BAREKENG: Journal of Mathematics and Its Applications
June 2025 Volume 19 Issue 2 Page 1057–1070
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss2pp1057-1070

INTEGRATION OF HIERARCHICAL CLUSTER, SELF-ORGANIZING MAPS, AND ENSEMBLE CLUSTER WITH NAÏVE BAYES CLASSIFIER FOR GROUPING CABBAGE PRODUCTION IN INDONESIA

Maulidya Maghfiro[1*], Ni Wayan Surya Wardhani[2], Atiek Iriany[3]

[1,2,3] *Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya*
Jl. Veteran No.10-11, Ketawanggede, Kec. Lowokwaru, Malang, 65145, Indonesia

*Corresponding author's e-mail: * **maulidyamaghfiroh56@gmail.com***

Article History:

Received: 3rd September 2024
Revised: 25th January 2025
Accepted: 26th February 2025
Published: 1st April 2025

Keywords:

Cabbage;
Cluster Ensembles;
Hierarchical Analysis;
Naïve Bayes;
SOM.
.

ABSTRACT

The purpose of this study is to evaluate and compare different clustering techniques, including hierarchical cluster analysis (using complete linkage, average linkage, and single linkage methods), Self-Organizing Maps (SOM) clustering, and ensemble clustering, within the framework of integrated cluster analysis combined with Naïve Bayes analysis, specifically applied to cabbage production in Indonesia. The data utilized in this study are on cabbage production from various districts and cities in Indonesia, obtained from the 2023 publications of the Central Statistics Agency (BPS). The variables used in this study are cabbage harvest, cabbage production, area height, and rainfall. The data size used is 157 districts/cities in Indonesia. This research is a quantitative analysis employing integrated cluster analysis combined with Naïve Bayes. Cluster analysis is used to obtain classes in each district/city. Different clustering methods, including hierarchical clustering, Self-Organizing Map (SOM), and ensemble clustering, are compared to determine the best approach for grouping districts based on cabbage production. Naïve Bayes analysis is then used to classify cabbage production in Indonesia and identify the optimal clusters. This comparison aims to find the most effective clustering method for improving grouping accuracy and understanding cabbage production patterns. The best method for classifying cabbage production in Indonesia is the ensemble clustering approach integrated with Naïve Bayes, resulting in three distinct clusters: high, medium, and low production clusters.

How to cite this article:

M. Maghfiro, N. W. S. Wardhani and A. Iriany., "INTEGRATION OF HIERARCHICAL CLUSTER, SELF ORGANIZING MAPS, AND ENSEMBLE CLUSTER WITH NAÏVE BAYES CLASSIFIER FOR GROUPING CABBAGE PRODUCTION IN INDONESIA," *BAREKENG: J. Math. & App.*, vol. 19, iss. 2, pp. 1057-1070, June, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article· Open Access

1. INTRODUCTION

Data mining is the application of mathematics, statistical techniques, machine learning, and artificial intelligence to describe and identify potential and useful knowledge information contained in a database. Data mining is also described as a component of the process of extracting knowledge from a database, commonly known as Knowledge Discovery in Databases (KDD) [1]. As data collection grows, incomplete labels or response variables can hinder analysis, introducing semi-supervised learning. Semi-supervised learning is a machine learning approach that integrates both supervised and unsupervised learning methods [2]. One of the techniques in semi-supervised learning involves integrating clustering and classification algorithms. Unlabeled data is grouped using clustering techniques, and the resulting clusters are used as labels for further classification. This Cluster process provides significant advantages over classification techniques that help identify groups of data that have the same characteristics at the beginning and increase accuracy and detection rates [3].

Cluster analysis is widely used in multivariate analysis to group objects based on their similarities, employing either hierarchical or non-hierarchical methods. In hierarchical clustering, similar data points are grouped within the same hierarchy, while dissimilar data points are placed in separate hierarchies [4]. This method also offers the benefit of displaying results in a dendrogram, making the grouping more informative [5]. On the other hand, Self-Organizing Maps (SOM) is a non-hierarchical clustering technique rooted in artificial neural networks (ANN) [6]. Previous studies [7] have shown that SOM achieves relatively high accuracy in clustering various entities such as locations, areas, regions, and objects. This makes SOM a powerful tool for clustering analysis across a wide range of applications.

Different clustering methods may yield varying results, and cluster ensemble techniques, such as those introduced by Strehl, aim to combine outcomes from multiple methods for better results. Cluster ensemble is a technique used to merge various clustering outcomes without relying on the initial data's characteristics [8]. The core idea of cluster ensemble is integrating results from different clustering methods. Strehl's research suggests that cluster ensemble can lead to higher-quality clustering results.

One of the classification algorithms is the Naïve Bayes Classifier. This algorithm employs probability and statistical methods to forecast future outcomes based on past experiences, utilizing Bayes' Theorem [9]. A key advantage of this method is its ability to perform classification with a relatively small sample or dataset [10]. Moreover, despite its complexity, Naïve Bayes is an efficient algorithm capable of managing multiclass data. This makes Naïve Bayes a suitable and contemporary choice for classification analysis. Previous research on the Naïve Bayes Classifier algorithm was conducted by [11] to classify cases of stunting in toddlers. This study achieved an accuracy of 88.46%, so it can be concluded that the Naïve Bayes Classifier model is considered sufficiently effective for classifying data on the nutritional status of stunting toddlers. Research on integrating clustering with classification was conducted by [12] in the study titled "A Hybrid Approach: Utilizing K-means Clustering and Naïve Bayes for IoT Anomaly Detection." The objective of this study was to combine the K-Means clustering algorithm and the Naïve Bayes classification algorithm. The study achieved an accuracy ranging from 90% to 100%.

