



Comparison of single, complete, and average linkage algorithms in clustering instant noodle products based on nutritional content

Fitri Novita Sari*, Arief Agoestanto

Department of Mathematics, Faculty of Mathematics and Natural Science,
Universitas Negeri Semarang, Semarang, 50229, Indonesia

Article Info

Article History:

Received 18-06-2025

Accepted 20-04-2026

Published 30-04-2026

Keywords:

Instant noodles, hierarchic agglomerative clustering, nutritional content, linkage algorithm, silhouette coefficient

Abstract

Instant noodles are consumed on a very large scale in Indonesia, the second largest consumer in the world, while many products are high in sodium, fat, and sugar, so that consumers need a simple way to see which products have a heavier nutritional profile. This study clusters 138 instant noodle products sold in Semarang by five label nutrients (fat, protein, carbohydrate, sugar, and sodium), standardized by z-scores, using Hierarchical Agglomerative Clustering with Euclidean distance and three linkage algorithms: single, complete, and average. The number of clusters from 2 to 10 was evaluated with the silhouette coefficient in Python. Complete linkage was the best with a silhouette score of 0.4445 at two clusters of 98 and 40 products, followed closely by average linkage (0.4365), while single linkage was clearly lower (0.3567) and stayed below 0.10 for four or more clusters because of chaining. The second cluster has higher means for all five nutrients: 32% more fat, 54% more protein, 59% more carbohydrate, 83% more sugar, and 47% more sodium, with a mean sodium of 1,630 mg against 1,109 mg in the first cluster. Because the silhouette score indicates a weak to moderate structure, the clusters should be read as a gradient of nutritional load. Consumers are advised to prefer products of the lighter cluster and to check sodium on the label, and producers to reduce sodium.

How to cite:

Sari, F. N., & Agoestanto, A. (2026). Comparison of single, complete, and average linkage algorithms in clustering instant noodle products based on nutritional content. *UNNES Journal of Mathematics*. 15(1): 25-33

1. Introduction

Machine learning is a branch of science at the intersection of statistics, artificial intelligence, and computer science that focuses on developing algorithms to discover patterns in data and make automatic predictions (Giordani et al., 2020). In general, machine learning is divided into two main types, supervised learning and unsupervised learning. One of the widely used techniques in unsupervised learning is clustering, a method for grouping data based on similarities in characteristics. Clustering has been applied across various fields such as medicine, marketing, economics, and social sciences due to its ability to uncover hidden patterns in data (Giordani, Ferraro, & Martella, 2020).

A popular approach in clustering is Hierarchical Agglomerative Clustering (HAC), which organizes data into a hierarchical structure based on the degree of similarity between objects without requiring the number of clusters to be specified in advance (Muhima et al., 2021). HAC operates in a bottom-up manner, starting with each data point as an individual cluster and gradually merging them based on inter-point distances. The final result is visualized in a dendrogram, providing a comprehensive overview of the cluster structure (Kassambara, 2017). Compared to methods like k-means or DBSCAN, HAC is considered more flexible as it does not require predefining the number of clusters and can identify natural structures in the data (Murtagh & Contreras, 2017).

In practice, HAC includes several linkage algorithms that determine how clusters are merged, including single linkage, complete linkage, and average linkage (Kusumastuti et al., 2022). Each method has its own strengths and weaknesses in forming cluster structures. Therefore, it is important to evaluate the performance of each linkage method to determine the most suitable algorithm for a given dataset.

On the other hand, Indonesia is one of the countries with the highest consumption of instant noodles. According to the World Instant Noodles Association, Indonesia ranked second globally with 14.54 billion servings in 2023 (Fahira et al., 2024). The popularity of instant noodles is driven by their affordability, wide availability, and ease of preparation. However, many products contain high levels of sodium, fat, and sugar (Wibowo et al., 2024), and an unbalanced nutritional intake over a long period may increase the risk of non-communicable diseases (Cahyani et al., 2024). Even in Thailand, the study of Chailek et al. (2026) showed that most instant foods remain above the sodium benchmarks of the World Health Organization despite a modest decline in sodium content.

