Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10, Number 2, 2026
DOI: 10.30829/zero.v10i2.26756
Page: 549-558

Clustering Indonesian Traditional Foods by Nutritional Profiles using the K-Means Algorithm for Health Policy

[1]Irene Devi Damayanti
Department of Informatics Engineering, Universitas Kristen Indonesia Toraja, 91811, Indonesia

[2]Semuel Yacobus Padang
Department of Informatics Engineering, Universitas Kristen Indonesia Toraja, 91811, Indonesia

[3]Isak Tandi
Department of Informatics Engineering, Universitas Kristen Indonesia Toraja, 91811, Indonesia

[4]Melda Duma'
Department of Agrotechnology, Universitas Kristen Indonesia Toraja, 91811, Indonesia

Article Info

Article history:

Accepted: 10 July 2026

Keywords:

Clustering;
Health policy;
Indonesian traditional foods;
K-Means algorithm;
Nutritional profiles.

ABSTRACT

Indonesia has a wide variety of traditional foods; however, systematic mapping of their nutritional composition remains limited. This study aims to cluster Indonesian traditional foods based on nutritional profiles using the K-Means algorithm to support health policy development. The analysis focuses on calories, protein, fat, and carbohydrates. A quantitative approach was applied, including data selection, normalization, and determination of optimal number of clusters using the Elbow Method. The results show that four clusters (k = 4) were obtained. Cluster 3 contains foods high in calories and protein, Cluster 2 is dominated by carbohydrates, Cluster 0 shows moderate nutritional values, and Cluster 1 represents low-energy foods. Clustering quality was evaluated using the Silhouette Coefficient (0.45) and Davies–Bouldin Index (0.94), indicating moderate and acceptable clustering performance. PCA visualization retained 88.77% of total data variance. Clusters inform policy via dietary grouping, guiding interventions; limited to macronutrients, excluding micronutrients and portion variability.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Irene Devi Damayanti,
Department of Informatics Engineering
Universitas Kristen Indonesia Toraja, Indonesia
Email: irenedamayanti@ukitoraja.ac.id

1. INTRODUCTION

Indonesia is known as a country rich in traditional culinary diversity. Typical foods and beverages from various regions not only reflect local culture, but also have diverse nutritional compositions. However, although traditional foods are believed to be healthier than modern processed foods [1], there are still challenges in understanding the nutritional content of these various types of food systematically and scientifically. Therefore, a data-driven approach is needed to accurately and efficiently categorize Indonesian foods and beverages based on their nutritional profiles.

In recent years, artificial intelligence and machine learning technologies have been widely used to support nutritional data analysis [2], [3]. One of the most commonly used methods is the K-Means Clustering algorithm, which is effective for grouping numerical data such as calorie, protein, fat, and carbohydrate content [4]. By applying this method, foods can be grouped based on the similarity of their nutritional values, which can ultimately be used to develop diet plans, personalized dietary recommendations[5], and prevent diet-related diseases such as obesity [6] and diabetes mellitus [7], [8].

Previous studies have shown the effectiveness of K-Means in grouping foods based on nutritional value [9]. For example, foods can be classified into three categories: high, medium, and low in nutrition. Other studies have integrated K-Means with other methods such as Fuzzy C-Means or Neural Networks for personalized diet recommendations. The application of this method has also been used in assessing dietary patterns in various countries, such as Ireland and Korea [10].

In addition to grouping foods based on nutritional content, the cluster analysis approach has also been used in mapping nutritional status [11], predicting food safety based on raw ingredients, and recommending culinary products based on user preferences [12], [13]. However, research that specifically categorizes Indonesian foods and beverages based on their nutritional profiles is still very limited. Previous studies have focused on calorie estimation and classification of modern foods, but systematic clustering of Indonesian traditional foods based on key macronutrient attributes (calories, protein, fat, and carbohydrates) remains scarce. Unlike previous studies that focus on general food classification or modern dietary datasets, this study specifically examines Indonesian traditional foods using a consistent macronutrient-based clustering approach, which remains underexplored in the context of local dietary patterns.

This study addresses this research gap by applying K-Means clustering to Indonesian traditional foods within the context of health, cultural preservation, and sustainable diets [14]. Furthermore, this approach has potential to support national food policies and disease risk mapping based on food consumption [15]. The availability of clustering-based nutrition information systems can also assist users in selecting foods according to individual nutritional needs [16].

This study uses a quantitative approach by applying the K-Means clustering algorithm to group Indonesian foods and beverages based on calorie, protein, fat, and carbohydrate content. The optimal number of clusters is evaluated using the Elbow Method, Davies–Bouldin Index, and Silhouette Score [17]. The dataset used is sourced from the nutritional data of 1,346 Indonesian traditional foods as reported in previous studies [18].