One application of clustering is in the grouping of cabbage production. Cabbage is a widely cultivated horticultural commodity. It is classified as a vegetable plant with a single growing season or a short lifespan, producing only once before completing its life cycle. Harvesting of cabbage generally occurs when the plants are 60-80 days old from planting [13]. Grouping of cabbage plants is necessary because cabbage production can vary significantly between regions and time. By grouping cabbage, groups of regions with similar production characteristics can be identified so that management strategies can be adjusted to the specific needs of each group. In addition, clustering analysis can show several regions that have production far above average, thus allowing for cabbage business opportunities and maximizing the profits of cabbage farmers.

This study extends previous research by integrating ensemble clusters and Naïve Bayes classification. This study aims to provide an enhanced analysis through the integration of both algorithms. The purpose of clustering and classification of cabbage production levels in Indonesia is to obtain optimal clusters and high classification accuracy values, which can be used to assist the government in addressing planning and decision-making related to increasing cabbage production in the region. This study is also expected to provide deeper insights for policymakers in developing sustainable programs to support cabbage farmers in Indonesia.

2. RESEARCH METHODS

2.1 Data

This study utilizes secondary data on cabbage production in Indonesia for 2023. The research data is sourced from publications by the Central Statistics Agency, covering various districts and cities across the country. Specifically, the data includes cabbage production figures from 157 districts and cities in Indonesia. The study incorporates the following variables.

Table 1. Explanation of Variables

Variables	Variable Name	Unit
X_1	Harvest Area	Hectare
X_2	Production	Quintal
X_3	Area Height	Masl
X_4	Rainfall	Millimeters/year

2.2 Data Standardization

Data standardization is a technique used to normalize variables, ensuring that differences in scale do not affect the grouping of objects. A widely used method for standardization is the z-score, which standardizes data by subtracting the mean and dividing by the standard deviation. The z-score method's advantage is its ability to evaluate the relative quality of information compared to the group average based on the standard deviation [14]. The z-score is calculated using the following formula:

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

where:

Z_i : variable standardization
x_i : the i-th data point
$\bar{x}$: the average of all data for each variable
s : standard deviation

2.3 Euclidean Distance

Euclidean distance is a commonly used method for measuring distance due to its simplicity and ease of interpretation. It calculates the distance between a data object and the center of a cluster, making it effective for determining the nearest distance between two data points. This metric represents the geometric distance between two data objects. The Euclidean distance between two points is calculated using the following equation [15].

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

where:

$d(x_i, y_j)$: distance between object i and object j
x_{ik} : the value of object i in variable k
y_{jk} : the value of object j in variable k
p : number variables are observed

2.4 Hierarchical Clusters

Hierarchical clustering groups similar data within the same hierarchy, while dissimilar data is placed in more distant hierarchies. The hierarchical clustering process begins by treating each object as its own cluster. Then, the two objects with the closest distance are combined into a single cluster. Subsequently, a third object either joins the existing cluster or forms a new cluster with another object, continuing to group based on proximity. This process continues until all objects are grouped into a single cluster, producing a dendrogram.

Two commonly used methods in hierarchical clustering are agglomerative hierarchical clustering and divisive clustering **[16]**. Agglomerative clustering starts with N individual clusters (where N is the total number of data points) and merges them progressively into a single cluster. In contrast, divisive clustering begins with one large cluster and splits it into N smaller clusters.

2.4.1 Single Linkage Method

The single linkage clustering method is one of the clustering algorithms that groups data based on the nearest neighbor distance **[17].** The input for this algorithm can be in the form of similarity distances between pairs of objects. Groups are formed from individual entities by merging those with the shortest distance or highest similarity. The formula for the single linkage method is as follows:

$$d_{(AB)C} = min(d_{\mathrm{AC}}, d_{BC}) \tag{3}$$

where:
$d_{(AB)C}$: the distance between cluster (AB) and cluster C
d_{AC} : the distance of object A within cluster C
d_{BC} : the distance of object B within cluster C

2.4.2 Complete Linkage Method

Complete linkage provides certainty that all items in one cluster are the furthest distance (least similarity) from each other **[18]**. Find the furthest distance but take the smallest value. The formula for complete linkage is as follows:

$$d_{(AB)C} = ma\,x(d_{AC}, d_{BC}) \tag{4}$$

where:
$d_{(AB)C}$: the distance between cluster (AB) and cluster C
d_{AC} : the distance of object A within cluster C
d_{BC} : the distance of object B within cluster C

2.4.3 Average Linkage Method

Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of the pair belongs to each cluster **[19]**. The formula for average linkage is as follows:

$$d_{(AB)C} = \frac{\sum_A \sum_B d_{AB}}{N_{AB} N_c} \tag{5}$$

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

2.5 Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) was first introduced by Teuvo Kalevi Kohonen in 1982. SOM is a technique within artificial neural networks (ANN) designed to organize data into distinct clusters or groups. Data points with similar characteristics are grouped together. The algorithm for grouping network patterns using SOM involves the following stages [20]:

1. **Initialize the Weights:** Randomly initialize the weights W_{ji}, where the columns of the weight matrix denote the number of elements in a vector, and the rows correspond to the maximum number of clusters to be created.
2. **Calculate the Distance:** Compute the distance D_j between the input values and the weights using the Euclidean distance, as outlined in **Equation (2)**.
3. **Determine the Minimum Value:** Identify the minimum value by analyzing the results of the distance vector D_j. The weights are then updated using the following formula [21]:

$$W_{ji}(new) = W_{ji}(old) + \alpha\left[x_i - W_{ji}(old)\right] \tag{6}$$

4. **Update Weights:** During the weight update stage, the learning rate is α required, where $0 \leq \alpha \leq 10$. The learning rate can be adjusted using the following formula [22]:

$$\alpha(t) = \alpha_i\left(1 - \frac{t}{t_{max}}\right) \tag{7}$$

where:

α_i : initial learning rate value i
t : number of iterations
t_{max} : maximum number of iterations

5. **Convergence Test:** The stopping condition is assessed by comparing the new weights W_{ji} (new) with the old weights W_{ji} (old). If the weights have not changed significantly, or the changes are minimal, it indicates that the test has stopped and the algorithm has reached convergence.

2.6 Ensemble Cluster

In 2002, Strehl introduced a method called Cluster Ensemble, designed to combine multiple clustering solutions to enhance the quality and robustness of cluster outcomes. This approach in cluster analysis is known as Cluster Ensemble or consensus clustering [23]. The technique works by integrating various solutions from different clustering methods to arrive at an improved final cluster solution [24]. Cluster ensemble offers advantages in four key aspects:

1. Robustness: Enhances performance quality.
2. Novelty: Generates solutions that are independent of any single algorithm.
3. Stability and confidence estimation: Exhibits low sensitivity to noise, outliers, and variations caused by sampling.
4. Parallelization and scalability: Enables the integration of multiple clustering results using the parallelization technique.

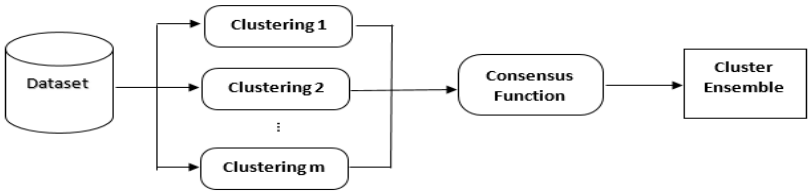

Figure 1. Clustering Steps with Cluster Ensemble
Source : [25]

The following outlines the algorithm and provides an illustration of the calculation for the Cluster-based Similarity Partitioning Algorithm [26]:

1. Forming ensemble members involves the relabelling of datasets $x = \{x_1, x_2, \dots, x_p\}$ and applying clustering methods $\prod = \{\pi_1, \pi_2, \dots, \pi_m\}$.
2. The relabeling process in the transformation results in an ordered matrix of size $n\ x\ p$ for each ensemble cluster member, denoted by S_m with $m = 1, 2, \dots, n$. Each entry in this matrix represents the relationship between two data points.
3. Form a weighting matrix with the following steps:
   a. Forming a matrix (W) with the equation

$$W_{ij} = \frac{\left|X_{c_i} \cap X_{c_j}\right|}{\left|X_{c_i} \cap X_{c_j}\right|} \tag{8}$$

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

$$WCT_{IJ} = \sum_{k=1}^{q} \min\left(W_{ik}, W_{jk}\right) \tag{9}$$

$$Sim^{WCT}(i,j) = \frac{WCT_{IJ}}{WCT_{\max}} \tag{10}$$

   where:

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

   c. Creating a similarity matrix

   If this cluster entry is associated with the same cluster, the entry will be marked as 1; otherwise, it will be marked as 0. The similarity between the two data points $x_i, x_j \in X$, derived from m members of the cluster ensemble, can be calculated as follows:

$$S_m\left(X_i, X_j\right) = \begin{cases} 1, if\ C(X_i) = C(X_j) \\ 0, \quad otherwise \end{cases} \tag{11}$$

   where:

   $S_m\left(X_i, X_j\right)$: the similarity value between object i and object j in the clustering algorithm of method m
   $C(X_i) = C(X_j)$: the similarity value between the label of object i and the label of object j

Consequently, the m matrices are combined to form a CO matrix. Each element in the CO matrix represents the degree of similarity between two data points based on how frequently they have been assigned to the same cluster across all members of the ensemble cluster. Formally, the similarity between $x_i, x_j \in X$ is defined as follows:

$$CO\left(x_i, x_j\right) = \frac{1}{m} \sum_{m=1}^{m} S_m\left(x_i, x_j\right) \tag{12}$$

with m is the number of cluster members formed.

2.7 Naïve Bayes Classifier

Naïve Bayes classification is a method grounded in Bayes' theorem [27]. Developed by British statistician Thomas Bayes, Naïve Bayes applies probabilistic and statistical techniques for categorization. It estimates the likelihood of future events based on past experiences [9]. The underlying assumption of Naïve Bayes is that attributes are conditionally independent, meaning that the presence or absence of specific characteristics of a class does not affect the characteristics of other classes. The Bayes' theorem equation is expressed as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{13}$$

where:

X : data with an unknown class or label
H : the hypothesis of data X is a specific class
$P(H|X)$: the probability of hypothesis H conditional on X
$P(X|H)$: the probability of hypothesis X conditional on H
$P(H)$: probability H (prior probability)
$P(X)$: probability of X (evidence)

The advantage of this method is that it requires only a small sample size or limited data to perform the classification process for predictive purposes [10]. To clarify Bayes' theorem, it is important to recognize that the classification process relies on several indicators to determine the most appropriate class for the sample under analysis. Consequently, Bayes' theorem is adjusted as follows:

$$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1, \dots, F_n|C)}{P(F_1,\dots,F_n)} \tag{14}$$

2.8 Model Performance Measures

To assess the performance of a classification model, a confusion matrix is commonly employed. This matrix is a table that summarizes the performance of a specific model or algorithm. Each row in the matrix represents the actual classifications, while each column represents the predicted classifications, or vice versa [28]. Below is a general form of a confusion matrix for k labels or classes.

