Clustering Snack Products Based on Nutrition Facts Using SOM and K-Means for Diabetic Dietary Recommendation

Maritza Adelia[1], Arum Handini Primandari[2]*
Department of Statistics, Faculty of Mathematics and Natural Science[1], [2]
Universitas Islam Indonesia
Yogyakarta, Indonesia
maritza.adelia@students.uii.ac.id[1] primandari.arum@uii.ac.id [2

Abstract*— The number of diabetics in Indonesia continues to rise, with Type II Diabetes Mellitus (DM) dominating 90% of cases. One of the main contributors is the excessive consumption of snack products high in Sugar, Salt, and Fat (SSF), which increases health risks, particularly for diabetics. However, the current nutrition facts provided in the product package is not easy to understand. Creating label for the product can make an effective information to assist people on buying decision. This study aims to segment snack products based on their nutritional facts, particularly focusing on their SSF content, to identify products that are potentially high-risk for diabetics. In this study, data on the nutritional facts of snack products were analyzed. Utilizing a hexagonal Self-Organizing Map (SOM) topology with a 5 × 9 grid, the best clustering method identified was k-means. This method yielded two clusters, with a silhouette index of 0.44, a Dunn index of 0.09, and a connectivity index of 11.14. The first cluster comprises 165 products that have low levels of total fat, saturated fat, sugar, and salt. In contrast, the second cluster consists of 46 products with high total fat and saturated fat content, and this cluster is of particular concern due to its elevated levels of these unhealthy fats. The segmentation results can serve as a reference for more intuitive food labeling, potentially improving consumer awareness and aiding in dietary decision-making, particularly for diabetics.

Keywords*— Clustering SSF, nutrition facts, snack healthy label, SOM

I. INTRODUCTION

Diabetes Mellitus (DM) is a serious problem in Indonesia and around the world. DM is a chronic condition characterized by the body's inability to produce adequate insulin or effectively utilize the insulin it produces, resulting in elevated blood glucose levels. According to IDF (International Diabetes Federation), number of people with DM in Indonesia reached 19.47 million in 2021 and is expected to increase over time. DM consists of 4 types, namely type I, type II, gestational DM, and other DM, but as reported by IDF, 90% of diabetics suffer from type II DM [1].

Type II DM is influenced by unhealthy lifestyle factors or triggered by other conditions such as high blood pressure or obesity. In addition, the consumption of packaged foods and beverages high in Sugar, Salt, and Fat (SSF, Indonesian: *Gula, Garam, Lemak*-GGL) is a risk factor [2]. According to the Individual Food Consumption Survey, approximately 77 million Indonesians have consumed SSFs above the daily limit, 53.1% of whom are adolescents aged 13-18 years [3]. This indicates a significant health risk that needs immediate attention.

To overcome this problem, the Food and Drug Administration (BPOM RI), through BPOM Regulation No. 26 of 2021, has required the inclusion of the nutrition facts on the label of processed packaged products and urges the public to always read the nutrition facts table correctly and carefully [4]. However, the low level of public awareness regarding the nutrition facts table has driven BPOM to introduce the Nutri-Level program. This program focuses on labeling the risk level of SSF content by indicating high and low SSF levels in packaged products [5]. To support this initiative, clustering packaged snack products based on their SSF content can be a powerful strategy. Through cluster analysis, it is possible to systematically group products with similar SSF content, identify high-risk products, and develop consistent labeling schemes. This not only helps policymakers implement targeted regulations but also increases public awareness of health risks associated with excessive SSF consumption. The Self-Organizing Map (SOM) method is one of the clustering techniques that enables the grouping of products based on the similarity of their SSF profiles.

First introduced by Professor Teuvo Kohonen in 1982, SOM is a technique for visualizing and clustering data according to its characteristics [6]. This method is able to cluster high dimensional data and is resistant to noise and outliers [7]. In the SOM method, there are output neurons that can be regrouped to simplify the clustering results and make them easier to understand. This method was chosen because it can cluster high-dimensional data and is resistant to noise and outliers [8]. The methods used in this research are hierarchical agglomerative methods in the form of complete linkage and average linkage and the k-means method because these three methods can be used in clustering output neurons in the SOM topology.

Previous research has explored the application of the SOM method for clustering various datasets. For instance, Hardika K.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

(2018) for clustering social and population data from 33 provinces in Indonesia, which are indicators of remote and disadvantaged areas, and 2 clusters were formed with SOM and k-means methods as advanced clustering methods. There is also research on grouping 38 packaged products based on nutrition facts by Husna et al. (2019) using k-means method and 2 clusters were formed. However, there has been no research specifically focusing on segmenting snack products using SOM with SSF content as the primary basis for clustering. This research aims to fill that gap by identifying high-risk snack product groups, which can further support effective public health strategies through product labeling and increased consumer awareness.

