) showing how the two main components (Comp.1 and Comp.2) are distributed across the SOM grid. Each circle in the figure represents a neuron on the grid, and the green and white segments within the circle represent the average value of each component in that neuron. Neurons with large green segments (Comp.1) indicate that the value of the first component is high in that neuron, while white segments indicate the value of the second component (Comp.2). For example, neurons in the bottom left and top left have large green segments, indicating high values for Comp.1. In contrast, the neurons in the upper right have very small green segments, indicating low values for Comp.1. This distribution helps identify patterns or clusters in the original data. Neurons with similar segments tend to represent similar data. Thus, these plots provide insight into how the original data is structured and help understand the results of clustering performed by SOM.

G. Cluster Ensembles

Ensemble clustering forms a set of clustering results obtained from different methods as ensemble members. Clustering using the ensemble method was carried out using RStudio software with the 'diceR' package. Some important parameters that need to be set are the clustering algorithm and the consensus function that will be used. The clustering algorith m used is a hierarchical cluster with complete linkage and SOM. The chosen consensus function is CSPA (Cluster-based Similarity Partitioning Algorithm).

Table 11 Cluster ensemble result metric index

Clusters	SC	DI	CI
2	**0.9124162**	**1.37344**	**2.928968**
3	0.7828166	0.1837872	9.740476
4	0.7942755	0.4425576	10.74048
5	0.7942755	0.6556859	19.79524

Based on Table 11, the cluster results show that grouping with 2 clusters is the most optimal. The Silhouette Coefficien t (SC) value is 0.9124162, the Dunn Index (DI) value is 1.37344, and the Connectivity Index (C I) value is 2.928968. It can be concluded that ensemble clustering with 2 clusters shows the most optimal results, with better SC, DI, and CI values compared to a larger number of clusters. This shows that the ensemble method can produce better and more stable grou ping in 2 clusters compared to a larger number of clusters.

H. Determination of the best cluster

To obtain optimal grouping results among the five methods that have been used to categorize districts and cities in East Java Province based on the level of cabbage production, the five methods need to be compared based on the level of goodness of the grouping results. The goodness of the grouping results can be reviewed based on the Silhouette Score (SC), Dunn Index (DI), and Connectivity Index (CI) values. The resulting values for the goodness of fit of the cluster method are presented in Table 12.

Table 12 Comparison of cluster indices

Method	Clusters	SC	DI	CI
Complete Linkage Hierarchy Analysis	2	0.8427	1.4938	2.9290
Average Linkage Hierarchy Analysis	2	0.8492	1.5178	2.9290
Single Linkage Hierarchy Analysis	2	0.8492	1.5178	2.9290
SOM	2	0.6095	0.4818	4.8579
Cluster ensembles	**2**	**0.9124162**	**1.37344**	**2.9289**

Table 12 displays a comparison of cluster indices from several clustering methods, namely Hierarchical Analysis (Complete Linkage, Average Linkage, and Single Linkage), Self-Organizing Map (SOM), and Ensemble Cluster, each for 2 clusters. The indices compared include the Silhouette Coefficient (SC), Dunn Index (DI), and Connectivity Index (CI). The Ensemble Cluster shows the best performance with a SC value of 0.9124, DI of 1.3734, and CI of 2.9290. This method has the highest SC value among all methods, indicating excellent cluster separation. Although the DI value is slightly lower compared to the hierarchical method, it is still within a good range. The CI generated by the Ensemble Cluster is similar to the hierarchical method, showing optimal inter-cluster connectivity.

Based on these results, it can be concluded that the Ensemble Cluster method is the most optimal for grouping cabbage production data in East Java, followed by Hierarchical Analysis with the Complete, Average, and Single Linkage methods. The SOM method has the lowest performance in this context. Ensemble Cluster produces excellent cluster separation with optimal internal cohesion and external separation. The following are the results of the members of the cluster ensemble. After knowing the best method and the optimum number of clusters for grouping districts and cities in East Java, the distribution is then carried out for each cluster. The distribution results for each cluster group are presented in Table 13.

Table 13 Distribution of cluster members

Clusters	Cluster Member
1	Pacitan, Treggalek, Tulungagung, Blitar, Kediri, Jember, Banyuwangi, Mojokerto, Jombang, Ngawi, Pamekasan, Batu City, Lumajang, Bondowoso, Pasuruan, Magetan
2	Malang

Based on the grouping results obtained in Table 13 it can be seen that Cluster 2 only consists of one member, namely Malang Regency. This is because Malang Regency is a district with outlier data conditions on the cabbage production variable. The regional grouping based on the results of cluster analysis using cabbage production levels is visualized on the map of East Java Province in Figure 3.