Table 2. Confusion Matrix for Sum $k = K$

Current (A)	Prediction (P)			
	Class 1	**Class 2**	**...**	**Class n**
Class 1	x_{11}	x_{12}	...	x_{1n}
Class 2	x_{21}	x_{22}	...	x_{2n}
⋮	⋮	⋮	⋮	⋮
Class n	x_{n1}	x_{n2}	...	x_{nn}

Source: [29]

1. True Positive (TP) represents the number of instances where the actual class is positive, and the model correctly predicts it as positive.
2. True Negative (TN) represents the number of instances where the actual class is negative, and the model accurately predicts it as negative.
3. False Positive (FP) represents the number of instances where the actual class is negative, but the model incorrectly predicts it as positive.
4. False Negative (FN) represents the number of instances where the actual class is positive, but the model mistakenly predicts it as negative.

Table 2 illustrates a confusion matrix for a multi-class classification problem with K classes [25]. The total number of FN, FP, and TN for each class is calculated using the following equation:

$$TFN_i = \sum_{j=1, j\neq i}^{n} x_{ij} \tag{15}$$

$$TFP_i = \sum_{j=1, j\neq i}^{n} x_{ji} \tag{16}$$

$$TTN_i = \sum_{j=1, j\neq i}^{n} \sum_{k=1, k\neq i}^{n} x_{ji} \tag{17}$$

$$TTP_i = \sum_{j=1}^{n} x_{jj} \tag{18}$$

where:

TFN_i : total false negatives of the *i*-th
TFP_i : total false positives of the *i*-th
TTN_i : total true negative of the *i*-th
TTP_i : total true positive of the *i*-th

Classification accuracy is assessed using two main criteria: accuracy and error rate. Each is defined as follows **[30]**:

1. Accuracy measures the proportion of correctly identified instances among all instances. It indicates the overall effectiveness of a diagnostic test or model. The formula for calculating accuracy is:

$$Accuracy = \frac{TP + TN}{Total} \tag{19}$$

Error Rate Indicates how frequently the model makes incorrect predictions. It is used to evaluate the performance of prediction and classification models. The formula for calculating the error rate is:

$$Error\ rate = \frac{TFN + TFP + TTN}{Total} \tag{20}$$

3. RESULTS AND DISCUSSION

3.1 Standardization Data

The initial step in performing cluster analysis is to standardize the data. Data standardization is essential because the variables in the dataset may have different units and scales, which can affect the clustering results. Standardization ensures that each variable contributes equally to the analysis by transforming the data to a common scale. By converting data to the same unit, we can compare information directly and accurately which is important in statistical analysis and data processing. Data standardization can be done by using the formula in **Equation (1)**

3.2 Hierarchical Cluster

After conducting the assumption test, hierarchical cluster analysis is performed. This analysis employs three linkage methods: complete linkage, single linkage, and average linkage. The calculations for single linkage, complete linkage, and average linkage are detailed in **Equation (3)**, **Equation (4)**, and **Equation (5)**, respectively. Euclidean distance is used for measurement, as specified in **Equation (2)**. The resulting dendrogram from the hierarchical cluster analysis is illustrated in the following figure.

