JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 9 (ISSUE 1) JUNE 2025: 65 – 78
https://doi.org/10.21009/JSA.09106

FOOD AND BEVERAGE PRODUCT SEGMENTATION BASED ON NUTRITION FACTS USING THE DBSCAN METHOD

Dhika Nurul Fadlilah[1], Arum Handini Primandari[2*]

[1,2]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia
Jl. Kaliurang KM 14.5, Yogyakarta, 55584, Indonesia

Corresponding author's e-mail: ***primandari.arum@uii.ac.id***

Article History:

Received: April 12, 2025
Revised: June 26, 2025
Accepted: June 29, 2025
Published: June 30, 2025
Available online.

Keywords:

Cluster Analysis; DBSCAN; Diabetes; Nutrition Facts; Candy; Chocolate.

ABSTRACT

Type 2 diabetes mellitus is increasingly affecting not only teenagers and adults in Indonesia but also children. This serious issue is linked to high-sugar foods, particularly candy and chocolate products consumed by children. The aim of this research is to categorize these products based on their nutritional information, specifically total fat, saturated fat, sugar, and salt (SSF) content per serving, using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method. By doing so, the study seeks to produce simplified product labels that offer clearer nutritional insights compared to conventional nutrition facts labels. Data was collected through purposive sampling from three retail stores. The clustering results, using parameters Eps 0.4 and MinPts 10, revealed two distinct clusters and 133 noise points. Cluster 1 consists of 215 products with low levels of total fat, saturated fat, sugar, and salt, while Cluster 2 includes 27 products that are high in these nutrients. The clustering quality is validated with a Silhouette Coefficient of 0.77 and a Davies-Bouldin Index of 0.345.

How to cite this article:
D.N. Fadlilah, A.H. Primandari, "FOOD AND BEVERAGE PRODUCT SEGMENTATION BASED ON NUTRITION FACTS USING THE DBSCAN METHOD", Jurnal Statistika dan Aplikasinya, vol. 9, iss. 1, pp. 65 – 78, June 2025

Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Modern society often overlooks healthy living habits by consuming instant foods and beverages without considering their nutritional value and engaging in minimal physical activity. This trend is known to increase the risk of non-communicable diseases (NCDs), particularly diabetes mellitus [1]. Diabetes mellitus is a chronic condition characterized by elevated blood glucose levels, which can result from impaired insulin production or utilization. The condition is categorized into type 1, type 2, gestational diabetes, and other specific types related to different causes [2].

Type 2 diabetes is especially concerning due to its rising prevalence and its strong connection to lifestyle choices. The International Diabetes Federation (IDF) predicts that the number of global diabetes cases will reach 643 million by 2030 and 784 million by 2045. Additionally, Indonesia ranks fifth in the world with 19.5 million diabetes cases reported in 2021 [3].

Although type 2 diabetes generally affects adults, there is a rising number of cases in children and adolescents. According to the chairman of the Indonesian Pediatric Association (IDAI), children and adolescents are vulnerable to type 2 diabetes due to their consumption of artificially sweetened foods and beverages and excessive sugar intake [4]. At the beginning of 2023, IDAI also reported a 70-fold increase in diabetes cases among children since 2010. The IDAI chairman acknowledged that the increase in type 2 diabetes cases in children is due to unhealthy lifestyle patterns [5].

As is well known, sweet foods and beverages are among the favorite items for children. Confectionery/candy and chocolates are sweet foods that are high in sugar and calories but low in nutrients. The sugar in these sweet foods is generally artificial sweeteners, which are sweeter than natural sweeteners. Frequently consuming sweet foods can lead to addiction, causing children to continuously crave them. When children have a habit of consuming sweet foods and do not adopt a healthy lifestyle, such as consuming a balanced diet and exercising regularly, they are at risk of becoming obese. If this continues, the risk of pancreas damage and insulin resistance increases, which can lead to diabetes [6].

In this context, the importance of reading and understanding nutrition facts on food and beverage products becomes even more critical. The Indonesian Food and Drug Authority (BPOM RI) emphasizes the importance of consumers being diligent in reading nutrition facts on food and beverage product labels, especially for diabetic patients who need to limit their sugar intake [7]. In response, BPOM RI has initiated the labeling of nutrition levels on processed food products, which will be implemented gradually. This initiative is driven by the high rates of non-communicable diseases in Indonesia, with diabetes being one of the leading causes of death. Contributing factors include excessive consumption of sugar, salt, and fat (SSF, Indonesian: *Gula, Garam, Lemak* — GGL). According to a 2014 Ministry of Health survey, around 29.7% of Indonesia's population consumes SSF above the standard. As a solution, BPOM will implement a nutri-level system with several tiers indicating the SSF content in the product. This policy aims to educate and help the public understand the nutritional content of products they consume, with an emphasis on labels that are easy to read and understand for Indonesians [8].

To classify food and beverage products based on their high or low SSF factor, particularly within the confectionery and chocolate categories, we commonly use clustering analysis. Clustering is an unsupervised learning approach that allows us to gain insights into data distribution and examine the characteristics of each cluster for further analysis.

There are various methods of clustering, and in this research, we applied the density-based clustering method known as Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN groups neighboring data points into a single cluster, making it especially useful for datasets that contain many outliers or noisy data, as the method automatically disregards points considered outliers and does not include them in any cluster. This algorithm was introduced by Ester et al. in 1996. The input parameters for the DBSCAN method are epsilon (ε) and minimum points (MinPts) [9].

A study by Bej et al. (2022) demonstrated the use of DBSCAN in identifying meaningful clusters of Type 2 Diabetes Mellitus patients based on diverse epidemiological features such as nutritional habits, lifestyle, and socio-demographics. This highlights DBSCAN's capability in handling mixed data types and detecting outliers in complex health-related datasets [10]. However, research specifically applying clustering methods to food and beverage products that pose a risk of causing diabetes remains limited. Another study clustered packaged food products that should be avoided by diabetic patients based on

their fat, protein, sugar, and sodium content using the K-Means algorithm [11]. Nonetheless, the study lacked a clear explanation regarding the sampling design and selection criteria for the products analyzed.