II. METHODOLOGY

A. Material and Data

The population in this study were food products included in the category 15.0 of snack products based on BPOM Regulation Number 13 of 2023. The category includes all types of savory or other flavored snacks: 15.1 snacks – potato, tubers, cereals, flour or starch (from tubers and nuts); 15.2 nut preparations, including coated nuts and nut mixtures (examples with dried fruit); and 15.3 fish-based snacks. A total of 211 samples were taken using purposive sampling technique because it was based on the retrieval criteria: products included in the category of snack products having nutrition facts tables in their packaging and are sold at Manna Kampus Godean, Toko Agung Grosir Yogyakarta, and KN Putra Toserba Magelang. Data were collected by photographing the nutrition facts table of products that fall under the snacks category using a mobile phone camera as shown in Fig 1. The variables used in this study were total fat, saturated fat, sugar, and salt. The following in Table 1 are operational definition of research variables.

TABLE I. OPERATIONAL DEFINITION OF RESEARCH VARIABLE

No.	Variable	Definition	Unit	Scale
1	Total Fat	All fatty acids in food and expressed as triglycerides	Gram	Ratio
2	Saturated Fat	All fatty acids without double bonds	Gram	Ratio
3	Sugar	The sum of all monosaccharides and disaccharides found in processed foods	Gram	Ratio
4	Salt	Amount of salt (sodium) listed as total sodium	Gram	Ratio

Fig. 1. Example of Sample Data Collection

B. Research Method

This research consists of several stages. The following in Fig 2 is a research flowchart.

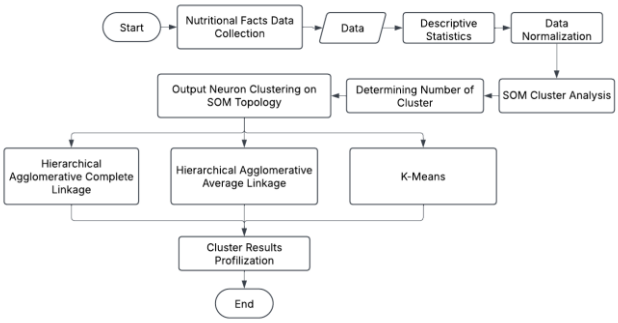

Fig. 2. Research Flowchart

1) Descriptive Statistics

Descriptive statistics is a stage of data analysis in which the data is described without the aim of making general conclusions or generalizations [9]. Descriptive statistical techniques used in this study are maximum, minimum, mean, and visualization in the form of boxplots.

2) Data Normalization

Data normalization is used to rescale the data so that the analysis results are more representative. This is because each variable in a data set often has a range of values that are very different. The method used is min-max normalization, where the data scale is changed to a range of 0 to 1. The following is the formula for min-max normalization [10].

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x' is the normalized data, x_i is the actual data, x_{min} is the minimum value of data per variable, and x_{max} is the maximum value of data per variable.

3) Self-Organizing Map (SOM)

SOM represent a form of unsupervised Artificial Neural Network that functions to both reduce dimensionality and group similar data points into clusters according to their characteristics. The architecture of SOM consists of an input layer with input neurons and an output layer with output neurons. SOM itself uses competitive learning in its algorithm, which means that the output neurons compete to determine the closest distance to the input neuron until a winning neuron is obtained [11]. To determine the size of the grid or the number of output neurons to form, the researcher uses the Kohonen formula, which states that the maximum number of output neurons should be $5 \times \sqrt{N}$, where N is the number of observations in the data [12]. The number of iterations and the type of topology used must also be determined. The optimal number of iterations is reached when the map has reached a stable state, or as can be seen from the average distance to the nearest unit value that is stable at each iteration [13]. SOM also has two types of topology such as hexagonal, where each neurons has at most 6 neighbors and rectangular, where each neurons has at most 4 neighbors. but hexagonal topology is preferred because it allows better visualization of the overall

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

data structure [14].

Then, after the input vectors are successfully clustered in each output neuron, a partitioning of the output neurons formed by the vector weights on each output neuron is performed using another clustering method, such as the k-means method. This is done to make the boundaries between clusters clearer and to simplify the clustering results [8]. The steps for performing SOM clustering are as follows [15].

1. Initialize the weight vector between the input neuron and the output neuron with a random number from 0 to 1.
2. Calculate the distance between the input vector and the weight vector for each output neuron using Euclidean distance, and take the output neuron with the smallest distance value as the winning neuron. The formula for the Euclidean distance is as follows
$$d_j = \sum_{i=1}^{p}(w_{ij} - x_{ki}), k = 1, 2, \dots, n \quad (2)$$
where w_{ij} is the weight vector with $i = 1,2,\dots,p$ and $j = 1,2,\dots,m$ where p is the number of variables and m is the number of output neurons and x_{ki} is the value of the k-th input vector in the i-th variable.
3. Use the following formula to update the weight vector of the winning neurons.
$$w_{ij}(new) = w_{ij}(old) + \alpha(x_{ki} - w_{ij})(old) \quad (3)$$
where α is the learning rate, which has a value of $0 \leq \alpha \leq 1$ and will decrease with the number of iterations performed.
4. For each input vector x, repeat steps two through three.
5. Update the learning rate (α) at the t-th iteration with $t = 1,2,\dots,T$ with the following equation.
$$\alpha(1+t) = \alpha(t)\left(1 - \frac{t}{T}\right) \quad (4)$$
where α is the learning rate, which has a value of $0 \leq \alpha \leq 1$ and will decrease with the number of iterations performed.
6. Until the maximum iteration is reached and the learning rate converges to zero, repeat steps four through five.
7. Group each observation object or input vector into the output neuron with the closest distance or the one with the smallest distance value.