The concrete problem addressed here is that consumers face many instant noodle products whose labels list five nutrients, but there is no simple way to see which products have a similar and heavier nutritional profile. The nutrients are on different scales, and sodium, which is given in milligrams and is the nutrient of most concern, dominates the raw values. The urgency is that instant noodles are consumed on a very large scale and a single pack can provide a large part of the recommended maximum sodium intake of 2,000 mg per day (World Health Organization, 2023), so that a data-driven grouping of products can help consumers and regulators to identify the products of concern.

Hierarchical Agglomerative Clustering was used because the number of groups of products is not known in advance and because the dendrogram shows the structure of the products at every level of similarity (Murtagh & Contreras, 2017). Three linkage algorithms were compared because the linkage determines the shape of the clusters: single linkage merges by the nearest members and may chain products into one long cluster, complete linkage merges by the farthest members and gives compact clusters, and average linkage is a compromise between them (Murtagh & Contreras, 2017). The silhouette coefficient was used because there are no true labels and because it can be computed for each number of clusters (Shahapure & Nicholas, 2020), and the data were standardized by z-scores so that sodium does not dominate the distance.

Previous studies applied hierarchical clustering to food data. Nurhayati et al. (2023) grouped noodles by nutritional similarity with hierarchical cluster analysis, Dalimunthe and Hanafiah (2021) used agglomerative clustering with average linkage on the nutrient content of foods, and Yusniyanti et al. (2021) compared average linkage with K-means for provinces. These studies used one linkage or compared a hierarchical method with a partitioning method. To the authors' knowledge, few studies compare single, complete, and average linkage on instant noodle products using several numbers of clusters.

The novelty of this study is therefore threefold: (1) the comparison of three linkage algorithms for 138 instant noodle products sold in Semarang, evaluated for 2 to 10 clusters and not only for the number of clusters chosen in advance; (2) the diagnosis of the behavior of each linkage, including the chaining of single linkage; and (3) the interpretation of the clusters in terms of the share of the recommended sodium intake that a product provides.

This study aims to cluster instant noodle products by their nutritional content with Hierarchical Agglomerative Clustering and to compare single, complete, and average linkage. The research questions are: (1) which linkage algorithm gives the best cluster quality according to the silhouette coefficient; (2) how many clusters are optimal; and (3) what are the nutritional characteristics of the clusters, and what do they imply for consumers and producers?

2. Method

2.1 Data Collection and Sources

This study utilized data from 138 instant noodle products sold in three local markets in Semarang City, Goori Supermarket (Gunungpati), Indomaret (Sekaran), and SuperIndo (Ungaran). Nutritional information was obtained directly from the product packaging labels, covering carbohydrate, protein, fat, sugar, and sodium content. The data were recorded and processed using Microsoft Excel to facilitate further analysis.

2.2 Tools and Analysis Platform

Data processing and analysis were carried out using the Python programming language through the Google Colaboratory platform. Several libraries were employed, including *pandas*, *scikit-learn*, *scipy.cluster.hierarchy*, and *matplotlib* for data manipulation, clustering, and visualization purposes.

2.3 Preprocessing and Standardization

Prior to analysis, feature selection was performed to filter relevant variables, specifically the five main nutritional components. The data were then standardized using the Z-score method to equalize the scale among variables. The standardization formula used is as follows (Giordani et al., 2020):

$$z = \frac{x - \mu_x}{\sigma_x} \quad (2.3)$$

Where:

- z : Z-score standardized value
- x : Original data value
- μ_x : Mean of x
- σ_x : Standard deviation of x