Based on previous research, the use of K-Means has proven effective in modern and international food clustering. Several studies have also demonstrated the integration of K-Means with other methods such as neural networks, fuzzy clustering, and principal component analysis to improve data segmentation accuracy [19]. Accordingly, this study is explicitly positioned as an applied mathematical modeling and data-driven classification study, rather than a methodological advancement of the clustering algorithm itself.

2. RESEARCH METHOD

This study uses a quantitative approach with data mining methods to cluster traditional Indonesian foods and beverages based on their nutritional profiles. This study aims to identify patterns in traditional foods based on macronutrient similarities, providing an initial descriptive nutritional classification.

This study employs the K-Means clustering algorithm to group numerical nutritional data into clusters based on similarity. The determination of the number of clusters is conducted using the Elbow Method as an initial exploratory approach in unsupervised clustering. Although K-Means is known to be sensitive to centroid initialization, this issue was mitigated by applying the K-Means++ initialization method and performing multiple runs to ensure solution stability. To ensure clustering stability, the K-Means algorithm was executed 20 times using different random seed initializations. The inertia values obtained from each run were recorded, and their variation was evaluated using the standard deviation.

The flowchart for this study is shown in Figure 1 below:

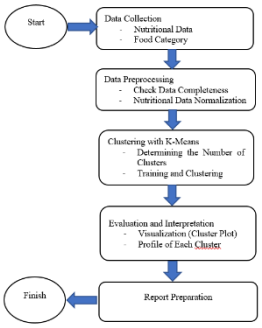

Figure 1. Research Flowchart Using the K-Means Algorithm

2.1 Data Collection

The first step of this study is data collection. The data were obtained from an open-access dataset containing nutritional composition information of typical Indonesian foods and beverages, including energy (calories), protein, fat, and carbohydrate content per 100 grams. The dataset was sourced from a publicly available Kaggle repository (https://www.kaggle.com/datasets/anasfikrihanif/indonesian-food-and-drink-nutrition-dataset), uploaded by Anas Fikri Hanif on November 1, 2023, and consists of 1346 nutritional records of Indonesian foods and beverages. This dataset has also been utilized in previous studies for nutritional clustering and analysis [18], supporting its reliability and relevance for this research. This dataset was obtained from the Tabel Komposisi Pangan Indonesia (Indonesian Food Composition Table) data published by the Ministry of Health of the Republic of Indonesia (www.panganku.org). This dataset has gone through a modification process in the form of data cleaning and adding image links. Since the original data were derived from the official Indonesian Food Composition Table, the dataset can be considered reliable for nutritional analysis purposes. Figure 2 presents a snapshot of the nutritional data obtained from the Kaggle platform.

id	calories	proteins	fat	carbohydr	name	image
1	280	9.2	28.4	0	Abon	https://img-cdn.medkomtek.com/PbrY9X3ignQ8sVuj_LU
2	513	23.7	37	21.3	Abon haru	https://img-global.cpcdn.com/recipes/cbf330fbd1ba631
3	0	0	0.2	0	Agar-agar	https://res.cloudinary.com/dk0z4ums3/image/upload/v
4	45	1.1	0.4	10.8	Akar tonjo	https://images.tokopedia.net/img/cache/200-square/p
5	37	4.4	0.5	3.8	Aletoge se	https://nilaigizi.com/assets/images/produk/produk_157
6	85	0.9	6.5	7.7	Alpukat se	https://katakabar.com/assets/images/upload/news/me
7	96	3.7	0.6	19.1	Ampas kel	https://images.tokopedia.net/img/cache/215-square/sh
8	414	26.6	18.3	41.3	Ampas Tal	https://palpres.disway.id/upload/9e9c1ba592cac7270d
9	75	4.1	2.1	10.7	Ampas tal	https://cdns.diadona.id/diadona.id/resized/640x320/ne
10	67	5	2.1	8.1	Ampas tal	https://cdn-image.hipwee.com/wp-content/uploads/20

Figure 2. Excerpt of Indonesian Food and Beverage Nutrition Data from the Kaggle Website

The original dataset obtained from Kaggle consisted of 1346 records of Indonesian foods and beverages. Since this study employed an exploratory clustering approach, a representative subset was selected to improve computational efficiency while preserving data variability. The minimum sample size was estimated using the Slovin formula with a 5% margin of error:

$$n = \frac{N}{1+N(e)^2} \tag{1}$$

where n is the required sample size, N is the population size (1346 records), and e is the tolerated error level (0.05). Based on this calculation, the minimum recommended sample size was approximately 308 observations. The use of 300 observations was considered practically sufficient, representing 22.3% of the total population. The selected records were obtained through purposive keyword-based filtering to retain traditional Indonesian foods and beverages relevant to the research objectives. This approach is consistent with exploratory data mining studies, where representative subsets are commonly used when the full dataset contains heterogeneous entries.