Although previous studies have applied DBSCAN for food clustering based on general nutrition facts such as carbohydrates, proteins, calories, and fats, this study introduces a new approach by focusing specifically on confectionery and chocolate products. These products are closely associated with the increasing prevalence of type 2 diabetes in children. In contrast to studies that only form clusters without further interpretation, this research provides added value by profiling each cluster according to its sugar, salt, and fat (SSF) content. The study also assigns simplified nutrition-based labels intended to be easier to understand than conventional nutrition facts labels. These simplified labels are designed to help the public, especially individuals with diabetes and caregivers of children at risk, in making better-informed food choices. In addition, this study addresses a research gap related to the lack of clarity in sampling methods by using a purposive sampling technique. This technique ensures that the selected products are those that are frequently consumed by children. The uniqueness of this research lies not only in the use of the DBSCAN method but also in the focus on high-risk products, the development of understandable product labeling, and its relevance to Indonesia's national nutrition labeling program known as the nutri-level policy.

Based on these considerations, this study applies the DBSCAN method to group confectionery and chocolate food and beverage products based on their nutritional facts. The aim is to help consumers, especially those who have or are at risk of diabetes, make healthier purchasing decisions. The results of this segmentation can also serve as input for the Indonesian Food and Drug Authority (BPOM RI) in supporting the implementation of nutrition labeling based on SSF content. Moreover, while previous studies on product classification and nutritional analysis often lacked transparency in their sampling procedures, this study improves upon those efforts by applying a purposive sampling method. This ensures that the data collected are relevant to the context of the research and representative of products commonly found in the market.

2. METHODS

Material and Data

This research focuses on food and beverage products categorized as confectionery/candy and chocolate based on BPOM Regulation No. 13 of 2023 (Category 05.0). The sample consists of 375 products collected from three stores: Manna Kampus Simanjuntak Yogyakarta, Toko Agung Grosir Yogyakarta, and KN Putra Toserba Magelang. We employed a non-probability sampling technique with purposive sampling, selecting products based on criteria that included being packaged, commercially available, and possessing a nutrition facts label. Data were collected by photographing the brand and nutrition facts label using a mobile phone. The variables analyzed include total fat, saturated fat, sugar, and salt in grams per serving size.

Table 1. Operational Definition of Research Variables

Variable	Definition	Unit	Scale
Total Fat	Total fat is all types of fat contained in food and drinks, including saturated fat, unsaturated fat, and trans-fat.	Grams	Ratio
Saturated Fat	Saturated fat is a type of fat that generally comes from animal products.	Grams	Ratio
Sugar	Sugar is a simple carbohydrate that is used as a natural sweetener or addition in food and beverage.	Grams	Ratio
Salt	Salt is a mineral that is used to give a salty taste to food and beverage.	Grams	Ratio

Figure 1. Example of Sample Data Collection

Research Method

The following is the research flowchart.

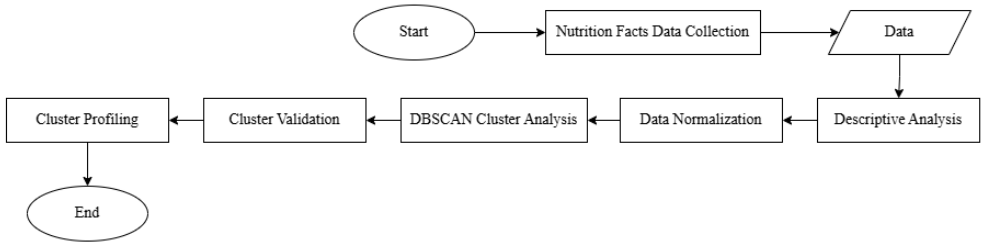

Figure 2. Research Flowchart

Descriptive Statistics

Descriptive statistics is a statistical method that aims to collect, organize, and process data to provide a clear picture of a specific condition or event from which the data was taken. In general, data presented in descriptive statistics takes the form of tables, graphs, or diagrams, and measures of central tendency such as mean and median [12]. In this research, the data will be presented in the form of data summaries to understand the general overview of the data, including minimum values, maximum values, mean, and boxplots.

Data Normalization

In the process of data analysis, it is common to encounter variables with differing value ranges. Such differences can influence the outcome of the analysis; therefore, a data normalization or standardization method is required. Data normalization is a technique used to standardize the scale of attribute values to a smaller, uniform range with equal weighting. Among various normalization methods, Z-score normalization is one of the most widely used. This method normalizes data based on its mean and standard deviation. One of its advantages is its robustness in handling outliers, as it produces values that are more stable in their presence. The Z-score normalization can be calculated using the following formula [13]:

$$z_I = \frac{x_I - \bar{x}}{s} \quad (1)$$

where z_I is the result Z-score normalization value for data x_I, x_I is the data value to be normalized, $\bar{x}$ is the mean value in a variable, and s is the standard deviation value in a variable.

Density Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a clustering method based on the concept of density-connected points. This algorithm was designed by Ester et al. in 1996 and is useful for identifying clusters in large spatial datasets by examining the local density of data elements using two parameters, Epsilon and MinPts. The DBSCAN process is also fast and highly efficient for databases of various sizes [14]. DBSCAN offers several

advantages that make it an effective clustering method. It does not require prior information about the number of clusters to be formed, so it can automatically determine the appropriate number of clusters based on the data. It can identify objects that are considered noise. It only requires two main parameters, epsilon (ε) and MinPts, which are not sensitive to the order of points/objects in the data [15]. Furthermore, the DBSCAN algorithm is not only suitable for spatial data but can also be applied to various non-spatial datasets, such as image processing, fraud detection, or health monitoring, and recommendation systems [16].