4) Hierarchical Clustering

Hierarchical clustering is a clustering method that uses a hierarchical structure or level in the process. This method is divided into two types, namely divisive and agglomerative. Divisive means that objects are placed in one cluster and then divided into several clusters, while agglomerative means that adjacent objects are combined into separate clusters and then adjacent clusters are combined until all objects are included in one cluster [16]. In this study, the agglomerative method will be used in the form of complete linkage and average linkage.

a) Complete Linkage

Grouping in the complete linkage method is based on the greatest distance between objects in different clusters. The steps for the calculation are as follows [17].

1. Using the Euclidean distance size equation, calculate the distance matrix D between objects using the equation.
$$d_{i,j} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \text{ with } k = 1, 2, \dots, n \quad (5)$$
where d_{ij} is the Euclidean distance between the i-th object and the j-th object, x_{ik} and x_{jk} are the values of the i-th and j-th objects in the k-th variable, and p is the number of variables observed.
2. Select the closest distance in the distance matrix $D = \{d_{ij}\}$, then combine the two closest objects, for example objects U and V form a cluster (UV).
3. Update the distance matrix D by calculating the distance between clusters (UV) and other objects using the following formula.
$$d_{(UV)W} = \max(d_{UW}, d_{VW}) \quad (6)$$
where, $d_{(UV)W}$ is the distance between cluster (UV) and object W, d_{UW} is the distance between object U and W, and d_{VW} is the distance between object V and W.
4. Repeat the third step until all objects are placed in one cluster.

b) Average Linkage

Grouping in the average linkage method is based on the average between objects in different clusters. The steps for the calculation are as follows [17].

1. Using the Euclidean distance size equation (5), calculate the distance matrix D between objects using the equation.
2. Select the closest distance in the distance matrix $D = \{d_{ij}\}$, then combine the two closest objects, for example objects U and V form a cluster (UV).
3. Update the distance matrix D by calculating the distance between clusters (UV) and other objects using the following formula.
$$d_{(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{n_{(UV)} n_W} \quad (7)$$
where, $d_{(UV)W}$ is the distance between cluster (UV) and object W, $d_{(UW)}$ is the distance between objects U and W, $d_{(VW)}$ is the distance between objects V and W, $n_{(UV)}$ is the number of members in cluster (UV), and n_W is the number of members in cluster W.
4. Repeat the third step until all objects are placed in one cluster.

5) K-Means

The k-means method is a non-hierarchical method that aims to group objects into clusters based on their characteristics. The following are the calculation steps of the k-means method [18].

1. Determine the number of clusters or the value of k and randomly initialize the center of the cluster (centroid) as many as k.
2. Calculate the distance of each object to the centroid using the Euclidean distance equation (5) until the closest distance of each object to the centroid is found.
3. Assign the object to the cluster with the closest centroid.
4. Perform iterations from step 3. The new centroid value is calculated using the following equation.
$$C_k = {}^{1}/_{n_k} \sum d_i \quad (8)$$
Where n_k is the number of data in cluster k and d_i is the

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

sum of the distance values contained in each cluster.
5. Iterate until the centroid value and members of each cluster do not change. If the condition is not met, repeat from step two.

6) Cluster Validation

Cluster validation is a step to quantitatively and objectively evaluate the results of cluster analysis [19]. This research uses the silhouette index, Dunn index, and connectivity index methods.

Silhouette index used to evaluate the quality and strength of clusters, or how accurately an object is placed in a cluster [20]. The silhouette index has a value between 1 and -1, the closer the value is to 1, the more correct is the clustering structure produced, the closer the value is to -1, the more overlapping is the clustering structure produced. The following is the calculation formula [21].

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (9)$$

where $s(i)$ is the silhouette value for the i-th data, $a(i)$ is the average distance of the i-th object to all objects in the same cluster, and $b(i)$ is the minimum value of the average distance between the i-th object and objects in other clusters.

Dunn index is the ratio of the smallest distance between observations in different clusters to the largest distance between observations in the same cluster. Dunn index has a value of $0 \leq D \leq \infty$, and the higher the dunn index value, the better the resulting cluster. The following is the calculation formula [21].

$$D = \min_{j=i+1,\ldots,n_c} \left(\frac{d(c_i,c_j)}{\max_{k=1,\ldots,n_c}(diam(c_k))} \right) \quad (10)$$

where, D is the Dunn index value, $d(c_i, c_j)$ is the distance between clusters c_i and c_j, and $\max_{k=1,\ldots,n_c}(diam(c_k))$ is the maximum distance between objects in one cluster with n_c being the total number of clusters.