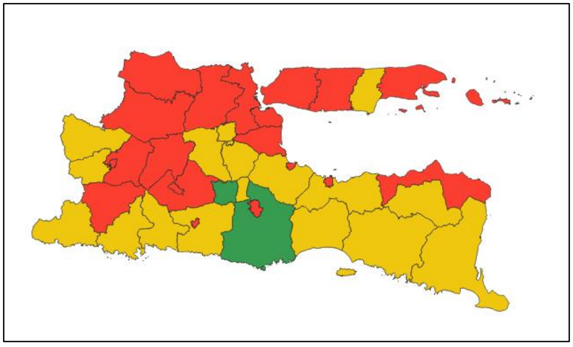

Figure 3 Regional cluster mapping based on cabbage production levels in East Java Province

The following is an interpretation of the clustering map of cabbage production in East Java Province based on different cluster colors. Cluster 0, marked in red, includes districts and cities with no cabbage production. Cluster 1, marked in yellow, indicates areas with low levels of cabbage production. Cluster 2, marked in green and only including Malang Regency, stands out with very high or significantly different levels of cabbage production. The distribution of clusters 0 and 1 is spread across the province, while cluster 2 only consists of one district, indicating the unique characteristics of cabbage production in each region. This grouping provides guidance for planning and decision-making regarding cabbage farming in East Java, focusing on the specific needs of each cluster for optimal development.

V. CONCLUSIONS AND SUGGESTIONS

Based on the analysis and discussion above, it can be concluded that to determine the optimal number of clusters and the best method, three cluster validation metrics were utilized: Silhouette Score (SC), Dunn Index (DI), and Connectivity Index (CI). The ensemble cluster method was identified as the best method with two clusters as the optimal number in this study. This conclusion is based on an SC value of 0.9124, a DI of 1.3734, and a CI of 2.9290. The characteristics of cabbage production levels in each region vary greatly. Regions in Cluster 1 have low production numbers and harvested area, while regions in Cluster 2 have high cabbage production numbers and harvested area.

By understanding the clustering of regencies and cities in East Java and formulating recommendations for each cluster, this study is expected to serve as a consideration for the government and relevant institutions to address the main problems that still occur in each regency and city. Thus, the potential cabbage production in each area can be identified, optimizing regions with high cabbage production, and exploring the potential of other crops in areas with low cabbage production. This study applies optimal clustering from multivariate data sets using the best cluster method. The results of this study are logical and consistent when compared to the actual conditions of cabbage production in regencies and cities in East Java at present.

REFERENCES

[1] S. Nugroho, *Statistika Mutivariat Terapan*, Edisi 1. Bengkulu: UNIB Press, 2008.

[2] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.: Pearson Prentice Hall*, Sixth Edit. upper saddle river: Perason Education Inc, 2007.

[3] W. Widyawati, W. L. Y. Saptomo, and Y. R. W. Utami, "Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan," *J. Ilm. SINUS*, vol. 18, no. 1, p. 75, 2020, doi: 10.30646/sinus.v18i1.448.

[4] U. Wagschal, *Cluster analysis*. John Wiley & Sons, 2016.

[5] I. Wahyuni and S. P. Wulandari, "Pemetaan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesejahteraan Rakyat Menggunakan Analisis Cluster Hierarki," *J. Sains dan Seni*, vol. 11, no. 1, pp. D70–D75, 2022.

[6] S. Kania, D. Rachmatin, and J. A. Dahlan, "Program Aplikasi Pengelompokan Objek Dengan Metode Self Organizing Map Menggunakan Bahasa R," *J. EurekaMatika*, vol. 7, no. 2, pp. 17–29, 2019.

[7] M. H. Ghaseminezhad and A. Karami, "A novel self-organizing map (SOM) neural network for discrete groups of data clustering," *Appl. Soft Comput. J.*, vol. 11, no. 4, pp. 3771–3778, 2011, doi: 10.1016/j.asoc.2011.02.009.

[8] K. Millati, C. Suhaeni, and B. Susetyo, "Penggerombolan Daerah 3T di Indonesia Berdasarkan Rasio Tenaga Kesehatan dengan Metode Penggerombolan Berhierarki dan Cluster Ensemble," *Xplore J. Stat.*, vol. 10, no. 2, pp. 197–213, 2021, doi: 10.29244/xplore.v10i2.744.

[9] A. Strehl, "Cluster Ensembles – A Knowledge Reuse Framework for," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[10] Y. Reinaldi, N. Ulinnuha, and M. Hafiyusholeh, "Comparison of Single Linkage, Complete Linkage, and Average Linkage Methods on Community Welfare Analysis in Cities and Regencies in East Java," *J. Mat. Stat. dan Komputasi*, vol. 18, no. 1, pp. 130–140, 2021, doi: 10.20956/j.v18i1.14228.