Next, a data selection process was conducted to identify local Indonesian traditional foods and beverages relevant to the research focus. The selection was performed using keyword-based filtering applied to food and beverage names commonly associated with Indonesian traditional cuisine, including nasi, ayam, ikan, soto, rendang, tempe, tahu, sayur, rawon, bakso, lontong, opor, gado, pecel, gulai, lele, bebek, mie, ketupat, sambal, dodol, jenang, es, jamu, kerupuk, abon, telur, tongkol, terong, jengkol, and petai. This heuristic approach was applied to both solid foods and beverages without explicitly separating physical form or consumption patterns. This keyword-based filtering was adopted as an operational definition of Indonesian traditional foods for exploratory analysis, acknowledging a degree of subjectivity in the classification process. A subset of 300 samples was selected using purposive filtering to balance computational efficiency and data diversity. However, this approach may introduce sampling bias, and future studies should consider stratified sampling to improve representativeness. From the overall dataset, the first 300 records that met these criteria were retained for analysis, as illustrated in Figure 3.

id	calories	proteins	fat	carbohydrate	name	image	name_lower
1	280	9.2	28.4	0	Abon	https://im	abon
2	513	23.7	37	21.3	Abon haruwan	https://im	abon haruwan
8	414	26.6	18.3	41.3	Ampas Tahu	https://pa	ampas tahu
9	75	4.1	2.1	10.7	Ampas tahu kukus	https://cd	ampas tahu kukus
18	126	3.4	7.9	10.3	Anyang sayur	https://st	anyang sayur
22	113	0.9	7.2	11.2	Ares sayur	https://as	ares sayur
29	98	3.9	2.6	14.8	Asinan Bogor sayuran	https://wi	asinan bogor sayuran
36	275	37.4	12.2	1.3	Ayam goreng kalasan paha	https://th	ayam goreng kalasan paha
39	295	39.2	13.6	1	Ayam goreng mbok berek dada	https://im	ayam goreng mbok berek dada
41	246	37.9	9	0.7	Ayam goreng pasundan dada	https://d1	ayam goreng pasundan dada
42	245	33.1	11.4	0.3	Ayam goreng pasundan paha	https://d1	ayam goreng pasundan paha
44	244	36.7	9.2	1	Ayam goreng sukabumi dada	https://i0.	ayam goreng sukabumi dada
45	283	35.7	14.3	0.5	Ayam goreng sukabumi paha	https://i0.	ayam goreng sukabumi paha
46	261	27.4	16.1	1.6	Ayam hati segar	https://as	ayam hati segar
47	264	18.2	20.1	2.7	Ayam taliwang masakan	https://as	ayam taliwang masakan

Figure 3. Excerpt of the First 300 Local Indonesian Traditional Food and Beverage Entries

2.2 Data Preprocessing

The data preprocessing stage ensures that the dataset used in the analysis is of high quality and ready for clustering. The first step is to prepare the numerical data by opening the filtered file containing the first 300 data entries of local Indonesian traditional foods and beverages. The data is then read and displayed to ensure that all data has been stored correctly.

The next step was attribute selection, which was done by removing irrelevant columns, taking only columns containing numerical data needed in the clustering process, namely: calories, proteins, fat, and carbohydrates, as shown in Table 1.

Table 1. Selection of Numeric Data Columns

Selected Column	Explanation
calories	Calories per 100 grams/food or beverage (cal)
proteins	Protein per 100 grams of food or beverage (grams)
fat	Fat per 100 grams/food or beverage (grams)
carbohydrate	Carbohydrates per 100 grams/food or beverage (grams))

The next step is to check and handle missing values, which involves checking whether there are empty values (NaN) in the dataset. If empty values are found, they are handled by filling in the values using the average nutritional value from the relevant column, or by deleting the data row if it is considered unrepresentative.