Steps in performing cluster analysis using the DBSCAN method [9]:

1. Determine the values of Epsilon (ε) and MinPts.
2. Select an arbitrary starting point p.
3. Calculate the distance from point p to all other points using the Euclidean distance formula.

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad (2)$$

   where $d(i,j)$ is the distance between point i and point j; $x_{i1}, x_{i2}, \ldots, x_{ip}$ is the value from point i; and $x_{j1}, x_{j2}, \ldots, x_{jp}$ is the value from point j.
4. Select all points that are density reachable from point p.
5. If p is a core point, meaning it has at least MinPts neighbouring points within a radius of ε, then a cluster is formed starting from p and all density-reachable points from p will be included in the same cluster.
6. If p is a border point, which is a point located within the ε-neighborhood of a core point but has fewer than MinPts neighbors itself, and there are no density-reachable points from p, DBSCAN will move to another unvisited point in the dataset.
7. Repeat the process until all points have been processed.

The DBSCAN method was selected in this study due to its ability to identify clusters of arbitrary shape and its robustness in handling noise and outliers, which are common in real-world nutritional data. Although the dataset consists of 375 observations and four nutritional variables, preliminary exploration showed that several products contain zero or extreme values that could potentially distort the clustering outcome if traditional methods such as K-Means were applied. Unlike K-Means, which requires the number of clusters to be specified in advance and is sensitive to noise, DBSCAN does not rely on such assumptions. This makes it more suitable for uncovering natural groupings in a dataset where the cluster boundaries are not clearly defined and where the presence of noise points is expected. Furthermore, DBSCAN has been proven effective in nutrition-related clustering problems, particularly when the focus is on identifying meaningful groups while filtering out atypical or inconsistent product profiles.

Cluster Validation

Cluster validation is a procedure that quantitatively and objectively evaluates the results of cluster analysis. It is used to assess whether the resulting clusters can accurately represent and explain the overall population. Cluster validity is essential for addressing the fundamental issue of determining the optimal number of clusters [17]. In this research, the metrics used for validation are the Silhouette Coefficient and the Davies-Bouldin Index.

The silhouette coefficient is a clustering evaluation technique that measures how similar an object is to its own cluster compared to other clusters, as well as the separation between clusters. The silhouette coefficient ranges from -1 to 1; the closer the value is to 1, the better the clustering result [18]. A positive value approaching 1 indicates that a data point is closer to its own cluster than to the nearest neighboring cluster, suggesting that the object is appropriately clustered. A negative value approaching -1 implies that the data point may have been assigned to the wrong cluster. A value near 0 indicates that the data point lies close to the boundary between two clusters, implying unclear separation between clusters [19]. The silhouette coefficient can be calculated using the following formula [18].

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

where $s(i)$ is silhouette coefficient for data point i. The values of $a(i)$ and $b(i)$ are calculated using the Euclidean distance formula shown in Equation (2). Specifically, $a(i)$ is the average distance between data point i and all other points in the same cluster, while $b(i)$ is the minimum average distance between data point i and all points in the nearest neighboring cluster.

The Davies-Bouldin Index (DBI) is an internal evaluation metric used to assess the quality of clustering results. This metric measures the distance between each cluster and all other clusters. DBI takes into account the average distance between the centroids of each pair of clusters and the relative size of the two neighboring clusters. The lower the DBI value, or the closer it is to zero, the more optimal the clustering result [20]. DBI can be calculated using the following formula [21]:

$$DBI = \frac{1}{k}\sum_{m}^{k} max_{m \neq n}(R_{m,n}) \tag{4}$$

where k is the number of clusters and $R_{m,n}$ is the similarity measure between cluster m and n, which is defines as [21]:

$$R_{m,n} = \frac{SSW_m + SSW_n}{SSB_{m,n}} \tag{5}$$

SSW_m and SSW_n represent the average intra-cluster distance (within-cluster scatter) for cluster m and n, respectively. Meanwhile $SSB_{m,n}$ denotes the inter-cluster distance between the centroids of clusters m and n, computed using the Euclidean distance in Equation (2).

Mann-Whitney U Test

The Mann-Whitney U test is a nonparametric statistical test that serves the same purpose as the t-test in parametric statistics, which is to compare two independent samples to determine whether there is a significant difference between them. In parametric statistics, the analysis of two samples is conducted using the t-test, but only if certain assumptions are met. If any of these assumptions are not fulfilled, the t-test must be replaced with a nonparametric statistical test specifically designed for two independent samples [22]. The null hypothesis for the Mann-Whitney U test is there is no difference (in terms of central tendency) between the two groups in the population.

3. RESULTS

Descriptive Analysis

Table 2. presents a summary of the data for each variable in grams, based on 375 products in the confectionery/candy and chocolate category.

Table 2. Data Summary

Variable	Minimum (g)	Maximum (g)	Mean (g)
Total Fat	0	15	2.349
Saturated Fat	0	9	1.526
Sugar	0	23	6.1
Salt	0	0.135	0.011

According to Table 2, all variables have a minimum value of 0 grams, indicating the presence of products with no total fat, saturated fat, sugar, or salt content. Total fat content reaches a maximum of 15 grams, with an average of 2.349 grams, while saturated fat has a maximum of 9 grams and an average of 1.526 grams. Compared to the recommended daily fat intake limit of 67 grams, the levels of total fat and saturated fat in confectionery/candy and chocolate products are still within a reasonable range.