Connectivity index evaluates the homogeneity of the cluster. The connectivity index has a value of $0 \leq C \leq \infty$, and the smaller the value, the better the resulting cluster. The following is the calculation formula [21].

$$Conn = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}} \quad (11)$$

where, $nn_{i(j)}$ is the nearest neighbor of the i-th object to the j-th object, N is the number of objects, and L is the number of clusters. $x_{i,nn_{i(j)}}$ is 0 if i and $nn_{i(j)}$ are in the same cluster and $1 / j$ if they are in different clusters.

7) Independent Samples t-Test

An independent samples t-test is used to determine whether there are significant differences in the means of a variable when comparing two unrelated groups. The null hypothesis for this test is that there is no significant difference between the means of the two groups. The test statistics for the independent samples t-test are as follows [22].

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (12)$$

Where $\bar{x}_1$ is the group 1 mean, $\bar{x}_2$ is the group 2 mean, s_1^2 is the group 1 variance, s_2^2 is the group 2 variance, n_1 is the number of group 1 observations, and n_2 is the number of group 2 observations.

III. Result and Analysis

A. Descriptive Statistics

Descriptive statistics include mean, minimum, maximum, and boxplot visualization. The following are descriptive statistics of the nutrition facts data for snack products.

TABLE II. DESCRIPTIVE STATISTICS

Variable	Measure		
	Mean	Minimum	Maximum
Total Fat (g)	5.811	1	21
Saturated Fat (g)	2.3	0	6
Sugar (g)	1.818	0	15
Salt (g)	0.133	0.005	1.095

According to Table II, the highest total fat is 21 g contained in Aceh Fish Skin Salted Egg, the highest saturated fat is 6 g within Chitato Lite Onion Cream Sauce and Japota Spicy Lime, the the highest sugar content is 15 g contained in Sunbay Snack Spicy Crispy Squid, and the the highest salt content is 1,095 g within Maxicorn Roasted Corn. Moreover, based on the average value of each variable, none of them exceeded the daily intake limit according to the Ministry of Health. The maximum recommended daily intake for SSL is 67 g of fat, 50 g of sugar, and 5 g of salt. However, the total fat and sugar content has a maximum value of 21 g and 15 g, indicating a fairly high number for snack products when compared to the daily consumption limit.

The following are the results of boxplot visualization of nutrition facts data for snack products.

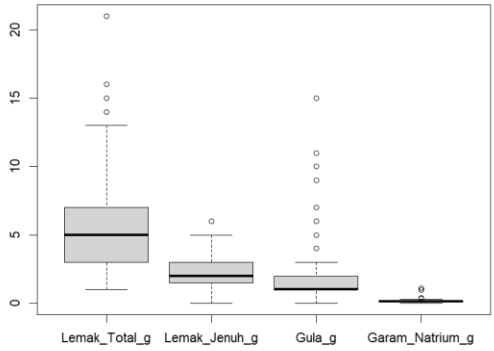

Fig. 3. Boxplot Visualization

According to the boxplot visualization in Fig 3, there are a total of 45 outliers. Since these outliers contain important information, the cluster analysis is performed using the SOM method which is insensitive to outliers.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

B. Data Normalization

Data normalization is essential because each variable in the dataset has a significant range. Standardizing the scale of the data is necessary to facilitate cluster analysis. The following presents the nutritional information of snack products after applying min-max normalization. The results, illustrated in Table III, demonstrate that the normalized values are uniformly distributed within a range of 0 to 1. Consequently, this data is suitable for further cluster analysis using SOM.

TABLE III. DATA AFTER NORMALIZED

No.	Product Name	Total Fat (g)	Saturated Fat (g)	Sugar (g)	Salt (g)
1	Nissin Sagu Keju	0.286	0.167	0.286	0.004
2	Happy Tos Corn Chips Merah	0.286	0.119	0.048	0.005
...	...	...	...	...	...
210	Aceh Fish Skin Salted Egg Spicy	1	0	0.048	0.018
211	Aceh Fish Skin Salted Egg	1	0	0.048	0.018

Based on Table III, the data is successfully normalized with a range between 0 and 1, so the data is ready for clustering analysis using SOM.

C. Cluster Analysis with SOM

In this analysis, a hexagonal topology was employed, and a total of 1500 iterations were conducted. To determine the optimal grid size and the number of output neurons, experiments were performed utilizing the silhouette index as a validation method. Below are the results of these experiments.

TABLE IV. RESULTS OF SOM TOPOLOGY GRID SIZE EXPERIMENT

No.	*Grid*	*Silhouette Index*	Number of Empty Output Neurons
1.	1 × 7	0.37	0
2.	1 × 8	0.36	0
...	...	...	...
27.	5 × 9	0.51	0
...	...	...	...
36.	8 × 9	0.48	11
37.	9 × 9	0.55	17

Based on the experimental results presented in Table IV, the optimal grid size for SOM analysis consists of a grid with up to 45 output neurons, yielding a silhouette index value of 0.51. This value indicates that the grid size is adequate for analysis. Although there are other grid sizes with higher silhouette index values, the chosen grid size is preferred because it avoids empty output neurons, which can lead to overfitting. Therefore, a grid of this size will be used in the SOM analysis. Below is a graph illustrating the training progress of the resulting SOM model.