The final step in preprocessing is data normalization using the MinMaxScaler method. This process is carried out to standardize the scale of each variable so that they have the same value range, ensuring that no attribute dominates the clustering process. Thus, the normalized data is ready for use in the next stage of clustering analysis. MinMaxScaler normalization formula [18]:

$$V_{norm} = \left(\frac{V_i - V_{min}}{V_{max} - V_{min}}\right) \quad (2)$$

where:

V_{norm} = Normalized value
V_{min}, V_{max} = Minimum and maximum absolute values of the column (feature)
V_i = Original value of each entry in the data

2.3 Clustering with K-Means

The clustering process is performed using the K-Means algorithm through the following steps [20], [18]:

a. Determination of the Number of Clusters
The number of clusters, denoted as K, is determined prior to clustering.

b. Initialization of Centroids
K data points are initially selected using the K-Means++ initialization strategy rather than purely random selection. This method distributes the initial centroids more systematically across the data space, reducing the risk of poor local minima and improving clustering stability.

c. Distance Calculation and Cluster Assignment
Each data point is assigned to the nearest centroid by calculating the Euclidean distance, defined as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

where:

n = number of attributes (features)
x_i = value of the i-th attribute of data point x
y_i = value of the i-th attribute of data point y

A data point x is assigned to cluster C_k if the distance to centroid μ_k is minimal:

$$x \in C_k \text{ if } d(x, \mu_k) = \min_{1 \le j \le k} d(x, \mu_j) \quad (4)$$

where:

K = total number of clusters
C_k = the k-th cluster
μ_k = centroid of cluster C_k

d. Centroid Update

After all data points are assigned, each centroid is updated by computing the mean of all data points belonging to the corresponding cluster:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \epsilon C_k} x_i \quad (5)$$

where:

μ_k = centroid of the k-th cluster

C_k = set of data points in cluster k

$|C_k|$ = number of data points in cluster k

These steps are repeated iteratively until convergence is achieved, indicated by no change in cluster assignments or centroids. To further enhance robustness, the clustering process was executed multiple times using different random seeds, and the solution with the lowest inertia value was selected as the final clustering result.

Next, a testing process was carried out to assess the performance and accuracy of the clustering results produced by the K-Means algorithm. One commonly used approach to determine the best number of clusters is the Elbow Method [20]. In this study, the determination of the optimal number of clusters was carried out using the Elbow Method. The Elbow Method was used to determine the optimal number of clusters by calculating the WCSS values for K = 1 to 10. The optimal number of clusters is identified at the point where the reduction in WCSS begins to decrease more slowly, forming an elbow in the curve [21]. It then calculates the WCSS (Within-Cluster Sum of Squares) value for each K value. WCSS is the sum of the squares of the distances between each point and the cluster center in that cluster.

$$WCSS = \sum_{j=1}^{K} \sum_{x_i \epsilon C_j} \|x_i - \mu_j\|^2 \quad (6)$$

where:

K = total number of clusters

C_j = set of data points in cluster j

x_i = the i-th data point in cluster j

μ_j = centroid of cluster j

The optimal number of clusters is identified at the point where the reduction in WCSS begins to diminish significantly, forming an "elbow" in the curve.

Table 2 presents the number of clusters (K) and the corresponding WCSS (inertia) values used to identify the optimal number of clusters in this study.

Table 2. Number of Clusters (K) and Corresponding WCSS (Inertia) Values

Number of Clusters (K)	WCSS value
1	25
2	15
3	10.5
4	8.5
5	7.5
6	6.2
7	5.3
8	4.6
9	3.9
10	3.2

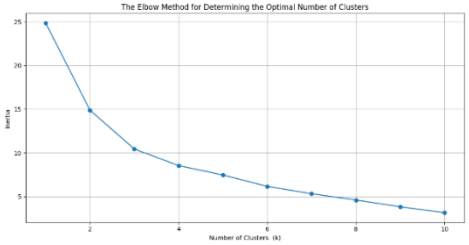

Figure 4. Determination of the Optimal Number of Clusters Using the Elbow Method

The graph in Figure 4 shows the relationship between the number of clusters (k) and the inertia value, which describes how tightly the data is grouped within each cluster. It can be seen that the inertia value decreases sharply from k = 1 to k = 3, then the decrease begins to level off at k = 4. The elbow point on the graph is around k = 3 to k = 4, indicating that the optimal number of clusters for this data is within that range. Thus, using 3 to 4 clusters is considered most efficient because it balances model complexity and the proximity of data within each group.

2.4 Silhouette Coefficient

The silhouette coefficient is a quantitative measure used to evaluate the quality of clustering results by comparing the degree of similarity of a data point to its own cluster and to the nearest neighboring cluster. The silhouette coefficient ranges from -1 to 1, where values close to 1 indicate optimal cluster separation and high intra-cluster cohesion, while values close to -1 suggest potential misclassification [22]. The silhouette coefficient for each observation $(s(i))$ is determined according to the mathematical formulation presented in Equation 6.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{7}$$

The average intra-cluster distance for the i-th observation is denoted by $a(i)$, while the average distance to the nearest inter-cluster is denoted by $b(i)$.