2.4 Clustering Method

The clustering process was conducted using the Hierarchical Agglomerative Clustering (HAC) method with Euclidean distance as the distance metric. Three types of linkage methods were applied and compared, where $D(\cdot, \cdot)$ denotes the distance between two clusters and C and C' are two non-empty clusters (Murtagh & Contreras, 2017):

- a) Single linkage (based on the minimum distance between members of two clusters)

$$D(C, C') = \min_{x \in C, y \in C'} d(x, y) \quad (2.4.1)$$

- b) Complete linkage (based on the maximum distance between members of two clusters)

$$D(C, C') = \max_{x \in C, y \in C'} d(x, y) \quad (2.4.2)$$

- c) Average linkage (based on the average distance between members of two clusters)

$$D(C, C') = \frac{1}{|C||C'|} \sum_{x \in C, y \in C'} d(x, y) \quad (2.4.3)$$

The Euclidean distance between two data points x and y in n -dimensional space is calculated using the following formula (Giordani et al., 2020):

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (2.4.4)$$

The clustering results were visualized using a dendrogram to illustrate the hierarchical structure of the groupings.

2.5 Cluster Validation

To assess clustering quality, the Silhouette Coefficient (SC) was used. This metric evaluates how well each data point lies within its assigned cluster compared to other clusters. The SC formula is defined as (Shahapure & Nicholas, 2020):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2.5)$$

Where:

- $a(i)$: The average distance from data point i to all other points in the same cluster
- $b(i)$: The average distance from data point i to all points in the nearest different cluster.

The linkage method with the silhouette score closest to 1 is considered the best-performing clustering approach in this study.

2.6 Result Interpretation

The clustering results were interpreted by analyzing the average nutritional characteristics carbohydrates, protein, fat, sugar, and sodium of each cluster. This interpretation aimed to identify patterns in nutrient distribution within clusters, thus providing insights into the typical characteristics of products grouped together, without explicitly labeling them into specific categories.

For interpretation, silhouette values above about 0.70 are usually read as a strong cluster structure, values from about 0.50 to 0.70 as reasonable, values from about 0.25 to 0.50 as weak, and values below about 0.25 as no substantial structure. Results and discussions

2.1. Data

The collected data were imported into the google colab environment using the pandas library to read an excel file containing 138 instant noodle product entries. The data were stored in a file named "DATA BARU COBA.xlsx" and assigned to a variable called "data". Once the data were successfully imported, the information to be analyzed was displayed in the following table:

Table 1. Data

	product	fat	protein	carbohydrate	sugar	sodium
0	mie sedap tasty beef yakiniku	23.0	13	69	5	1120
1	mie sedap tasty bakmi ayam daging ayam asli	22.0	16	84	6	820
2	mie sedap tasty chicken teriyaki	24.0	15	72	7	1010
...	...	...	...	...	...	...
136	samyang buldak 2x spicy hot chicken flavor ramen	12.0	12	98	12	1880
137	samyang buldak hot chicken flavor ramen	16.0	14	94	10	1900

3.2 Data Preprocessing

This stage begins with the display of descriptive analysis results. The analysis showed that there were no missing values in the dataset. Next, relevant features for the clustering process were selected by extracting the corresponding columns as a numerical array from the original dataset.

[[2.30e+01, 1.30e+01, 6.90e+01, 5.00e+00, 1.12e+03],

[2.20e+01, 1.60e+01, 8.40e+01, 6.00e+00, 8.20e+02],

.....

[1.20e+01, 1.20e+01, 9.80e+01, 1.20e+01, 1.88e+03],

[1.60e+01, 1.40e+01, 9.40e+01, 1.00e+01, 1.90e+03]]

The output of this extraction revealed that the variables had different scales, particularly because the sodium content was measured in milligrams (mg), unlike the others. Therefore, data standardization was performed to ensure that all variables had a comparable scale during the clustering process.