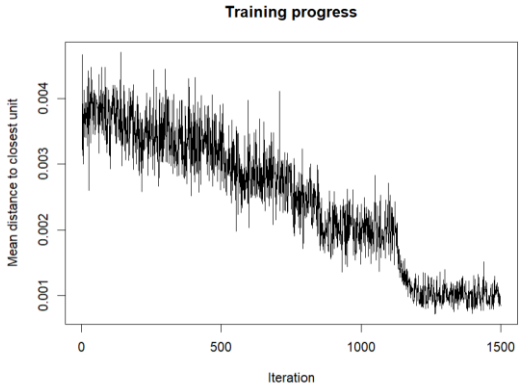

Fig. 4. SOM Model Training Progress Graph

The graph in Fig 4 shows that the average distance to the nearest output unit or neuron decreases as the number of iterations increases. After about 1100 iterations, the average distance to the nearest unit is less than 0.001 and remains stable or converges until the 1500th iteration. This indicates that the resulting cluster is good, because the smaller the average distance value to the nearest unit, the better the resulting cluster. After the iteration process, the resulting SOM topology with 45 output neurons is represented by a fan diagram as follows.

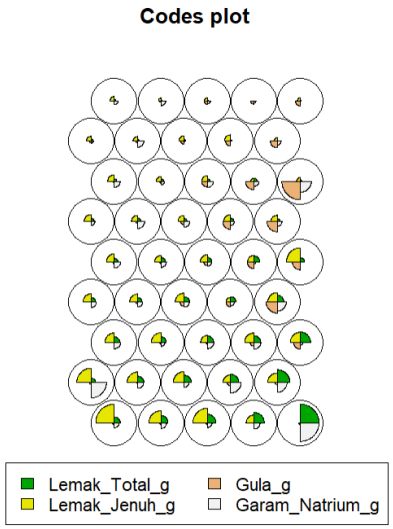

Fig. 5. Fan Diagram of SOM Analysis Results

The fan diagram in Fig 5 shows the distribution of each variable of the snack products in each output neuron. The larger the fan shape of a variable, the greater the content of that variable in the snack products included in the members of an output neuron. Below in Table V is a list of the snack product members in each output neuron.

TABLE V. LIST OF SNACK PRODUCTS IN SOM OUTPUT NEURONS

Output Neuron	Product Name
V1	Chitato Lite Saus Krim Bawang dan Japota Spicy Lime.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

Output Neuron	Product Name
...	...
V5	Aceh Fish Skin Salted Egg Spicy dan Aceh Fish Skin Salted Egg.
...	...
V35	Oishi Caramel Popcorn, Oishi Chocolate Popcorn, Krizz Chocolate, etc.
...	...
V45	Oishi Pillows Ubi, Oishi Pillows Keju, Oishi Pillows Durian, etc.

According to Fig 5, there is an output neuron associated with products that have a high total fat and salt content. This neuron is identified as neuron V5, which has an average total fat content of 21 g and a salt content of 0.38 g. Additionally, there is another output neuron that indicates high sugar content; this neuron is designated as neuron V35, which has an average sugar content of 11 g.

To clarify the interpretation of the formed clusters, further analysis is conducted through clustering the output neurons. This clustering process is based on the vector weights of the 45 output neurons. Table VI presents the vector weights for each output neuron as established in the SOM topology.

TABLE VI. SOM OUTPUT NEURON VECTOR WEIGHT

Output Neuron	Total Fat (g)	Saturated Fat (g)	Sugar (g)	Salt (g)
V1	0.359629	0.285714	0.047619	0.00607
V2	0.425696	0.184134	0.044744	0.006705
...	...	...	...	...
V45	0.094989	0.047619	0.142857	0.001766
V41	0.144807	0.064383	4.41E-09	0.004938

D. Output Neuron Clustering

a) Complete Linkage

Before clustering output neurons using the complete linkage method, it is crucial to determine the optimal number of clusters. The results of cluster validation are shown below.

TABLE VII. CLUSTER VALIDATION OF COMPLETE LINKAGE METHOD

Method	Number of Cluster			
	2	3	4	5
Silhouette Index	0.53	0.43	0.38	0.30
Dunn Index	0.22	0.18	0.27	0.28
Connectivity Index	7.18	13.33	15.29	15.79

According to Table VII, the optimal number of clusters is determined to be 2, as it has the highest silhouette index and the lowest connectivity index. However, the Dunn Index indicates that the best results are achieved with 5 clusters. Despite this, we have decided to proceed with the clustering of 2 groups. The values of the all validation method indicate a good cluster structure, where objects in the same cluster have a high degree of similarity. Below are the results of clustering the output neurons of the SOM model, along with a visualization using a fan diagram.