[11] P. Melin, J. C. Monica, D. Sanchez, and O. Castillo, "Analysis of Spatial Spread Relationships of Coronavirus (COVID-19) Pandemic in the World using Self Organizing Maps," *Chaos, Solitons and Fractals*, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109917.

[12] S. Y. Hadist and A. P. Utomo, "Pengelompokan Kabupaten/Kota di Pulau Jawa Berdasarkan Kondisi Sosial Ekonomi Sebelum dan Setelah Memasuki Pandemi COVID-19 Penerapan Metode Cluster Ensemble," *Semin. Nas. Off. Stat.*, vol. 19, no. 2020, pp. 322–332, 2021.

[13] A. N. Haloho, "RESPON PERTUMBUHAN DAN PRODUKSI KUBIS (Brassica oleraceae.L) DENGAN PEMBERIAN BERBAGAI JENIS DAN DOSIS PUPUK KANDANG," *Agroprimatech*, vol. 4, no. 1, pp. 10–17, 2020, doi: 10.34012/agroprimatech.v4i1.1325.

[14] Basri and Syarli, "AHP-Standar Score: Pendekatan Baru Dalam Sistem Pemeringkatan," *J. Keteknikan dan Sains – LPPM UNHAS*, vol. 1, no. 1, pp. 1–6, 2018.

[15] R. E. Walpole, R. H. Myers, and Sharon L. Myers Radford, *Probability and Statistics*, 9th ed. Boston, MA: Pearson Education, Inc., 2012.

[16] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, "Multivariate Data Analysis," *Polymers*, vol. 12, no. 12. Annabel Ainscow, United Kingdom, pp. 1–18, 2019, doi: 10.3390/polym12123016.

[17] J. Costales, J. J. J. Catulay, J. Costales, and N. Bermudez, "Kaiser-Meyer-Olkin Factor Analysis: A Quantitative Approach on Mobile Gaming Addiction using Random Forest Classifier," *ACM Int. Conf. Proceeding Ser.*, no. March, pp. 18–24, 2022, doi: 10.1145/3546157.3546161.

[18] F. Jabnabillah and N. Margina, "ANALISIS KORELASI PEARSON DALAM MENENTUKAN HUBUNGAN ANTARA MOTIVASI BELAJAR DENGAN KEMANDIRIAN BELAJAR PADA PEMBELAJARAN DARING," *J. Sintak*, vol. 1, no. 1, pp. 14–18, 2022.

[19] M. I. R. Suyantiningsih, "KORELASI ANTARA PERSEPSI PENGELOLAAN DAN LAYANAN PUSTAKA DENGAN MOTIVASI BELAJAR DI DIGITAL LIBRARY UNY," *Comput. Handb. Two-Volume Set*, vol. 1, pp. 1–23, 2020, doi: 10.1201/b16768-21.

[20] S. Hussain, N. Z. Quazilbash, S. Bai, and S. Khoja, "Reduction of variables for predicting breast cancer survivability using principal component analysis," *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2015-July, pp. 131–134, 2015, doi: 10.1109/CBMS.2015.62.

[21] T. H. Dang, T. D. Pham, H. L. Tran, and Q. Le Van, “Using dimension reduction with feature selection to enhance accuracy of tumor classification,” *BME-HUST 2016 - 3rd Int. Conf. Biomed. Eng.*, pp. 14–17, 2016, doi: 10.1109/BME-HUST.2016.7782082.

[22] M. Z. Nasution, “PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA) DALAM PENENTUAN FAKTOR DOMINAN YANG MEMPENGARUHI PRESTASI BELAJAR SISWA (Studi Kasus : SMK Raksana 2 Medan),” *J. Teknol. Inf.*, vol. 3, no. 1, p. 41, 2019, doi: 10.36294/jurti.v3i1.686.

[23] D. P. Isnarwaty and I. Irhamah, “Text Clustering pada Akun TWITTER Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN) dan K-Means,” *J. Sains dan Seni ITS*, vol. 8, no. 2, pp. 2–9, 2020, doi: 10.12962/j23373520.v8i2.49094.

[24] M. Ozcalici and M. Bumin, “An integrated multi-criteria decision making model with Self-Organizing Maps for the assessment of the performance of publicly traded banks in Borsa Istanbul,” *Appl. Soft Comput. J.*, vol. 90, p. 106166, 2020, doi: 10.1016/j.asoc.2020.106166.