3.3 Data Standardization

The standardization process was carried out using formula (2.3), where the mean (μ_x) of the fat variable is 16.452899 and the standard deviation (σ_x) is 4.481465. The standardized value (z-score) for the first data point of the fat variable, with an original value of $x = 23.0$, is calculated as follows:

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{23 - 16.452899}{4.481465} = 1.4609287$$

This value matches the result obtained from the computation in google colab, as shown in the following standardized array output:

[[1.46092872e+00, 1.73640429e+00, 6.59294883e-01, 1.84223778e-02, -3.28269738e-01],

[1.23778742e+00, 2.96105740e+00, 1.58141678e+00, 3.81606398e-01, -1.03025151e+00],

....

[-9.93625567e-01, 1.32818658e+00, 2.44206388e+00, 2.56071052e+00, 1.45008409e+00],

[-1.01060371e-01, 2.14462199e+00, 2.19616471e+00, 1.83434248e+00, 1.49688287e+00]]

3.4 Euclidean Distance

In this study, distance calculations were performed automatically using the `pdist` function from the `scipy.spatial.distance` library. Additionally, a manual Euclidean distance calculation was conducted between the first two data points using formula (2.4.4):

$$\begin{aligned} d1 &= \sqrt{(1.46092872 - 1.23778742)^2 + \dots + (-0.328269738 - (-1.03025151))^2} \\ &= \sqrt{3.024554} \\ &= 1.739125 \end{aligned}$$

This manual result matches the output produced by the Python implementation. The Euclidean distance values between data points are displayed in Table 2 below:

Table 2. Euclidean Distance

	0	1	...	136	137
0	0.000000	1.739125	...	4.358384	3.405462
1	1.739125	0.000000	...	5.6828	3.96928
...	...	...	...	...	...
136	4.358384	4.391668	...	0.000000	1.433006
137	3.405462	3.366578	...	1.433006	0.000000

3.5 Linkage Algorithm

In this study, the clustering process using the single linkage, complete linkage, and average linkage algorithms was conducted with the assistance of the SciPy library. The result of each method is represented in the form of a dendrogram, as shown below for each linkage method:

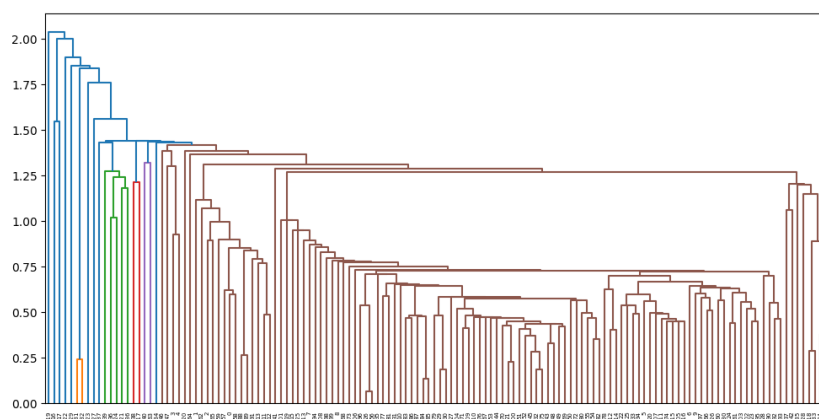


Figure 1. Single Linkage Dendrogram

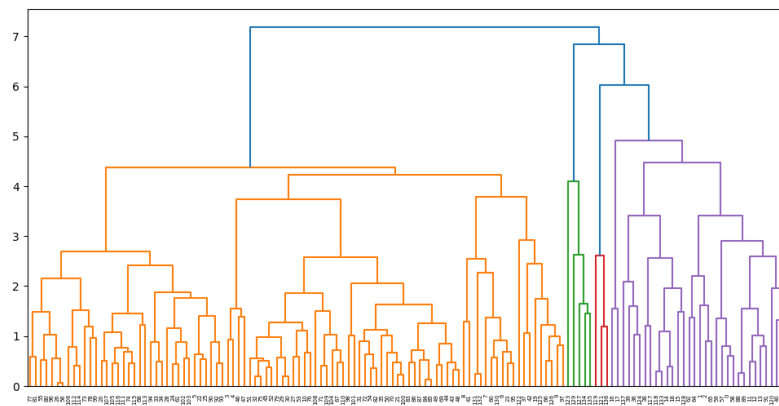


Figure 2. Complete Linkage dendrogram

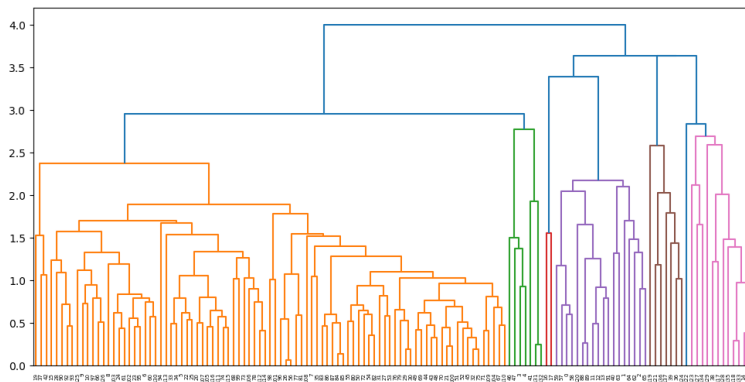


Figure 3. Average Linkage Dendrogram

3.6 Silhouette Score

Performance evaluation is essential to assess and compare the effectiveness of the three clustering algorithms used in this study, single linkage, complete linkage, and average linkage in grouping instant noodle products based on their nutritional content. One of the evaluation methods employed in this research is the Silhouette Coefficient:

Table 3. Silhouette Score

Cluster	Single Linkage	Complete Linkage	Average Linkage
2	0.3567	0.4445	0.4365
3	0.2791	0.4307	0.4286
4	0.0790	0.4139	0.4316
5	0.0636	0.4033	0.4289
6	0.0456	0.4016	0.3716
7	0.0536	0.1865	0.3354
8	0.0610	0.2230	0.3213
9	0.0453	0.2154	0.3224
10	0.0997	0.2157	0.3186

3.7 Result Interpretation

Based on the results presented in Table 3, the complete linkage algorithm demonstrated the most effective performance with a silhouette score of 0.445 and produced two optimal clusters. Therefore, the subsequent analysis in

this study focuses on this algorithm. Figure 2 illustrates the resulting dendrogram, where Cluster 1 consists of 98 instant noodle products, and Cluster 2 contains 40 products.

The two clusters formed exhibit distinct characteristics. These differences can be observed from the mean of each nutritional variable within each cluster, as shown in the following table:

Table 4. Mean of Variable Cluster

cluster	fat	protein	carbohydrate	sugar	sodium
1	15.061224	7.561224	49.806122	3.989796	1109.387755
2	19.862500	11.650000	79.025000	7.300000	1630.000000

Based on the analysis of average nutritional content across clusters reveals that products in cluster 1 contain lower levels of fat, protein, carbohydrates, salt, and sugar compared to those in Cluster 2. In contrast, Cluster 2 encompasses products with relatively higher levels of these nutritional components, indicating a distinct nutritional profile between the two clusters.

Table 3 shows that all three linkages have their highest silhouette at two clusters, and that the best value of complete linkage (0.4445) is only 0.008 higher than that of average linkage (0.4365), which is a difference of 1.8%, so these two linkages are practically equal at two clusters. Single linkage is clearly lower (0.3567 at two clusters) and falls to between 0.045 and 0.10 from four clusters onward, which is the typical sign of chaining: it separates a few outlying products and puts the rest into one large cluster. Complete linkage keeps values between 0.40 and 0.44 up to six clusters and drops to 0.19 at seven clusters, and average linkage keeps values of about 0.43 up to five clusters, so both support a small number of clusters. Because the best silhouette is about 0.44, the structure is weak to moderate, and the products should be seen as a gradient from lighter to heavier products rather than as sharply separated groups.