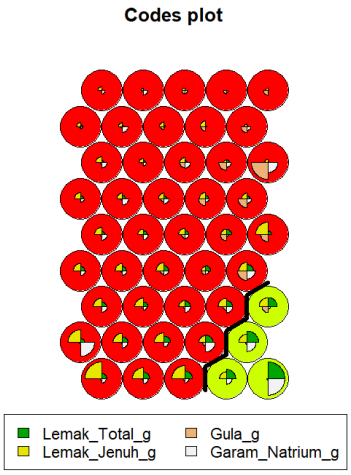

Fig. 6. Fan Diagram of Complete Linkage Method Results

Fig 6 shows that the output neurons are successfully grouped into 2 clusters, where the first cluster in red circles consists of 196 products with high sugar content and cluster 2 in yellow circles consists of 15 products with high total fat, saturated fat, and salt content. The clustering results show an extreme imbalance in the number of members in both clusters.

b) Average Linkage

The number of clusters associated with its measurement for average linkage is presented the following table.

TABLE VIII. CLUSTER VALIDATION OF AVERAGE LINKAGE METHOD

Method	Number of Cluster			
	2	3	4	5
Silhouette Index	0.65	0.46	0.37	0.31
Dunn Index	0.43	0.56	0.19	0.23
Connectivity Index	3.05	5.98	16.83	17.97

Based on Table VIII, the optimal number of clusters is 3 clusters because it has the largest silhouette index and Dunn index values and the smallest connectivity index value. The values of all validation methods show a good cluster structure, where objects in the same cluster have a high degree of similarity and are better than the previous method.

Then, the following is the results of clustering the output neurons of the SOM model with visualization using a fan diagram.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

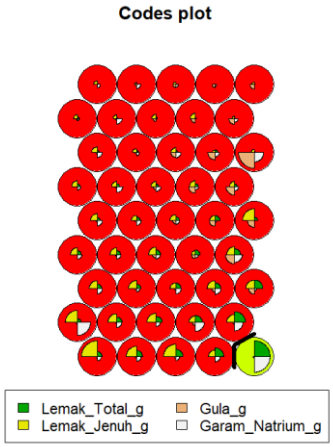

Fig. 7. Fan Diagram of Average Linkage Method Results

Fig 7 shows that the output neurons are successfully grouped into 2 clusters, where the first cluster shown in red circles consists of 209 products with high saturated fat and sugar content, and cluster 2 shown in yellow circles consists of 2 products with high total fat and salt content. This cluster result shows an extreme imbalance in the number of members in both clusters compared to the previous method.

c) K-Means

The following table represent the associate of the number of clusters and its evaluation measurements.

TABLE IX. CLUSTER VALIDATION OF K-MEANS METHOD

Method	Number of Cluster			
	2	3	4	5
Silhouette Index	0.44	0.43	0.37	0.30
Dunn Index	0.09	0.18	0.19	0.15
Connectivity Index	11.14	13.33	16.83	23.64

According to Table IX, the optimal number of clusters is determined to be 2, as this configuration yields the highest silhouette index and the lowest connectivity index. On the other hand, the Dunn index indicates that the optimal results are achieved with 4 clusters. Despite this, 2 clusters are prioritized since both evaluation methods indicate they produce the best outcomes. The values of all validation method indicate a weak cluster structure, but are still acceptable because this method is still able to identify data groups with certain similarity patterns. Below are the results of clustering the output neurons from the SOM model, along with a visualization using a fan diagram.

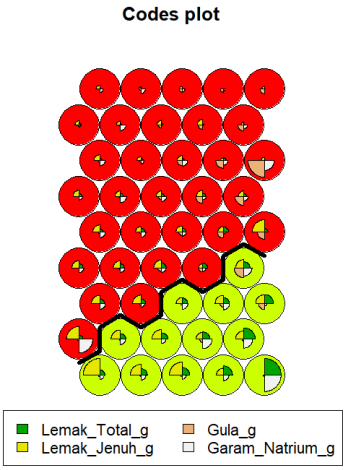

Fig. 8. Fan Diagram of K-Means Method Results

Fig 8 shows that the output neurons are successfully grouped into 2 clusters, where the first cluster in red circles consists of 165 products with high sugar content and cluster 2 in yellow circles consists of 46 products with high total fat, saturated fat, and salt content. This result shows a better distribution of cluster members compared to the two previous methods.

E. Results and Cluster Profilization

According to the output neuron clustering, the best method is k-means which forms 2 clusters. The following are the results of clustering the products into 2 clusters.

TABLE X. SNACK PRODUCT CLUSTERING RESULTS

Cluster	Number of Products	Products
1	165	Happy Tos Corn Chips Merah, TosTos Tortilla Chips Roasted Corn, Maxicorn Roasted Corn, Chitato Ayam Bumbu, Oishi Sponge Chocolate, etc.
2	46	Chitato Lite Saus Krim Bawang, Japota Sapi Panggang, Potabee Wagyu Beef Steak, Garuda Kacang Kulit Rasa Bawang, Krizz Cheese, etc.

Based on the product clustering results in Table X, the first cluster results in 165 products and the second cluster results in 46 products. For a complete list of products and to search for specific snack products by cluster, use the link https://bit.ly/CariMakananRingan. Then, to evaluate whether there are significant differences between the two clusters for each variable, an independent samples t-test is performed. Below are the results of the independent samples t-test.