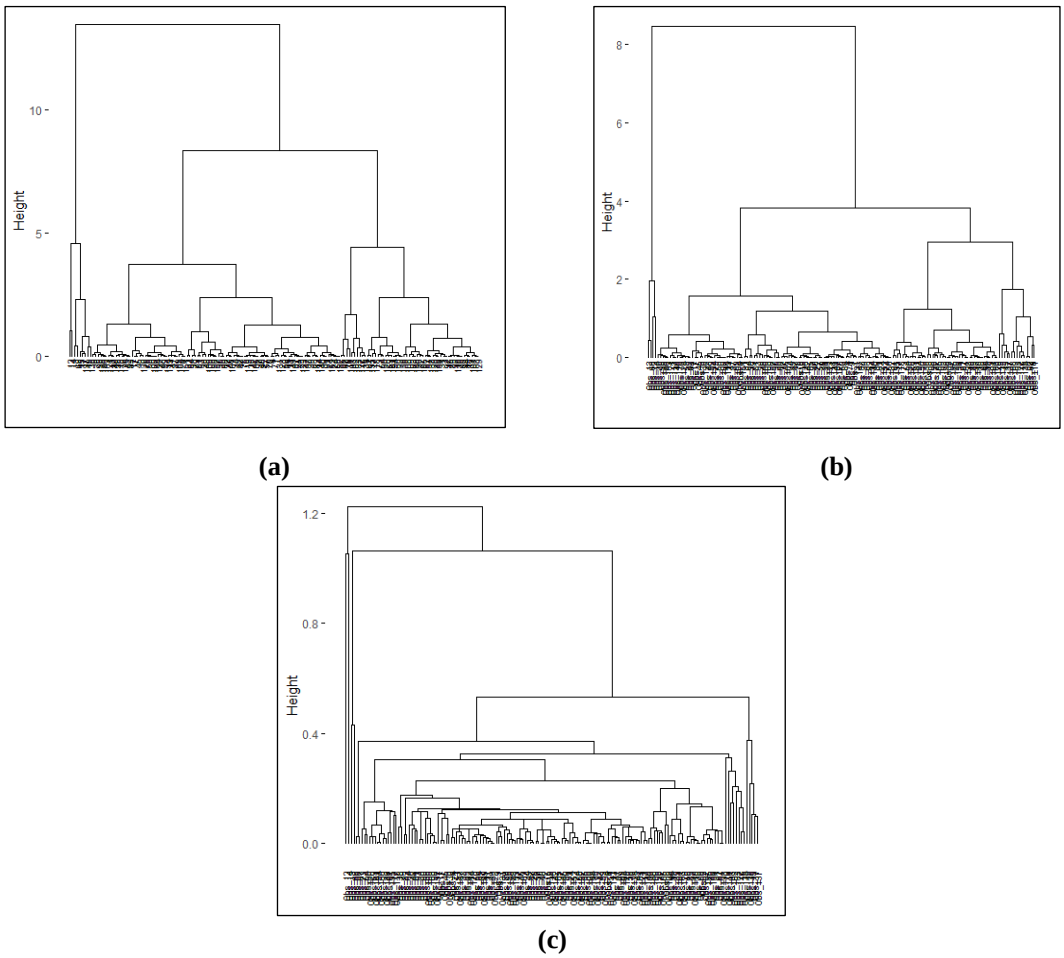

Figure 2. Dendogram of Hierarchical Cluster Results
(a) Complete Linkage, (b) Average Linkage, (c) Single Linkage

The dendrograms in **Figure 2** visualize the results of hierarchical clustering using three linkage methods there are complete linkage, average linkage, and single linkage. The vertical axis (height) represents the distance or dissimilarity between clusters. To determine the optimal number of clusters, a horizontal cut is made at a specific height based on the longest stem length in the dendrogram. This approach identifies the level at which the clusters are most distinct. For all three linkage methods shown, cutting the dendrogram at the appropriate height reveals three clusters, which are considered the optimal grouping. The selection process involves visually inspecting the dendrogram to find a balance between the number of clusters and the dissimilarity threshold, ensuring meaningful and interpretable results. The distribution of members within each cluster is detailed in **Table 3**.

Table 3. The Number of Members of Each Linkage on Hierarchical Cluster

Cluster	The Number of Cluster Members		
	Complete Linkage	Average Linkage	Single Linkage
1	71	95	154
2	29	58	1
3	57	4	2

Table 3 shows the distribution of members within each cluster for different hierarchical clustering methods. Using the Complete Linkage Method, Cluster 1 consists of 71 members, Cluster 2 has 29 members, and Cluster 3 has 57 members. The Average Linkage Method results in Cluster 1 containing 95 members, cluster 2 having 58 members, and Cluster 3 with just 4 members. The Single Linkage Method produces three clusters there are cluster 1 with 154 members, Cluster 2 with 1 member, and Cluster 3 with 2 members.

When comparing the average indicators for each cluster, it is evident that both the average linkage and single linkage methods produce the same number of members per cluster. Across all methods, there are complete linkage, average linkage, and single linkage. Cluster 1 consistently exhibits the highest average indicator values, identifying it as the "high" cluster. Cluster 2 demonstrates moderate average indicator values, classifying it as the "medium" cluster. Meanwhile, Cluster 3 consistently shows the lowest average

indicator values, making it the "low" cluster. The high cluster represents districts or cities with the largest cabbage harvest areas, highest production levels, and greatest productivity, while the medium and low clusters reflect progressively lower values for these indicators.

3.3 Cluster Self-Organizing Method (SOM)

Self-Organizing Map (SOM) Cluster Analysis is an effective method for grouping data because it utilizes neural networks and enables the visualization of complex structures within the data while preserving the original topological relationships. In this study, the optimal number of clusters is determined using an elbow diagram. **Figure 3** illustrates the elbow diagram used for this purpose.

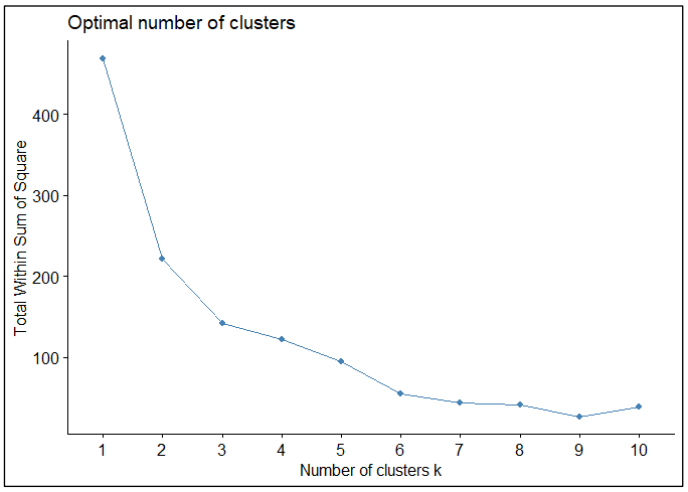

Figure 3. Elbow Diagram

In **Figure 3** above, it can be seen that in this figure, the graph forms a right angle. So, the optimal cluster obtained is 3 clusters. The next step is to initialize the grid parameters for SOM using the som grid function. This grid has 3 units (neurons) on the X-axis and 1 unit on the Y-axis, with a hexagonal topology. This means that the neurons will be arranged in one horizontal row, with each neuron potentially connected to its neighboring neurons in a hexagonal pattern. The results of the SOM cluster with 3 clusters produce a mapping plot as in **Figure 4** below.