The cluster means in Table 4 show that Cluster 2 is higher than Cluster 1 in all five nutrients: 32% for fat (19.86 against 15.06 g), 54% for protein (11.65 against 7.56 g), 59% for carbohydrate (79.03 against 49.81 g), 83% for sugar (7.30 against 3.99 g), and 47% for sodium (1,630 against 1,109 mg). The weighted mean of the two clusters, for example 16.45 g of fat and 1,260 mg of sodium, agrees with the mean of the standardization step. Against the recommended maximum of 2,000 mg of sodium per day (World Health Organization, 2023), the mean product of Cluster 1 provides about 55% and that of Cluster 2 about 82%, so one pack of the heavier cluster provides most of the daily limit. Because all five nutrients increase together, the two clusters may mainly separate products by their overall size or portion, if the label values refer to a package; this should be checked by normalizing the values per 100 g.

3.8 Answers to the Research Questions

The research questions are answered by the results as follows. (1) Complete linkage gives the best cluster quality, with a silhouette score of 0.4445, followed by average linkage (0.4365) and single linkage (0.3567). (2) The optimal number of clusters is two, for all three linkages, with 98 products in Cluster 1 and 40 in Cluster 2 for complete linkage. (3) Cluster 2 has higher fat, protein, carbohydrate, sugar, and sodium than Cluster 1, with a mean sodium of 1,630 mg against 1,109 mg.

3.9 Discussion

The results show that the choice of linkage matters mainly through the failure of single linkage, and not through a large difference between complete and average linkage. This agrees with the expectation that single linkage is sensitive to chaining, and it means that, for this kind of data, either complete or average linkage can be used, with complete linkage giving the highest silhouette and compact clusters. The average linkage result is consistent with Dalimunthe and Hanafiah (2021), who obtained usable groups of foods by nutrient content with average linkage, and with Nurhayati et al. (2023), who grouped noodles by nutritional similarity with hierarchical clustering. Yusniyanti et al. (2021) compared average linkage with K-means for provinces; the present study complements them by comparing three linkages for food products over several numbers of clusters. Because the data and the validity indices differ among the studies, the comparison is limited to the direction of the findings.

The novelty of the findings is that the quality of the clustering of instant noodles depends strongly on the linkage only through single linkage, that all three linkages agree on two clusters, and that the two clusters differ by 32% to 83% in the five nutrients, with a mean sodium of 82% of the daily limit in the heavier cluster. This is consistent with the observation of Chailek et al. (2026) that instant foods often remain above the sodium benchmarks of the World Health Organization.

These findings lead to the following recommendations. (1) For consumers: prefer products of the lighter cluster, check the sodium value on the label, and limit the frequency of consumption of products of the heavier cluster, because one pack provides on average about 82% of the daily sodium limit. (2) For producers and regulators: reduce the sodium content of

the products of the heavier cluster and state the values per 100 g and per pack on the label. (3) For researchers: use complete or average linkage and avoid single linkage for such data, report the cophenetic correlation, and repeat the analysis with values per 100 g.

Several limitations should be noted. The silhouette score of about 0.44 indicates a weak to moderate structure; the differences between complete and average linkage are small; only 138 products from three stores and five nutrients were analyzed; the label values may refer to the package and not to a fixed weight, so that the clusters may reflect product size; and no other linkage (for example Ward) or validity index was tested. Future research should add other variables such as energy, fiber, and vitamins, other linkages, stability checks, and values per 100 g.

3. Conclusion

The three research questions are answered as follows. First, complete linkage gives the best cluster quality (silhouette score of 0.4445), followed by average linkage (0.4365), while single linkage is clearly lower (0.3567) and collapses for four or more clusters because of chaining. Second, the optimal number of clusters is two for all three linkages, with 98 and 40 products in complete linkage. Third, the second cluster contains products with higher fat, protein, carbohydrate, sugar, and sodium (a mean sodium of 1,630 mg, about 82% of the recommended daily maximum) than the first cluster (1,109 mg, about 55%), although the silhouette score indicates a weak to moderate structure.