Sugar content shows the highest maximum value among all variables, reaching 23 grams with an average of 6.1 grams. This indicates that certain products contain sugar levels approaching half of the recommended daily intake, which is 50 grams per day. Meanwhile, salt content has the lowest maximum value, only 0.135 grams, with an average of 0.011 grams. This suggests that most products have very minimal salt content, and considering the recommended daily salt intake of 5 grams, the salt levels in these products are well within safe limits. Overall, products in the confectionery and chocolate category tend to have higher sugar content compared to total fat, saturated fat, or salt.

To detect outliers in the data, a boxplot visualization was created for each variable. This visualization helps to observe the data distribution and interquartile range (IQR) and identify values outside the normal range, known as outliers. Figure 3 shows that for each variable, there are points located outside the whiskers, indicating the presence of outliers. This suggests the existence of extreme values or products with nutrient content significantly higher than the average for each variable.

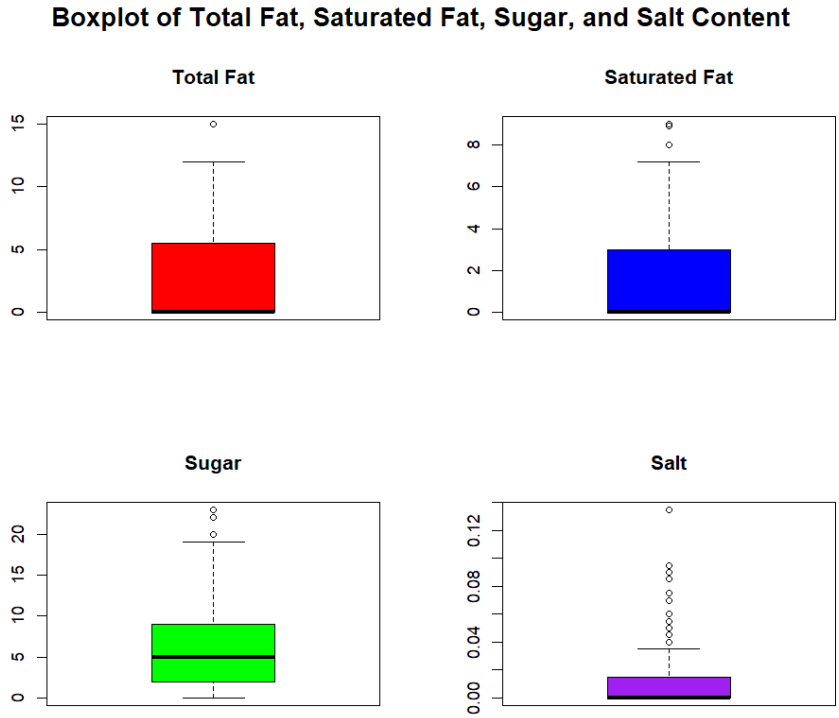

Figure 3. Boxplot Data

Data Normalization

Z-Score normalization was performed because the existing data lacked uniformity and exhibited significantly different value ranges across variables. The purpose of this normalization is to standardize the scale of each variable, making it easier to carry out following analysis.

Table 3. Data Normalization Result

No	Brands	Total Fat	Saturated Fat	Sugar	Salt
1.	Alba Pastilles Classic Flavour	-0.690	-0.645	-1.061	-0.603
2.	Alfie Crunchy with Bits Milk Chocolate Flavour	1.778	2.401	1.789	1.232
3.	Alfie Crunchy with Bits Strawberry & Milk Flavour	2.248	3.120	1.768	1.232
…	…	…	…	…	…
373.	Yupi Strawberry Kiss	-0.690	-0.645	0.811	-0.603
374.	Yupi Sweet Heart	-0.690	-0.645	0.187	-0.603
375.	Ziplong Eucalyptus Mint	-0.690	-0.645	-1.061	-0.603

The normalized data resulted in more standardized or uniform values, as it has a mean close to 0 and a standard deviation of 1. This facilitates further analysis without being affected by the original scale of the data.

DBSCAN Analysis

The DBSCAN method requires two main parameters to form a cluster: Epsilon and MinPts. Epsilon determines the maximum distance between objects within the same group, while MinPts sets the minimum number of neighboring objects required for a point to be considered part of a cluster. These two parameters greatly influence the resulting number of clusters. Therefore, various combinations of Eps (0.3 to 0.9) and MinPts (2 to 10) were tested to find the optimal values that yield the best clustering performance. Since DBSCAN does not have a standardized mechanism for determining these parameters automatically, a trial-and-evaluation approach was adopted by testing multiple combinations of Epsilon and MinPts. This experimental approach is widely applied in clustering research as it allows researchers to empirically determine parameter values that yield the most meaningful and valid cluster structures. By relying on objective internal validation metrics such as the silhouette coefficient and the Davies-Bouldin Index (DBI), this method enhances reproducibility and reduces the subjectivity that may arise from heuristic or visual selection techniques.

Table 4. Results of Optimal Eps and MinPts Parameter Combination Experiments

No	MinPts	Eps	Amount of Cluster	Amount of Noise	Silhouette Coefficient	Davies-Bouldin Index
1.	2	0.3	20	95	0.171	0.568
2.	2	0.4	24	75	0.185	0.562
3.	2	0.5	21	58	0.406	0.581
…	…	…	…	…	…	…
58.	10	0.3	2	151	0.77	0.286
59.	10	0.4	2	133	0.77	0.345
60.	10	0.5	2	123	0.762	0.372
…	…	…	…	…	…	…
64.	10	0.9	2	53	0.695	0.567

Based on Table 4, the parameter combination of Eps 0.4 and MinPts 10 is considered the optimal choice for product segmentations using the DBSCAN method. This combination yields a silhouette coefficient value of 0.77, indicating good cluster separation, and a Davies-Bouldin Index (DBI) of 0.345, which is relatively low compared to other combinations. Additionally, the number of noise points produced is 133, which is fewer than the 151 noise points generated by the combination of Eps 0.3 and MinPts 10. Considering the balance between a lower number of noise points, high-quality cluster separation, and a DBI value that remains within a reasonable range, Eps 0.4 and MinPts 10 are selected for cluster analysis using the DBSCAN method. The clustering results using this parameter combination are visualized in Figure 4 below.