2.5 Davies Bouldin Index (DBI)

The Davies–Bouldin Index (DBI) is a cluster validation metric that evaluates both separation and cohesion. Separation refers to the distance between cluster centroids, while cohesion reflects the similarity of data points within a cluster to their respective centroid. The quality of clustering results is determined by the distribution and proximity of data within each cluster. A lower DBI value (≥ 0) indicates better clustering performance [18].

$$DBI = \frac{1}{\mathrm{k}}\sum_{i=1}^{k} max_{i\neq j}(R_{i,j}) \tag{8}$$

where:

k = number of clusters
$R_{i,j}$ = inter-cluster ratio
max = the maximum inter-cluster ratio is selected.

3. RESULT AND ANALYSIS

The K-Means algorithm with K-Means++ initializationwas implemented through several preprocessing stages and the determination of the optimal number of clusters based on the results of observations on the Elbow graph. The number of clusters used in this implementation was set at 4 clusters (k = 4). This value was chosen to provide more detailed grouping results for the data variation analyzed. The clustering process was conducted with k = 4, assigning each data point to a cluster based on attribute similarity.

To evaluate the quality of the clustering results, the Silhouette Coefficient was calculated for the selected number of clusters (k = 4), yielding an average score of 0.45, which indicates a moderate clustering structure with acceptable separation and intra-cluster cohesion. A more detailed analysis of silhouette scores per cluster revealed variability in cluster quality. Cluster 1 achieved the highest score (0.57), indicating well-defined separation, while Clusters 2 and 3 showed lower scores (0.25 and 0.26), suggesting weaker separation and potential overlap. Cluster 0 exhibited moderate cohesion with a score of 0.35.

To further validate the differences between clusters, a one-way ANOVA test was conducted on all nutritional variables. The results revealed statistically significant differences across clusters for calories ($F = 300.47$, $p < 0.001$), proteins ($F = 323.54$, $p < 0.001$), fat ($F = 43.18$, $p < 0.001$), and carbohydrates ($F = 84.15$, $p < 0.001$), confirming that the clusters are meaningfully differentiated based on their nutritional characteristics.

Clustering stability was assessed through 20 runs with different random seeds, resulting in a mean inertia value of 8.82 and a standard deviation of 0.27, indicating relatively low variation and stable performance. This suggests that the clustering results are not highly sensitive to centroid initialization. In addition, the Davies–Bouldin Index (DBI) yielded a value of 0.94, indicating acceptable cluster separation, where lower values correspond to better clustering quality.

This result supports the suitability of using four clusters for nutritional data analysis, considering that traditional food items naturally exhibit overlapping nutritional characteristics. The results of the food grouping visualization using the K-Means algorithm are shown in Figure 5.

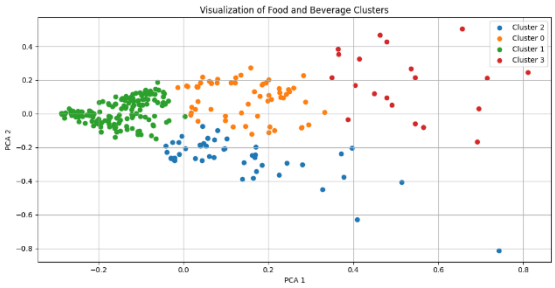

Figure 5. Visualization Results Using the K-Means Algorithm with 4 Clusters

Figure 5 shows the results of visualizing the grouping of traditional Indonesian foods and beverages based on locality using the K-Means algorithm with four clusters (k = 4). Each color represents a cluster formed based on similarities in calorie, protein, fat, and carbohydrate content. The two-dimensional visualization produced through Principal Component Analysis (PCA) shows that each cluster has a fairly separate data distribution, indicating that the clustering process successfully distinguishes food groups based on the similarity of their nutritional profiles. PCA reduces the four-dimensional data into two principal components, explaining 58.34% and 30.43% of the variance, respectively. Together, they retain 88.77% of the total variance. After the clustering process, the results are as shown in Tables 3 and 4 below.

Table 3. Amount of Food Data per Cluster

Cluster	Amount of Data
0	62
1	170
2	47
3	21

Table 4. Excerpts from the Clustering Results of Local Indonesian Traditional Foods and Beverages