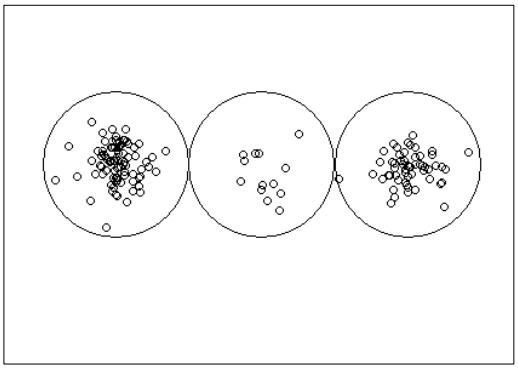

Figure 4. SOM Cluster Mapping Plot

Figure 4 shows that three circles each represent one neuron in a 3x1 SOM grid. The dots inside each circle represent data mapped to that neuron. Data that are closer in feature space will be mapped to the same or adjacent neurons. This plot shows that the data is divided into three main clusters, with two large clusters on the left and right sides and one small cluster in the middle.

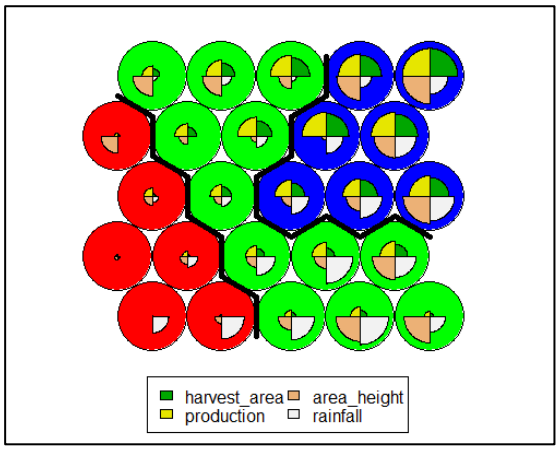

Figure 5. Codes Plot Cluster SOM

Figure 5 shows the visualization of a Self-Organizing Map (SOM) based on four variables: harvest area, production, area height, and rainfall. The node colors (red, green, blue) represent the dominant characteristics of each cluster, with blue nodes in the upper left indicating high harvest area and production, green nodes on the right reflecting the influence of altitude and rainfall, and red nodes in the lower left showing low values for production and harvest area. The pie charts within each node illustrate the relative contribution of each variable, while the proximity of nodes reflects the similarity of data between clusters.

Table 4. The Number of Members of Each Linkage on Cluster SOM

Cluster	The Number of Cluster Members
1	87
2	13
3	57

Table 4 reveals the results of the SOM cluster analysis, where Cluster 1 contains 87 members, Cluster 2 has 13 members, and Cluster 3 comprises 57 members. Upon comparing the average indicators for each cluster, it is evident that Cluster 1 exhibits the lowest average indicator value, categorizing it as a low cluster. Conversely, Cluster 2 has a medium average indicator, designating it as a high cluster. Cluster 3, with its higher average indicator value relative to the others, is identified as a medium cluster.

3.4 Cluster Ensemble

Ensemble clustering integrates multiple methods, such as complete linkage hierarchical analysis and SOM, to address differences in cluster memberships and enhance robustness. By leveraging the strengths of each method. Complete linkage Hierarchical analysis for well-separated clusters and SOM for preserving topological structures, the approach mitigates individual limitations. Using the CSPA algorithm via the diceR function in RStudio, a consensus matrix was created to harmonize cluster memberships, ensuring reliable groupings. The study identified two and three clusters as optimal configurations, with the cluster distributions summarized in **Table 5.**

Table 5. The Number of Members of Each Linkage on the Ensemble Cluster

Cluster	The Number of Cluster Members	
1	152	144
2	5	12
3	0	1

Table 5 displays the results of the ensemble cluster analysis, which produced two distinct cluster configurations: one with 2 clusters and one with 3 clusters. In the 2-cluster configuration, Cluster 1 includes 152 members, while Cluster 2 comprises 5 members. Upon comparing the average indicators for each cluster, Cluster 1, with the lowest average indicator value, is identified as a low cluster. Conversely, Cluster 2, having the highest average indicator, represents districts or cities with high cabbage production in Indonesia.

In the 3-cluster configuration, Cluster 1 contains 144 members, Cluster 2 has 12 members, and Cluster 3 includes 1 member. Analysis of the average indicators reveals that Cluster 1, with the lowest average

indicator value, is categorized as a low cluster. Cluster 2 has a medium average indicator, making it a high cluster. Cluster 3, with a higher average indicator value than Cluster 2, is designated as a medium cluster.