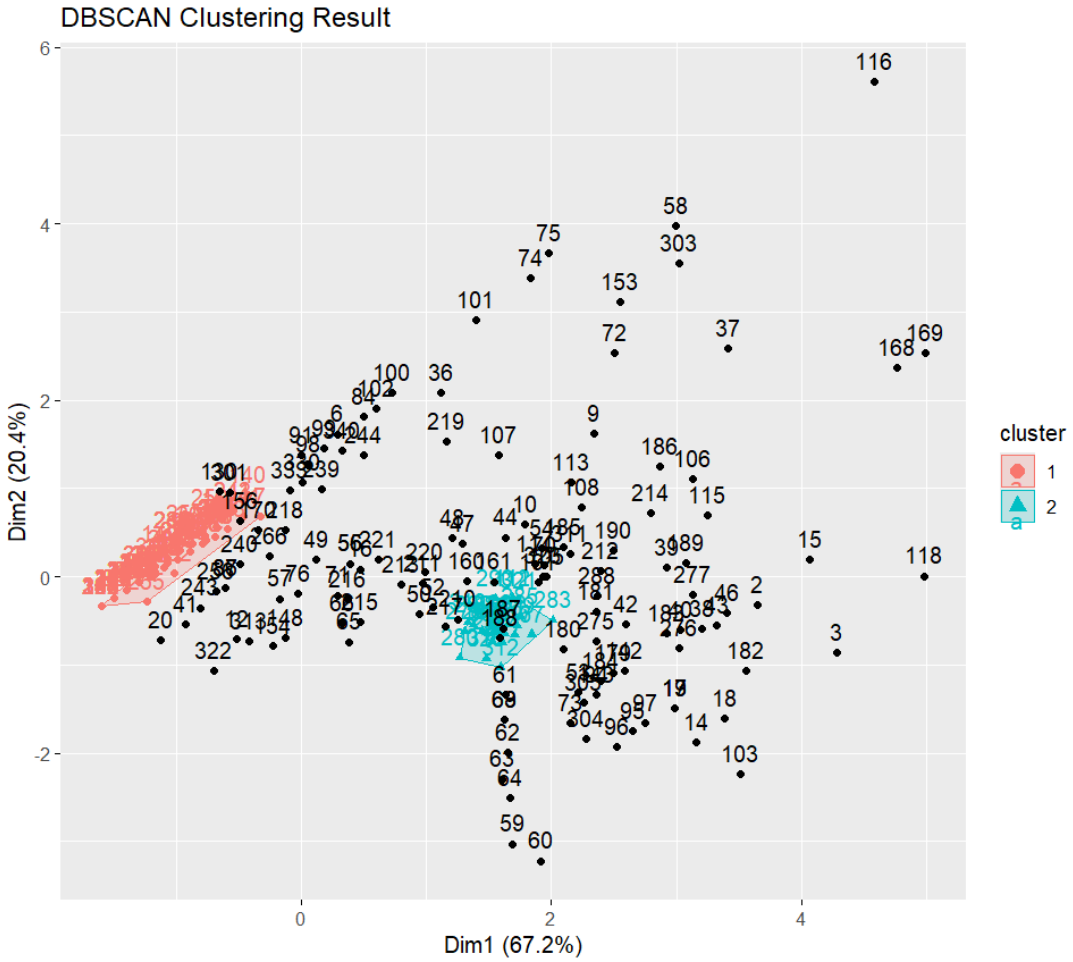

Figure 4. Visualization of DBSCAN Clustering Result

Figure 4 is a visualization of the clustering results using epsilon 0.4 and MinPts 10. The figure shows that the DBSCAN method successfully grouped the data into two distinct clusters, represented by red and blue colors. The black points scattered around the clusters indicate data classified as noise, which were not assigned to any cluster. To obtain this visualization, the four nutritional variables (total fat, saturated fat, sugar, and salt) were first standardized using Z-score normalization to ensure comparability across features. The DBSCAN algorithm was then applied with the selected parameters, and the resulting cluster assignments were visualized using a two-dimensional scatter plot. The axes represent the first two principal components derived from Principal Component Analysis (PCA), which reduces the four-dimensional dataset to two dimensions for easier visualization. Each point was plotted and color-coded based on its cluster label, enabling clear identification of grouped products and outliers.

Cluster Validation

Using the parameters Eps 0.4 and MinPts 10, a silhouette coefficient value of **0.77** was obtained. This value is close to 1, indicating that the clustering results have a well-defined structure, with data points within the same cluster being cohesive and clearly separated from other clusters. Meanwhile, the Davies-Bouldin Index (DBI) value is **0.345**, suggesting that the average ratio of inter-cluster distance to cluster size is relatively low. A DBI value closer to 0 indicates that the formed clusters are well-separated and compact in size. Overall, the validation scores from both the silhouette coefficient and the Davies-Bouldin Index show that the clustering results using the DBSCAN method with Eps 0.4 and MinPts 10 produced well-formed and representative clusters.

Product Segmentation Results

After performing clustering analysis using the DBSCAN method with parameters Eps 0.4 and MinPts 10 on 375 products in the confectionery/candy and chocolate category, the product segmentation results are presented in Table 5.

Cluster 0 consists of 133 products categorized as noise, referring to products that do not belong to any cluster due to not meeting the proximity criteria to other clusters. Cluster 1 contains the largest number of members, with 215 products, indicating that the majority share similar characteristics. Meanwhile, Cluster 2 includes 27 products, representing a smaller group with distinct characteristics from the majority.