Cluster	Foods and Beverages Data
0	Ayam Goreng Kalasan Paha, Ayam Goreng Mbok Berek Dada, Ayam Taliwang Masakan, Ikan Bandeng Presto Masakan, Ikan Cakalang Asap Mentah, Ikan Gabus Asap Mentah, Ikan Peda Banjar Mentah, Ikan Pindang Banjar Mentah, Ikan Pepetek, Ikan Sepat Kering, Ikan Teri Nasi Kering, Ikan Telur Asin Mentah, Peda Ikan Banjar, Pindang Ikan Layang, Pindang Ikan Selar Kecil, Kerupuk Udang, Ampas Tahu, Kacang Belimbing Tempe, Rendang Sapi Masakan, Sapi Abon.
1	Ayam Dideh, Anyang Sayur, Ares Sayur, Asinan Sayur, Gado-Gado, Gudeg Sayur, Gulai Kambing, Gulai Pakis, Gulai Pliek, Gulai Tiram, Ikan Bandeng, Ikan Gabus, Ikan Patin Bakar, Ikan Pepes Mas, Ikan Mujair Pepes, Ikan Kembung, Ikan Tongkol, Ikan Lele Bakar, Ikan Layur, Ikan Sepat Kering, Ikan Teri, Jamur Tiram, Jengkol Segar, Kalio Telur, Karedok Sayur, Ketupat Tahu, Kerupuk Sayong, Lamtoro Tempe, Mie Aceh Rebus, Mie Ayam, Mie Bakso, Nasi Gurih, Nasi Gemuk, Nasi Minyak, Nasi Rames, Nasi Tim, Nasi Goreng, Oncom Pepes, Oncom Kacang Tanah, Oncom Merah Tumis, Pempek, Petis Ikan, Pindang Ikan, Pundut Nasi, Rawon Masakan, Rendang Sapi, Sayur Asem, Sayur Bunga Pepaya, Sayur Kohu-Kohu, Sayur Lilin, Sayur Sop, Sayur Umbut Kelapa, Semur Jengkol, Soto Bandung, Soto Banjar, Soto Betawi, Soto Kudus, Soto Madura, Soto Padang, Tempe Bongkrek, Tahu Isi, Tahu Semur, Ketupat Ketan, Ampas Tahu Kukus, Koro Tempe Benguk, Ikan Baung Bakar, Ikan Belida Bakar, Ikan Papuyu Bakar, Ikan Lais Bakar, Ikan Nasu Metti.
2	Abon, Dodol, Dodol Galamai, Dodol Kedondong, Emping (Kerupuk Melinjo), Ikan Asin Goreng, Jenang, Kacang Telur, Keripik Tempe, Kerupuk Kemplang Goreng, Kerupuk Melinjo Goreng, Kerupuk Udang Goreng, Nasi Jagung, Nasi Uduk, Sagu Rendang.
3	Ayam Usus Goreng, Ikan Asin Gabus Goreng, Ikan Asin Teri Goreng, Ikan Gabus Kering, Ikan Kayu Kering, Ikan Mujair Goreng, Ikan Sale Lais Mentah, Ikan Teri Kering Tawar, Kerupuk Kulit Kerbau, Kerupuk Urat.

After the grouping of food data using the K-Means algorithm was completed, the next step was to interpret the characteristics or features of each cluster. The objective of this stage was to identify differences in nutritional content patterns among the formed clusters of traditional foods and beverages. This interpretation was conducted by calculating the average value of each nutritional variable (calories, protein, fat, and carbohydrates) in each cluster. These average values represent the typical nutritional profile of each cluster. The formula used to calculate the average nutritional value for each cluster is expressed as follows:

$$\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_{ij} \quad (9)$$

where:

$\bar{x}_{kj}$ = the average value of nutritional variable j in cluster k
C_k = set of data points in cluster k
$|C_k|$ = number of data points in cluster k
x_{ij} = value of nutritional variable j for data point x_i in cluster k

After obtaining the average values for each nutritional variable, the characteristics of each cluster were identified, as presented in Table 5 and Figure 6.

Table 5. Average Nutritional Values and Characteristics of Each Cluster

Cluster	Calories (cal)	Protein (grams)	Fat (grams)	Carbohydrates (grams)	General Characteristics of Clusters
0	243.9032258	29.32903226	11.61774194	6.090322581	High protein, moderate calories, low carbohydrates
1	104.5158824	9.425294118	3.406470588	9.708235294	Low calorie, low fat, moderate carbohydrates
2	406.3404255	7.804255319	17.51489362	63.73191489	High in carbohydrates, high in calories, low in protein
3	444.047619	54.62857143	21.55714286	7.614285714	Very high in protein and calories, fairly high in fat

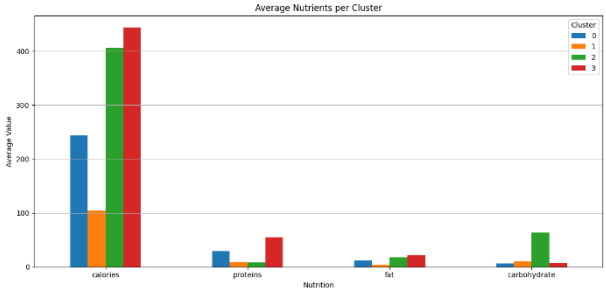

Figure 6. Graph of Average Nutrient Values per Cluster

The graph in Figure 6 illustrates the average nutritional content (calories, protein, fat, and carbohydrates) for each cluster formed using the K-Means algorithm. Each cluster demonstrates distinct nutritional characteristics. Based on the Elbow Method and the Silhouette Score evaluation (Silhouette Score = 0.45), the selection of four clusters (k = 4) provides a reasonable clustering structure for the analyzed nutritional data. The moderate Silhouette Score (0.45) suggests partial overlap between clusters, which is expected due to the inherent similarity of traditional foods sharing common ingredients such as rice, coconut milk, and spices.