Consumers are advised to prefer products of the lighter cluster and to check the sodium value on the label, and producers and regulators to reduce sodium in the heavier cluster. Future research is recommended to use values per 100 g, to add variables such as total energy, dietary fiber, and vitamins, and to test other linkages and validity measures.

References

- Cahyani, T. D., Puspitasari, D. I., & Sabini, D. (2024). Hubungan asupan zat gizi makro dan frekuensi makan dengan status gizi pada mahasiswa Fakultas Ilmu Kesehatan Universitas Muhammadiyah Surakarta. *Ghidza: Jurnal Gizi dan Kesehatan*, 8(2), 250–257. <https://doi.org/10.22487/ghidza.v8i2.1677>
- Chailek, C., Pavadhgul, P., Kittipichai, W., & Kantachuvesiri, S. (2026). Trends in sodium content of instant foods in Thailand, 2018–2021: Progress and remaining challenges. *Preventing Chronic Disease*, 23, Article 250382.
- Dalimunthe, S., & Hanafiah, A. (2021). Implementation of agglomerative hierarchical clustering based on the classification of food ingredients content of nutritional substances. *IT Journal Research and Development*, 6(1), 60–69. <https://doi.org/10.25299/itjrd.2021.6872>
- Fahira, D., Putri, D. C., Taufhan, L. S., & Nasution, Y. S. J. (2024). Analisis faktor-faktor penyebab konsumsi mie instan di kalangan mahasiswa UINSU Prodi Akuntansi Syariah tahun 2023/2024. *Jurnal Ekonomi Bisnis, Manajemen dan Akuntansi (JEBMAK)*, 3(2), 391–404.
- Giordani, P., Ferraro, M. B., & Martella, F. (2020). *An introduction to clustering with R*. Springer.
- Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA.
- Kusumastuti, R., Bayunanda, E., Rifa'i, A. M., Asgar, M. R. G., Ilmawati, F. I., & Kusriani. (2022). Clustering titik panas menggunakan algoritma Agglomerative Hierarchical Clustering (AHC). *Cogito Smart Journal*, 8(2), 501–513.
- Muhima, R. R., Kurniawan, M., Wardhana, S. R., Yudhana, A., Sunardi, S., Rahmawati, W. M., & Yulastuti, G. E. (2021). *Kupas tuntas algoritma clustering: Konsep perhitungan manual dan program*. Penerbit ANDI.
- Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, II. *WIREs Data Mining and Knowledge Discovery*, 7(6), Article e1219. <https://doi.org/10.1002/widm.1219>
- Nurhayati, A., Mardaweni, R., & Widiastuti, R. M. (2023). Noodle grouping based on nutritional similarity with hierarchical cluster analysis method. *Sainteks: Jurnal Sain dan Teknik*, 5(1), 112–125.
- Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747–748). IEEE. <https://doi.org/10.1109/DSAA49011.2020.00096>

- Wibowo, L., Andarwulan, N., & Indrasti, D. (2024). Kajian implementasi informasi “pilihan lebih sehat” label kemasan mi instan di Indonesia. *Jurnal Mutu Pangan*, 11(1), 63–70. <https://doi.org/10.29244/jmpi.2024.11.1.63>
- World Health Organization. (2023). *WHO global report on sodium intake reduction*. World Health Organization.
- Yusniyanti, A. L., Virgantari, F., & Faridhan, Y. E. (2021). Comparison of average linkage and K-means methods in clustering Indonesia's provinces based on welfare indicators. *Journal of Physics: Conference Series*, 1863(1), Article 012071. <https://doi.org/10.1088/1742-6596/1863/1/012071>