To facilitate access and product search in this research, a product database for the confectionery/candy and chocolate category is available on https://bit.ly/ProductSegmentationWithDBSCAN. This database contains detailed information on the 375 sample products used in the research, along with a feature that automatically identifies whether a product belongs to Cluster 0, Cluster 1, or Cluster 2.

Table 5. Product Segmentation Result

Cluster	Number of Members	Brands
0 (Noise)	133	Alfie Crunchy with Bits Milk Chocolate Flavour, Alfie Crunchy with Bits Strawberry & Milk Flavour, Alpenliebe Smooth Caramel Milk, Amanda Sweet KOKONA Brown Sugar, Amanda Sweet KOKONA Less Sugar, etc.
1	215	Alba Pastilles Classic Flavour, Alpenliebe Eclairs Choco, Alpenliebe Juicyfills, Alpenliebe Smooth Susu Stroberi, Alpenliebe Stick, etc.
2	27	Chocolate Monggo Matcha White Chocolate, Delfi Almond Dairy Milk Chocolate, Delfi Take it Black Chocolate dengan Wafer, Delfi Take it Milk Chocolate with Wafer, Delfi Take it Green Tea, etc.

Cluster Profiling

Cluster profiling was conducted to better understand the characteristics of each cluster. Through this action, it can be identified whether a cluster tends to have higher or lower average levels of total fat, saturated fat, sugar, and salt compared to the others. The profiling results are presented in Table 6. below.

Table 6. Cluster Profiling

Cluster	Total Fat	Saturated Fat	Sugar	Salt
1	0.016	0.002	3.395	0.002
2	6.741	3.852	7.852	0.017

The cluster profiling results in Table 6 show differences in the average content of total fat, saturated fat, sugar, and salt among confectionery/candy and chocolate products within each cluster. Based on these results, it can be concluded that Cluster 1 consists of products with lower levels of total fat, saturated fat, sugar, and salt compared to Cluster 2. In contrast, Cluster 2, which has fewer members than Cluster 1, consists of products with higher contents of total fat, saturated fat, sugar, and salt.

To determine whether there is a significant difference in the mean values of total fat, saturated fat, sugar, and salt between Cluster 1 and Cluster 2, an independent sample t-test was initially considered. However, since the assumptions of data normality and homogeneity of variances were not met in this research, the nonparametric Mann-Whitney U test, which serves the same purpose, was conducted instead. Here are the results of the Mann-Whitney U test.

I. Hypotheses

H_0 : there is no significant difference in the average nutritional content between Cluster 1 and Cluster 2.

H_1 : there is a significant difference in the average nutritional content between Cluster 1 and Cluster 2.

II. Significance Level
$\alpha = 5\% = 0.05$

III. Critical Region
Reject H_0 if *p-value* < α

IV. Test Statistics
In this research, the test statistic reported is the Wilcoxon Rank Sum statistic (W), as generated by the R software.
Total Fat: W = 0; Saturated Fat: W = 0; Sugar: W = 538; Salt: W = 162.

V. Decision and Conclusion

Table 7. Mann-Whitney U Test Result

Variable	*p-value*	Sign	α	Decision
Total Fat	$< 2.2 \times 10^{-16}$	<	0.05	Reject H_0
Saturated Fat	$< 2.2 \times 10^{-16}$	<		Reject H_0
Sugar	3.268×10^{-12}	<		Reject H_0
Salt	$< 2.2 \times 10^{-16}$	<		Reject H_0

Based on the Mann-Whitney U Test Result presented in Table 7., and using a 95% confidence level, it can be concluded that there is a significant difference in the average content of total fat, saturated fat, sugar, and salt between Cluster 1 and Cluster 2.

The results of the Mann-Whitney U test demonstrate that the two clusters have significant differences in the average content of total fat, saturated fat, sugar, and salt. These differences indicate that the clusters have distinct nutritional characteristics. Cluster 1 tends to have lower levels of total fat, saturated fat, sugar, and salt compared to Cluster 2. This variation in characteristics is due to the fact that Cluster 1 consists of confectionery/candy products, while Cluster 2 consists of chocolate products. Based on the profiling results, the characteristics of each cluster can be visualized as follows.