Within the clustering results, Cluster 3 exhibits the highest average calorie and protein values, followed by Cluster 2, while Cluster 1 shows the lowest values across most nutritional variables, representing low-energy food groups. Cluster 2 is characterized by high carbohydrate content, indicating carbohydrate-dominant food items. Overall, these findings demonstrate that the K-Means algorithm effectively groups traditional Indonesian foods based on dominant nutritional profiles, with each cluster reflecting a distinct pattern ranging from low to high energy and protein content. The clustering results reflect not only nutritional composition but also Indonesian culinary practices. For example, Cluster 3 (high protein and calorie) is dominated by fried and preserved animal-based foods, which are commonly prepared using oil-intensive methods. In contrast, Cluster 1 includes vegetable-based dishes and soups, which are typically lower in energy due to boiling or steaming processes.

Variations in cluster assignment for similar food items (e.g., "Rendang Sapi") may result from differences in preparation methods, ingredient proportions, and fat content, highlighting the importance of recipe-level standardization in nutritional datasets. Compared to previous studies [9], which categorized foods into broad nutritional levels, this study provides a more granular classification by identifying four distinct clusters, allowing more specific differentiation between protein-rich and carbohydrate-dominant food groups.

From a policy perspective, the identified clusters can support:
a. Development of food-based dietary guidelines by categorizing traditional foods into nutritional risk groups.
b. Public health interventions targeting obesity and diabetes by limiting high-calorie cluster consumption.
c. Nutritional labeling systems for traditional foods based on cluster classification.
d. Integration into digital nutrition recommendation systems for personalized diets.

This study has several limitations:
a. Only four macronutrient variables were used, excluding micronutrients such as vitamins and minerals.
b. The dataset does not account for portion size variability.
c. The use of a 300-sample subset may limit generalizability.
d. Variability in food preparation methods is not standardized.

4. CONCLUSION

The results demonstrate that the K-Means algorithm effectively clusters Indonesian traditional foods into four nutritionally distinct groups. Cluster 3 contains foods with the highest calorie and protein content, Cluster 2 is dominated by carbohydrate-rich foods, Cluster 1 represents low-energy foods, and Cluster 0 shows moderate nutritional values.

Beyond methodological contribution, the findings offer practical value for public health by enabling targeted dietary recommendations and supporting nutrition-sensitive policy design. However, the moderate clustering quality and limited feature set indicate that the results should be interpreted as exploratory rather than definitive. Future research should incorporate micronutrient data, larger datasets, and comparative clustering methods to improve robustness and policy relevance.

5. REFERENCES

[1] S. Arifah, E. R. Swedia, and M. R. D. Septian, "Analisis Perbandingan Algoritma Clustering dalam Melakukan Segmentasi Warna pada Citra Jajan Tradisional," *Sebatik*, vol. 27, no. 1, pp. 70–76, 2023, doi: 10.46984/sebatik.v27i1.2273.

[2] D. Tsolakidis, L. P. Gymnopoulos, and K. Dimitropoulos, "Artificial Intelligence and Machine Learning Technologies for Personalized Nutrition: A Review," *Informatics*, vol. 11, no. 3, p. 62, 2024, doi: 10.3390/informatics11030062.

[3] P.-M. Lu and Z. Zhang, "The Model of Food Nutrition Feature Modeling and Personalized Diet Recommendation Based on the Integration of Neural Networks and K-Means Clustering," *J. Comput. Biol. Med.*, vol. 5, no. 1, 2025, doi: 10.71070/jcbm.v5i1.60.

[4] M. M. Ridzki, I. Hadijah, M. Mukidin, A. Azzahra, and A. Nurjanah, "K-Means Algorithm Method for Clustering Best-Selling Product Data at XYZ Grocery Stores," *Int. J. Soc. Serv. Res.*, vol. 3, no. 12, pp. 3354–3367, 2023, doi: 10.46799/ijssr.v3i12.652.

[5] C. O'Hara, A. O'Sullivan, and E. R. Gibney, "A Clustering Approach to Meal-Based Analysis of Dietary Intakes Applied to Population and Individual Data," *Journal of Nutrition*, vol. 152, no. 10. pp. 2297–2308, 2022. doi: 10.1093/jn/nxac151.