I. Hypothesis
$H_0: \mu_1 = \mu_2$ (There is no significant difference in the average nutrient content between clusters 1 and 2)
$H_1: \mu_1 \neq \mu_2$ (There is significant difference in the average nutrient content between clusters 1 and 2)
II. Significance Level
$\alpha = 0.05$
III. Critical Area

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

Reject H_0 if $t_{hitung} < -t_{\alpha/2}$ or $t_{hitung} > t_{\alpha/2}$ or $p - value \leq \alpha$

IV. Test Statistics

$$t_{hitung} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

V. Decision and Conclusion

TABLE XI. INDEPENDENT SAMPLES T-TEST RESULT

Variable	$P - value$	Decision
Total Fat	$< 2.2 \times 10^{-16}$	Reject H_0
Saturated Fat	1.544×10^{-14}	Reject H_0
Sugar	0.2155	Fail to Reject H_0
Salt	0.09897	Fail to Reject H_0

Based on Table XI and using the 95% confidence level, it can be concluded that there is a significant difference in the average total fat and saturated fat content in clusters 1 and 2, while the average sugar and salt content is not significantly different.

Next, profiling is performed to determine the characteristics of each cluster as follows.

TABLE XII. CLUSTER RESULTS PROFILIZATION

Cluster	Average Total Fat (g)	Average Saturated Fat (g)	Average Sugar (g)	Average Salt (g)
1	4.45	1.99	1.91	0.13
2	10.69	3.4	1.49	0.16

According to the profiling results presented in Table XII, the first cluster includes products that contain low levels of total fat, saturated fat, sugar, and salt. In contrast, the second cluster comprises products that are high in total and saturated fat but low in sugar and salt. Snack products in the second cluster, which should be monitored by the general public and diabetics, have a high total fat content of 10.69 grams and a saturated fat content of 3.4 grams. It is important to note that the number of servings for each product may vary.

For example, in cluster 2 there is a Chitato Lite Onion Cream Sauce product with a net weight of 68 g and consisting of 3 portions. Each portion contains 7 g of total fat and 6 g of saturated fat. If the entire package is consumed in one meal, the total fat consumption entering the body will be 21 g and the saturated fat consumption will be 18 g. For comparison, according to the WHO (World Health Organization), the daily limit of total fat consumption is 67 g and the saturated fat consumption is 20-30 g. Thus, by consuming one package of Chitato Lite Onion Cream Sauce in one meal, approximately 31% of the daily limit for total fat consumption and 60% of the daily limit for saturated fat consumption will be met.

The results of this study are also in line with research conducted by Husna (2019), which found that the second cluster consisted of high-fat food products. Given the high levels of fat content found in cluster 2 snack products, it becomes crucial to consider preventive measures for public awareness. One effective strategy is food labelling, which has been successfully implemented in several countries. For instance, Singapore applies the Nutri-Grade system for beverage products, categorizing them from A to D according to sugar and saturated fat content. Meanwhile, Chile adopts Black Warning Labels for food and beverage products, indicating high levels of sugar, calories, saturated fat, or sodium. Inspired by these implementations, snack products in cluster 2 can benefit from a clear and prominent labelling system to highlight their high total fat and saturated fat content. This would allow consumers to make more informed dietary choices, reducing the intake of high-risk products and supporting diabetic management.

IV. CONCLUSION

Based on the research, snack products were segmented according to their nutritional facts using the hexagonal topology SOM method, which employed a grid and 1500 iterations. The best output neuron clustering was achieved using the k-means method, resulting in a silhouette index value of 0.44, a Dunn index of 0.09, and a connectivity index of 11.14. This analysis formed two significantly different clusters based on total fat and saturated fat content. The first cluster comprises 165 products with low saturated fat (SSF) content, while the second cluster includes 46 products with high levels of total and saturated fat. The second cluster consists of products that should be avoided by diabetics, as their consumption may exacerbate diabetes. Additionally, it is advisable for the general public to limit their intake of these products.

Given the differences in SSF content between the two clusters, proper labelling becomes crucial to raise consumer awareness and promote healthier dietary choices. The results of this study are expected to serve as a reference for designing Nutri-Level programs, taking into account the characteristics of SSF content in clusters 1 and 2. The labelling can be color-coded for better visibility, for example green indicating low-SSF snack products such as those in cluster 1, and red representing high-SSF snack products like those in cluster 2. Moreover, the labels may include additional indicators for products high in total fat, saturated fat, sugar, or salt. This approach is intended to help the public easily distinguish products that should be limited for consumption, thereby reducing the intake of SSFs beyond daily recommendations and lowering the risk of diabetes. Furthermore, this research could be further developed into a classification analysis of snack products based on segmentation results, enhancing its impact on public health awareness.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th edn., els, 2021.

[2] A. A. I. P. Wahyuni, "Konsumsi Gula, Garam, Lemak (GGL) Berlebihan ıh "MAUT"," 12 September 2024. [Online]. Available: //yankes.kemkes.go.id/view_artikel/3633/reqwest/index.

[3] E. Masri, N. S. Nasution and R. Ahriyasna, "Literasi Gizi dan Konsumsi Garam, Lemak pada Remaja di Kota Padang," *Jurnal Kesehatan*, vol.). 1, pp. 23-30, 2022.