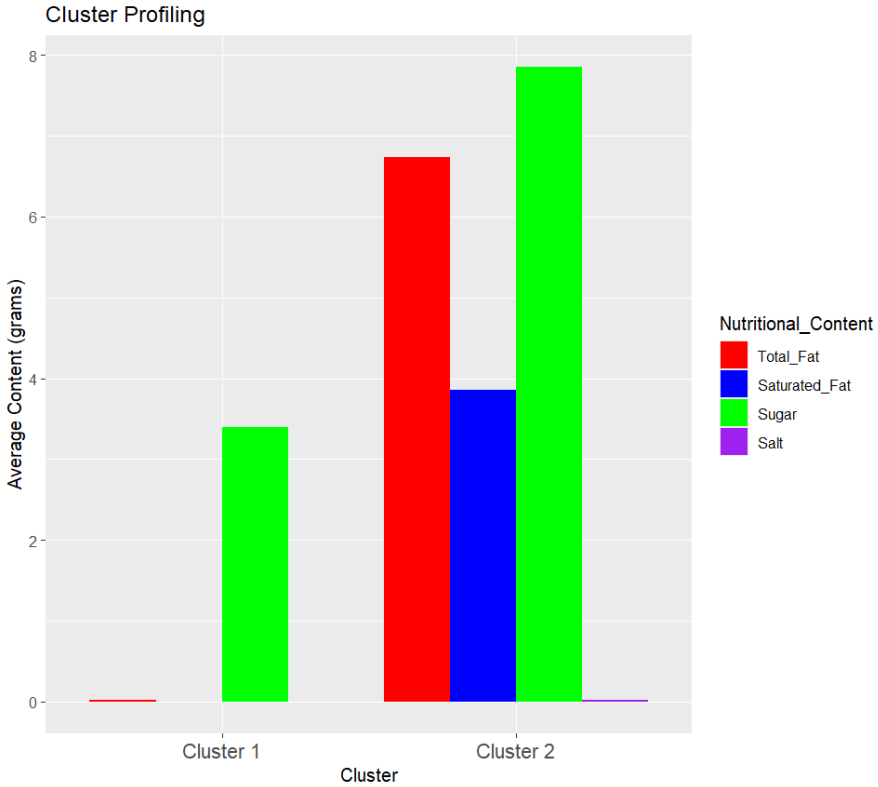

Figure 5. Cluster Profiling Visualization

Descriptive Analysis of Noise

In performing cluster analysis using the DBSCAN method, some data points were identified as noise. Out of 375 products in the confectionery/candy and chocolate category analyzed, 133 products were identified as noise and thus were not assigned to any cluster. To understand why these products were classified as noise, a descriptive analysis was performed by examining their characteristics and comparing them with those of the two formed clusters. This analysis aims to identify the main differences between noise products and those successfully grouped into clusters. The average content of total fat, saturated fat, sugar, and salt in the 133 noise products is presented in Table 8.

Table 8. Average Nutritional Content of Noise Products

Total Fat	Saturated Fat	Sugar	Salt
5.227	3.516	10.117	0.023

According to Table 8, the average levels of total fat, saturated fat, sugar, and salt in the noise products are relatively high. When compared to the nutritional averages of Cluster 1 and Cluster 2 in Table 6, the total fat and saturated fat content in the noise products is higher than in Cluster 1 but lower than in Cluster 2. Meanwhile, the sugar content in the noise products is higher than in both clusters, indicating that products categorized as noise tend to have a greater sugar level. The salt content is also slightly higher in the noise products compared to both clusters, although the difference remains relatively small. These differences suggest that the noise products exhibit characteristics that do not fully align with the patterns formed in Cluster 1 and Cluster 2, making them ungroupable by the DBSCAN method.

4. DISCUSSIONS

The clustering analysis using the DBSCAN method successfully segmented confectionery and chocolate products into two meaningful clusters, along with a substantial group of noise points. The results reflect the real-world variation in nutritional composition across product types, particularly in the context of diabetes risk.

The use of Eps 0.4 and MinPts 10 yielded the optimal clustering structure, as evidenced by a high Silhouette Coefficient (0.77) and a low Davies-Bouldin Index (0.345), indicating well-separated and compact clusters. Cluster 1, which contains 215 products, generally includes items with lower levels of total fat, saturated fat, sugar, and salt. These are mostly candies. In contrast, Cluster 2, consisting of 27 products, includes items with higher nutrient content, mostly chocolate-based products. These findings suggest a significant nutritional distinction between confectionery/candy and chocolate products. Interestingly, 133 products were categorized as noise, not fitting into either cluster. Descriptive analysis shows that these noise products often contain high sugar and salt levels, exceeding those in both clusters, which may indicate extreme or inconsistent nutritional profiles. Their exclusion highlights the strength of DBSCAN in filtering outliers that could distort pattern recognition.

The results of this study are in line with Bej et al. (2022), who showed that DBSCAN can form meaningful groups and detect data that does not follow common patterns [10]. Compared to the study by Husna et al. (2019), which used the K-Means algorithm to group food products that may pose risks to people with diabetes [11], this study gives clearer results because DBSCAN can separate products that have unusual nutrition patterns. Unlike K-Means, which puts all data into a group, DBSCAN can leave out products that don't fit into any group without saying whether those products are dangerous or not. In addition, the grouping in this study is based on sugar, salt, and fat (SSF) content, which matches BPOM RI's nutri-level labeling program. This approach is relevant to public health, especially with the increasing number of diabetes cases in children. Unlike other studies that look at many types of food, this study focuses only on candy and chocolate products that are more risky, so the results are more useful for consumers. The quality of the clustering is also supported by two statistical measures, the silhouette coefficient and Davies-Bouldin Index, which show that the product groups are well separated.

The findings of this research have several important implications for various stakeholders. For individuals with diabetes or those at risk, the segmentation results provide practical guidance: products classified in Cluster 2 should be consumed with caution, as they contain high levels of total fat, saturated

fat, sugar, and salt. In terms of policy implications, this study offers input for the implementation of the nutri-level labeling system in Indonesia. Specifically, confectionery/candy and chocolate products can be divided into two levels based on their nutritional profiles. Level 1 may represent products with lower contents of total fat, saturated fat, sugar, and salt, while Level 2 includes products with higher levels of these nutrients. However, policymakers should also consider the existence of products categorized as noise, which may not conform to the main clustering patterns but still require attention in labeling and regulation. Lastly, in terms of practical and social impact, the segmentation outcomes from this study could be integrated into a public-facing website or interactive application. Such a platform would allow consumers to access detailed information about confectionery/candy and chocolate products, categorized by their SSF content. This could support healthier purchasing decisions and increase public awareness regarding nutritional risks, particularly among vulnerable groups such as children.

5. CONCLUSION

This research employed the DBSCAN method to cluster confectionery/candy and chocolate products based on their nutrition facts labels. The data showed high sugar dominance and generally low salt levels, with several products containing zero nutrients. Using the optimal parameters of epsilon 0.4 and MinPts 10, the DBSCAN algorithm formed two distinct clusters and identified 133 products as noise. Cluster validation indicated strong performance, with a Silhouette Coefficient of 0.77 and a Davies-Bouldin Index of 0.345. Cluster profiling revealed that Cluster 1, consisting mainly of confectionery/candy products, had lower fat, sugar, and salt levels, while Cluster 2, composed mostly of chocolate products, contained higher levels of these nutrients.