[25] H. Hartatik and A. S. D. Cahya, “Clusterisasi Kerusakan Gempa Bumi di Pulau Jawa Menggunakan SOM,” *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 2, no. 02, pp. 25–34, 2020, doi: 10.46772/intech.v2i02.286.

[26] R. D. Kusumah, B. Warsito, and M. A. Mukid, “Perbandingan Metode K-Means Dan Self Organizing Map (Studi Kasus: Pengelompokan Kabupaten/Kota Di Jawa Tengah Berdasarkan Indikator Indeks Pembangunan Manusia 2015),” *J. Gaussian*, vol. 6, no. 3, pp. 429–437, 2017, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian.

[27] I. Hidayatin, S. Adinugroho, and C. Dewi, “Pengelompokan Wilayah berdasarkan Penyandang Masalah Kesejahteraan Sosial (PMKS) dengan Optimasi Algoritme K-Means menggunakan Self Organizing Map (SOM),” *J. Pengemb. Teknol. Inf. dan Ilmu Kompter*, vol. 3, no. 8, pp. 2548–964, 2019, [Online]. Available: http://j-ptiik.ub.ac.id.

[28] E. Fauziyari and D. U. Wustqa, “Pemetaan Kabupaten/Kota di Provinsi Papua Berdasarkan Indikator Daerah Tertinggal Dengan Metode Ensemble Clustering,” *J. Stat. Dan Sains Data*, vol. 1, pp. 40–55, 2023, [Online]. Available: https://journal.student.uny.ac.id/index.php/jssd.

[29] N. Aini, A. Lestari, M. N. Hayati, F. Deny, and T. Amijaya, “Analisis cluster pada data kategorik dan numerik dengan pendekatan Cluster Ensemble (Studi kasus : puskesmas di Provinsi Kalimantan Timur kondisi Desember 2017),” *J. EKSPONENSIAL Vol. 11*, vol. 11, pp. 117–126, 2020.

[30] A. Serra and R. Tagliaferri, “Unsupervised learning: Clustering,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January, pp. 350–357, 2018, doi: 10.1016/B978-0-12-809633-8.20487-1.

[31] X. Z. Fern and W. Lin, “Cluster ensemble selection,” *Soc. Ind. Appl. Math. - 8th SIAM Int. Conf. Data Min. 2008, Proc. Appl. Math. 130*, vol. 2, pp. 787–797, 2008, doi: 10.1137/1.9781611972788.71.

[32] A. A. Yusfar, M. A. Tiro, and S. Sudarmin, “Analisis Cluster Ensemble dalam Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Kinerja Pembangunan Ekonomi Daerah,” *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 3, no. 1, p. 31, 2020, doi: 10.35580/variansiunm14626.

[33] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal clustering validation measures,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 911–916, 2010, doi: 10.1109/ICDM.2010.35.

[34] S. Paembonan and H. Abduh, “Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat,” *PENA Tek. J. Ilm. Ilmu-Ilmu Tek.*, vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.

[35] B. Wira, A. E. Budianto, and A. S. Wiguna, “IMPLEMENTASI METODE K-MEDOIDS CLUSTERING UNTUK MENGETAHUI POLA PEMILIHAN PROGRAM STUDI MAHASIWA BARU TAHUN 2018 DI UNIVERSITAS KANJURUHAN MALANG,” *RAINSTEK J. Terap. Sains Teknol.*, vol. 1, no. 3, pp. 53–68, 2019, doi: 10.21067/jtst.v1i3.3046.

[36] A. F. Khairati, A. . Adlina, G. . Hertono, and B. . Handari, “Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA,” *Prism. Pros. Semin. Nas. Mat.*, vol. 2, pp. 161–170, 2019, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/article/view/28906.

[37] R. A. Hendrawan, A. Utamima, and D. A. Savitri, “Segmentasi Trafo Lisrik Menggunakan Algoritma K-Means untuk Mendukung Evaluasi Kapasitas Gardu Induk Listrik di Jawa Timur,” *J. Sist. Inf.*, vol. 5, no. 5, pp. 702–707, 2016.

[38] N. N. Halim and E. Widodo, “Clustering dampak gempa bumi di Indonesia menggunakan kohonen self organizing maps,” *Pros. SI MaNIS (Seminar Nas. Integr. Mat. dan Nilai Islam.*, vol. 1, no. 1, pp. 188–194, 2017, [Online]. Available: http://conferences.uin-malang.ac.id/index.php/SIMANIS/article/view/62.

[39] W. D. Ray, P. J. Brockwell, and R. A. Davis, *Time Series: Theory and Methods.*, vol. 153, no. 3. 1990.