3.5 Naïve Bayes Classifier

After categorizing cabbage production levels through clustering methods, namely hierarchical clustering, Self-Organizing Maps (SOM) clustering, and ensemble clustering. The subsequent step involves the classification process using the Naïve Bayes algorithm. This process includes testing various proportions of training and test data to evaluate their impact on the performance metrics of the Naïve Bayes classifier. The evaluation focuses on measuring accuracy, specificity, and sensitivity by averaging these metrics across different test scenarios. The test was carried out 2 times using the holdout method, involving 157 data whose percentage ratios were determined, namely, 70:30 and 80:20, with all test data taken through random sampling techniques and involving data classes originating from the results of the clustering process. The results of the first classification process are shown in **Table 6.**

Table 6. Accuracy and Error Rate in the Integrated Cluster Model with Naïve Bayes Classifier

Methods	Number Of Clusters	Dataset Division	Accuracy	Error Rate
Hierarchical Analysis of Single Linkage	3	70:30	0.9583	0.0417
		80:20	0.9375	0.0625
Hierarchical Analysis of Average Linkage	3	70:30	0.9787	0.0213
		80:20	0.9375	0.0625
Hierarchical Analysis of Complete Linkage	3	70:30	0.9787	0.0213
		80:20	0.9741	0.0259
SOM	2	70:30	0.9148	0.0852
		80:20	0.9365	0.0635
	3	70:30	0.9683	0.0317
		80:20	0.9574	0.0426
Ensemble Cluster	2	70:30	0.9583	0.0417
		80:20	0.9355	0.0645
	3	70:30	0.9757	0.0243
		80:20	0.9841	0.0159

This classification shows varying performance depending on the clustering method used and the proportion of data split between training and testing data. In general, the ensemble cluster method with 3 clusters proves to be the most effective approach in improving classification performance, especially when used with a larger training data proportion. Therefore, the best method is achieved with the ensemble cluster method using a 90:10 training and testing data split, resulting in an accuracy of 0.9841 and an error rate of 0.0159.

Cluster 1 consists of areas with a high planting area (X_1) and very high cabbage production (X_2). The area height (X_3) is high, and rainfall (X_4) is relatively high, indicating that these areas benefit from climate and topographic conditions that support efficient cabbage farming practices. Cluster 2 includes areas with smaller planting areas (X_1) compared to Cluster 1, but with similar area height (X_3) and moderate rainfall (X_4). These areas achieve a good level of production (X_2), showing efficient farming practices despite limited land resources and stable climate conditions. Cluster 3 includes areas with very small planting areas (X_1) and low cabbage production (X_2). The area height (X_3) and rainfall (X_4) in this cluster are also low, which may contribute to the poor production performance. Other factors, such as soil quality, geographical constraints, and less favorable climatic conditions, may explain the low yields in these areas.

4. CONCLUSIONS

Based on the process and results of the integration of several cluster methods and Naïve Bayes classifier, it was concluded that cluster methods such as hierarchical analysis with three linkages (complete, average, single), Self-Organizing Maps, and ensemble clusters were used for grouping cabbage production in 127 districts/cities in Indonesia. The results of the hierarchical analysis produced 2 clusters for complete

linkage, 2 clusters for average linkage, and 3 clusters for single linkage. The results of cluster analysis with the SOM method obtained 3 clusters, while the ensemble cluster obtained 2 and 3 clusters. The clusters obtained in each data will be used as labels, which will then be used for Naïve Bayes classification to classify cabbage production levels in Indonesia. From several classification processes that have been carried out, the classification model with the number of cluster classes 3 with a ratio of training data and testing data division of 90:10 produced the best model performance with an accuracy of 0.9987 and an error rate of 0.002. Cluster 1 is the cluster with high cabbage production, characterized by high planting area, very high cabbage production, and high area height along with high rainfall, indicating favorable conditions for efficient cabbage farming. Cluster 2 has moderate cabbage production, with a moderate planting area, similar area height, and moderate rainfall, indicating efficient farming practices despite limited land resources. Cluster 3 has low cabbage production, with low planting area, area height, and rainfall, suggesting less favorable conditions for cabbage farming.

REFERENCES

[1] Vikram Gude, Saravanan V, Ishwarya RJ, and Sathya M, "IDENTIFICATION OF SIGNIFICANT FEATURES AND DATA MINING TECHNIQUES IN PREDICTING HEART STROKE," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 2, no. 1, pp. 676–682, 2022, doi: 10.48175/ijarsct-7738.

[2] J. E. van Engelen and H. H. Hoos, "A SURVEY ON SEMI-SUPERVISED LEARNING," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020, doi: 10.1007/s10994-019-05855-6.

[3] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, "K-MEANS CLUSTERING AND NAIVE BAYES CLASSIFICATION FOR INTRUSION DETECTION," *J. IT Asia*, vol. 4, no. 1, pp. 13–25, 2016, doi: 10.33736/jita.45.2014.

[4] N. M. A. A. Badung, A. A. R. Fernandes, and W. H. Nugroho, "COMPARISON OF DISTANCE AND LINKAGE IN INTEGRATED CLUSTER ANALYSIS WITH MULTIPLE DISCRIMINANT ANALYSIS ON HOME OWNERSHIP CREDIT BANK IN INDONESIA," *Math. Stat.*, vol. 9, no. 6, pp. 958–975, 2021, doi: 10.13189/ms.2021.090612.

[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] V. Kotu and B. Deshpande, DATA SCIENCE CONCEPT AND PRACTICE, 2nd ed., vol. 19, no. 5. Chenna: Morgan Kaufmann, 2018.

[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] D. Aktaş, B. Lokman, T. İnkaya, and G. Dejaegere, "CLUSTER ENSEMBLE SELECTION AND CONSENSUS CLUSTERING: A MULTI-OBJECTIVE OPTIMIZATION APPROACH," *Eur. J. Oper. Res.*, vol. 314, no. 3, pp. 1065–1077, 2024, doi: 10.1016/j.ejor.2023.10.029.