For future research, it is recommended to expand the analysis through product classification techniques to achieve a clearer and more accurate segmentation of food products. Additionally, further examination can be conducted on the products identified as noise (Cluster 0) to explore their characteristics and potential grouping. The segmentation results from this study can also be developed into an interactive web or mobile application to inform the public about the sugar, salt, and fat (SSF) content levels in confectionery and chocolate products, thereby supporting healthier consumption choices.

6. REFERENCES

[1] I. D. A. E. C. Astutisari, A. Y. Darmini and I. A. P. Wulandari, "Hubungan Pola Makan dan Aktivitas Fisik dengan Kadar Gula Darah pada Pasien Diabetes Melitus Tipe 2 di Puskesmas Manggis I," *Jurnal Riset Kesehatan Nasional,* vol. 6, no. 2, pp. 79-87, 2022.

[2] PERKENI, Pedoman Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 Dewasa di Indonesia 2021, Jakarta: PB PERKENI, 2021.

[3] IDF, IDF Diabetes Atlas 10th Edition, 2021.

[4] M. Syafaruddin, "Diabetes pada Anak, Ancaman Nyata yang Perlu Diwaspadai," 18 July 2024. [Online]. Available: https://www.suarasurabaya.net/kelanakota/2024/diabetes-pada-anak-ancaman-nyata-yang-perlu-diwaspadai/.

[5] N. Fatiara, "IDAI Ungkap Makin Banyak Anak Idap Penyakit Diabetes dan Ginjal," 24 Juli 2024. [Online]. Available: https://kumparan.com/kumparanmom/idai-ungkap-makin-banyak-anak-idap-penyakit-diabetes-dan-ginjal-23Bq7hURKCQ/full.

[6] d. M. D. C. Pane, "Benarkan Makanan Manis Bisa Memicu Diabetes pada Anak?," 14 Februari 2023. [Online]. Available: https://www.alodokter.com/benarkah-makanan-manis-bisa-memicu-diabetes-pada-anak.

[7] borneonews, "Pentingnya Memahami Informasi Nilai Gizi pada Makanan Olahan," 16 Oktober 2024. [Online]. Available: https://www.borneonews.co.id/berita/395786-pentingnya-memahami-informasi-nilai-gizi-pada-makanan-olahan.

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

[9] R. R. Muhima, M. Kurniawan, S. R. Wardhana, A. Yudhana, Sunardi, W. M. Rahmawati and G. E. Yuliastuti, Kupas Tuntas Algoritma Clustering: Konsep Perhitungan Manual dan Program, Yogyakarta: ANDI, 2021.

[10] S. Bej, J. Sarkar, S. Biswas, P. Mitra, P. Chakrabarti and O. Wolkenhauer, "Identification and epidemiological characterization of Type-2 diabetes sub-population using an unsupervised machine learning approach," *Nutrition and Diabetes,* vol. 12, no. 27, 2022.

[11] N. Husna, F. Hanum and M. F. Azrial, "Pengelompokkan Produk Kemasan yang Harus Dihindari Penderita Diabetes Menggunakan Algoritma K-Means Clustering," *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan,* vol. 4, no. 1, pp. 167-174, 2019.

[12] L. D. Martias, "Statistika Deskriptif Sebagai Kumpulan Informasi," *FIHRIS: Jurnal Ilmu Perpustakaan dan Informasi,* vol. 16, no. 1, pp. 40-59, 2021.

[13] G. A. B. Suryanegara, Adiwijaya and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi),* vol. 5, no. 1, pp. 114-122, 2021.

[14] I. D. Id, Astrid and E. Mahdiyah, "Modifikasi DBSCAN (Density-Based Spatial Clustering With Noise) pada Objek 3 Dimensi," *Jurnal Komputer Terapan,* vol. 3, no. 1, pp. 41-52, 2017.

[15] R. R. A. Rahman and A. W. Wijayanto, "Pengelompokan Data Gempa Bumi Menggunakan Algoritma DBSCAN," *Jurnal Meteorologi dan Geofisika,* vol. 22, no. 1, pp. 31-38, 2021.

[16] R. Kumar, "A Guide to the DBSCAN Clustering Algorithm," 29 September 2024. [Online]. Available: https://www.datacamp.com/tutorial/dbscan-clustering-algorithm#rdl.

[17] H. A. Pradita, "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Faktor-Faktor Penyebab Kekerasan Seksual," 18 Mei 2022. [Online]. Available: https://rpubs.com/helvaaldha/PenerapanAnalisisCluster.

[18] Indra, N. Nur, M. Iqram and N. Inayah, "Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting," *INOVTEK Polbeng-Seri Informatika,* vol. 8, no. 2, pp. 356-367, 2023.

[19] Y. Cunanda, "Strategi Pemilihan Kluster dalam Analisis Data: Studi Kasus Metode Elbow dan Silhouette Dalam K-Means Clustering," 25 Oktober 2023. [Online]. Available: https://medium.com/@ycunanda/strategi-pemilihan-kluster-dalam-analisis-data-studi-kasus-metode-elbow-dan-silhouette-alam-5831b8413d24.

[20] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan Dbscan," *Jurnal Informatika dan Teknik Elektro Terapan,* vol. 12, no. 3, pp. 3517-3536, 2024.

[21] M. T. Jatipaningrum, S. E. Azhari and K. Suryowati, "Pengelompokan Kabupaten dan Kota di Provinsi Jawa Timur Berdasarkan Tingkat Kesejahteraan dengan Metode K-Means dan Density-Based Spatial Clustering of Application with Noise," *Jurnal Devirat,* vol. 9, no. 1, pp. 70-81, 2022.

[22] A. Quraisy and S. Madya, "Analisis Nonparametrik Mann Whitney Terhadap Perbedaan Kemampuan Pemecahan Masalah Menggunakan Model Pembelajaran Problem Based Learning," *VARIANSI: Journal of Statistics and Its Application on Teaching and Research,* vol. 3, no. 1, pp. 51-57, 2021.