[4] BPOM, "Gerakan Membaca Label Pangan," 8 Maret 2016. [Online]. able: https://www.pom.go.id/berita/gerakan-membaca-label-pangan.

[5] BPOM, "BPOM Dukung Penuh Pencantuman Nutri-Level pada Pangan ın Secara Bertahap," 23 September 2024. [Online]. Available: //www.pom.go.id/berita/bpom-dukung-penuh-pencantuman-nutri-level-pangan-olahan-secara-bertahap.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025

[6] I. Dermawan, A. Salma, Y. Kurniawati and T. O. Mukhti, ementation of the Self Organizing Maps (SOM) Method for Grouping ıces in Indonesia based on the Earthquake Disaster Impact," *UNP al of Statistics and Data Science*, vol. 1, no. 4, pp. 337-343, 2023.

[7] N. Y. Kusrahman, I. Purnamasari and F. D. T. Amijaya, "Optimasi Self-ıizing Map Menggunakan Particle Swarm Optimization untuk elompokkan Desa/Kelurahan Tertinggal di Kabupaten Kutai negara Provinsi Kalimantan Timur," *Jurnal Eksponensial*, vol. 11, no. 2, 39-144, 2020.

[8] B. Pangestu, D. Purwitasari and C. Fatichah, "Visualisasi Similaritas : Penelitian dengan Pendekatan Kartografi Menggunakan Self-ıizing Maps (SOM)," *Jurnal Teknik ITS*, vol. 6, no. 2, pp. 2337-3520,

[9] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif, dan R&D, ıng: ALFABETA, 2019.

[10] P. P. Allorerung, A. Erna, M. Bagussahrir and S. Alam, "Analisis ma Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada et Penyakit," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, 78-191, 2024.

[11] S. J. A. Sumarauw, Data Mining Model Self-Organizing Maps (SOMs), akarta: Bintang Semesta Media, 2022.

[12] I. Rojas, G. Joya and A. Catala, Advances in Computational Intelligence: nternational Work-Conference on Artificial Neural Networks, IWANN Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part II, Spain: ger, 2015.

[13] D. Miljković, "Brief review of self-organizing maps," *40th International ention on Information and Communication Technology, Electronics and electronics (MIPRO)*, pp. 1061-1066, 2017.

[14] I. G. Santos, V. Q. Carneiro, A. C. S. Junior, C. D. Cruz and P. C. Soares, organizing maps in the study of genetic diversity among irrigated rice ypes," *Acta Scientiarum. Agronomy*, vol. 41, pp. 1-9, 2019.

[15] S. Kania, D. Rachmatin and J. A. Dahlan, "Program Aplikasi elompokkan Objek dengan Metode Self Organizing Map Menggunakan sa R," *Jurnal Eurekamatika*, vol. 7, no. 2, pp. 17-29, 2019.

[16] V. Nellie, V. C. Marwadi and N. J. Perdana, "Implementasi Metode merative Hierarchical Clustering Untuk Sistem Rekomendasi Film," l Ilmu Komputer dan Sistem Informasi, vol. 11, no. 1, pp. 1-6, 2023.

[17] A. T. R. Dani, S. Wahyuningsih and N. A. Rizki, "Penerapan rchical Clusterinfg Metode Agglomerative pada Data Runtun Waktu," ura Journal of Mathematics, vol. 1, no. 2, pp. 64-78, 2019.

[18] R. Sarno, S. I. Sabilla, Malikhah, Purbawa and Ardani, Machine ing Deep Learning Konsep dan Pemrograman Python, Yogyakarta: bit ANDI, 2023.

[19] A. R. Gunawan, Sudarmin and Z. Rais, "Applied the Self-Organizing (SOM) Method for Clustering Educational Equity in South Sulawesi," *IS Journal of Mathematics and Applied Science*, vol. 4, no. 1, pp. 6-19,

[20] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode v dan Silhouette pada Algoritma Clustering K-Medoids dalam elompokan Produksi Kerajinan Bali," *Jurnal Matrix*, vol. 9, no. 3, pp. 09, 2019.

[21] N. Thamrin and A. W. Wijayanto, "Comparison of Soft and Hard ering: A Case Study on Welfare Level in Cities on Java Island," *esian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 141-2021.

[22] A. Muhid, Analisis Statistik Edisi ke-2, Sidoarjo: Zifatama Jawara, 2019.

[23] H. Khusnuliawati, "Algoritma Pengelompokkan Menggunakan Self-ıizing Map dan K-Means pada Data Sumber Daya Manusia Provinsi esia," *Jurnal Gaung Informatika*, vol. 11, no. 1, pp. 1-9, 2018.

[24] N. Husna, F. Hanum and M. F. Azrial, "Pengelompokkan Produk ısan yang harus Dihindari Penderita Diabetes Menggunakan Algoritma ans Clustering," *InfoTekJar: Jurnal Nasional Informatika dan Teknologi gan*, vol. 4, no. 1, pp. 167-174, 2019.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2342, Copyright ©2025
Submitted : April 28, 2025, Revised : May 11, 2025, Accepted : May 14, 2025, Published : May 26, 2025