[9] A. Z. Machfud, A. Pandu Kusuma, and W. Dwi Puspitasari, "ANALISIS ALGORITMA NAIVE BAYES CLASSIFIER (NBC) PADA KLASIFIKASI TINGKAT MINAT BARANG DI TOKO VIOLET CELL," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 87–94, 2023, doi: 10.36040/jati.v7i1.5692.

[10] M. Arifat, Wardiana Adinda Putri, and A. S. Mufida, "PENERAPAN METODE NAIVE BAYES CLASSIFIER UNTUK KLASIFIKASI INDEKS PEMBANGUNAN MANUSIA DI PROVINSI JAWA TIMUR," *J. Stat. dan Komputasi*, vol. 2, no. 1, pp. 31–43, 2023, doi: 10.32665/statkom.v2i1.1661.

[11] R. R. R. Arisandi, B. Warsito, and A. R. Hakim, "APLIKASI NAÏVE BAYES CLASSIFIER (NBC) PADA KLASIFIKASI STATUS GIZI BALITA STUNTING DENGAN PENGUJIAN K-FOLD CROSS VALIDATION," *J. Gaussian*, vol. 11, no. 1, pp. 130–139, 2022, doi: 10.14710/j.gauss.v11i1.33991.

[12] L. Best, E. Foo, and H. Tian, "UTILISING K-MEANS CLUSTERING AND NAIVE BAYES FOR IOT ANOMALY DETECTION: A HYBRID APPROACH," *Smart Sensors, Meas. Instrum.*, vol. 43, pp. 177–214, 2022, doi: 10.1007/978-3-031-08270-2_7.

[13] A. N. Haloho, "RESPON PERTUMBUHAN DAN PRODUKSI KUBIS (BRASSICA OLEARACEAE. 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] U. Braga-Neto, FUNDAMENTALS OF PATTERN RECOGNITION AND MACHINE LEARNING, 2nd ed. Switzerland: Springer, 2020.

[15] Haviluddin *et al.*, "A PERFORMANCE COMPARISON OF EUCLIDEAN, MANHATTAN AND MINKOWSKI DISTANCES IN K-MEANS CLUSTERING," *2020 6th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Disaster Manag. ICSITech 2020*, pp. 184–188, 2020, doi: 10.1109/ICSITech49800.2020.9392053.

[16] T. Li, A. Rezaeipanah, and E. S. M. Tag El Din, "AN ENSEMBLE AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM BASED ON CLUSTERS CLUSTERING TECHNIQUE AND THE NOVEL SIMILARITY MEASUREMENT," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3828–3842, 2022, doi: 10.1016/j.jksuci.2022.04.010.

[17] V. Vijaya, S. Sharma, and N. Batra, "COMPARATIVE STUDY OF SINGLE LINKAGE, COMPLETE LINKAGE, AND WARD METHOD OF AGGLOMERATIVE CLUSTERING," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 568–573, 2019, doi: 10.1109/COMITCon.2019.8862232.

[18] N. Satyahadewi, S. J. Sinaga, and H. Perdana, "HIERARCHICAL CLUSTER ANALYSIS OF DISTRICTS/CITIES IN NORTH SUMATRA PROVINCE BASED ON HUMAN DEVELOPMENT INDEX INDICATORS USING PSEUDO-F," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 3, pp. 1429–1438, 2023, doi: 10.30598/barekengvol17iss3pp1429-1438.
[19] S. Patel, S. Sihmar, and A. Jatain, "A STUDY OF HIERARCHICAL CLUSTERING ALGORITHMS," *2015 Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2015*, pp. 537–541, 2015.
[20] 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.
[21] 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.
[22] 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.
[23] 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.
[24] H. Nashir, A. Kurnia, and A. Fitrianto, "SUBDISTRICT CLUSTERING IN WEST JAVA PROVINCE BASED ON DISEASE INCIDENCE OF JKN PARTICIPANTS PRIMARY SERVICES," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 1, pp. 0295–0304, 2023, doi: 10.30598/barekengvol17iss1pp0295-0304.
[25] 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.
[26] 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.
[27] R. Agustin, V. M. Santi, and B. Sumargo, "METODE NAIVE BAYES DALAM MENDETEKSI SEL KANKER PAYUDARA," *J. Stat. dan Apl.*, vol. 3, no. 1, pp. 30–38, 2019, doi: 10.21009/jsa.03104.
[28] I. W. Saputro and B. W. Sari, "UJI PERFORMA ALGORITMA NAÏVE BAYES UNTUK PREDIKSI MASA STUDI MAHASISWA," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 1, 2020, doi: 10.24076/citec.2019v6i1.178.
[29] A. Diallo, L. Affognon, C. Diallo, and E. C. Ezin, "DEEP LEARNING BASED BINARY AND MULTI-CLASS CLASSIFICATION COMPARISON FOR ANOMALY DETECTION," *8th Int. Conf. Eng. Emerg. Technol. ICEET 2022*, no. October, pp. 1–6, 2022, doi: 10.1109/ICEET56468.2022.10007171.
[30] T. Wang and Y. Zhao, "CREDIT CARD FRAUD DETECTION USING LOGISTIC REGRESSION," *Proc. - 2022 Int. Conf. Big Data, Inf. Comput. Network, BDICN 2022*, pp. 301–305, 2022, doi: 10.1109/BDICN55575.2022.00064.