[6] Y. D. Pradvenanta and R. Prathivi, "Implementasi K-Means Untuk Pengelompokan Makanan Cepat Saji Bagi Penderita Penyakit Obesitas," vol. 6, no. 1, 2024, doi: 10.47065/bits.v6i1.5279.

[7] R. Gestavito, A. Id Hadiana, and F. Rakhmat Umbara, "Pengelompokan Tingkat Risiko Penyakit Diabetes Melitus Menggunakan Algoritma K-Means Clustering," *J.Masy.Inform.Unjani*, vol.8, no.1, pp.16–35, 2024.

[8] M. Wu and H. Jiang, "Research on Food Safety Prediction Method Based on K-Means Clustering Algorithm," *MATEC Web Conf.*, vol. 355, p. 03018, 2022, doi: 10.1051/matecconf/202235503018.

[9] Fito Mardianto, Faishal Nugraha, Federico Anggito Ryseno, Yoga Prasetyo Wibowo, and Eka Kusuma Pratama, "Nutritional Analysis Of Traditional Indonesian Food Using Machine Learning," *J. Artif. Intell. Eng. Appl.*, vol. 3, no. 3, pp. 644–649, 2024, doi: 10.59934/jaiea.v3i3.476.

[10] J. Kim *et al.*, "Dietary Patterns Derived by Cluster Analysis are Associated with Cognitive Function among Korean Older Adults," *Nutrients*, vol. 7, no. 6. pp. 4154–4169, 2015. doi: 10.3390/nu7064154.

[11] S. S. Nagari and L. Inayati, "Implementation of Clustering Using K-Means Method To Determine Nutritional Status," *Jurnal Biometrika dan Kependudukan*, vol. 9, no. 1. pp. 62–68, 2020. doi: 10.20473/jbk.v9i1.2020.62-68.

[12] A. K. Wardhani, C. E. Widodo, and J. E. Suseno, "Information System for Culinary Product Selection Using Clustering K-Means and Weighted Product Method," vol. 165, no. ICCSR, pp. 18–22, 2018, doi: 10.2991/iccsr-18.2018.5.

[13] Q. Thames *et al.*, "Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8899–8907, 2021, doi: 10.1109/CVPR46437.2021.00879.

[14] S. Biesbroek *et al.*, "Toward Healthy and Sustainable Diets for The 21st Century: Importance of Sociocultural and Economic Considerations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 26, pp. 1–9, 2023, doi: 10.1073/pnas.2219272120.

[15] J. Zhao *et al.*, "A Review of Statistical Methods for Dietary Pattern Analysis," *Nutr. J.*, vol. 20, no. 1, pp. 1–18, 2021, doi: 10.1186/s12937-021-00692-7.

[16] W. Min, C. Liu, L. Xu, and S. Jiang, "Applications of Knowledge Graphs for Food Science and Industry," *Patterns*, vol. 3, no. 5, p. 100484, 2022, doi: 10.1016/j.patter.2022.100484.

[17] A. Rykov, R. C. De Amorim, V. Makarenkov, and B. Mirkin, "Inertia-Based Indices to Determine the Number of Clusters in K-Means: An Experimental Evaluation," *IEEE Access*, vol. 12, pp. 11761–11773, 2024, doi: 10.1109/ACCESS.2024.3350791.

[18] P. Alga Vredizon, H. Firmansyah, N. Shafira Salsabila, and W. Eko Nugroho, "Penerapan Algoritma K-Means Untuk Mengelompokkan Makanan Berdasarkan Nilai Nutrisi," *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 2. pp. 108–115, 2024. doi: 10.37802/joti.v5i2.577.

[19] A. Maugeri *et al.*, "The Application of Clustering on Principal Components for Nutritional Epidemiology: A Workflow to Derive Dietary Patterns," *Nutrients*, vol. 15, no. 1, 2023, doi: 10.3390/nu15010195.

[20] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis based on k-means," vol. 18, no. 1, pp. 470–477, 2020, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[21] O. D.I., Durojaye; G.N., "Analysis and Visualization Market Segementation in Banking Sector Using KMeans Machine Learning Algorithm," *FUDMA J. Sci.*, vol. 6, no. 1, pp. 387–393, 2022, doi: https://doi.org/10.33003/fjs-2022-0601-910 8 FJS.

[22] F. Nurulhikmah and D. N. E. Abdi, "Classification of Foods Based on Nutritional Content Using K-Means and DBSCAN Clustering Methods," *TEKNIKA*, vol. 13, no. 3, pp. 481–486, 2024, doi: 10.34148/teknika.v13i3.1067.