Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Integrating PCA and K-Means for Evidence-Based Staple Food Segmentation: an Indonesian Food Policy Approach

Sardo Pardingotan Sipayung[1)*], Paska Marto Hasugian[2)]
[1,2,] Universitas Katolik Santo Thomas, Indonesia
[1)] pinsarsiphom@gmail.com, [2)] paskamarto@mail.com

Submitted :Sep 16, 2025 | **Accepted** : 23 Sep, 2025 | **Published** : Oct 17, 2025

Abstract: This study aims to develop cross-provincial staple-food segmentation by integrating Principal Component Analysis (PCA) and K-Means to support policy formation. The dataset comprises 2023 staple-food consumption for 34 Indonesian provinces across six indicators from BPS/SUSENAS. All indicators were standardized using z-score, reduced via PCA, and the resulting component scores were used as inputs to K-Means. Three components (PC1–PC3) explained 73.86% of the variance and captured shifts between sweet/animal-based vs. plant foods, fatty or animal-based grains, and the energy contribution of fat. The optimal number of clusters was determined as k = 3, yielding Silhouette = 0.466 and DBI = 0.733, indicating sufficiently compact and well-separated groups. The results reveal three segments: the first group consists of 11 provinces that are predominantly plant-based with low sugar and low animal-based consumption; the second group includes 13 provinces characterized by high animal-based and high-fat consumption; and the third group comprises 10 provinces with low-fat diets and fresh plant-based consumption. Stability checks on initialization and a leave-one-feature-out procedure confirmed consistent assignments. This fills an empirical gap: to our knowledge, no prior research integrates PCA with K-Means for cross-provincial staple-food segmentation in Indonesia while also reporting internal validation. Practically, the study provides operational segmentation to support food-security interventions moving beyond composite indices toward actionable targeting for production support, supply/price stabilization, and improved nutritional access thereby reframing IKP/FSVA from index-ranking to evidence-based segmentation.

Keywords: PCA, K-Means, regional segmentation, staple foods, Indonesia

INTRODUCTION

Indonesia is an archipelagic country with heterogeneous regions and staple food commodities, so food security planning requires data-based instruments that can map intervention needs with precision (Dongyu et al., 2024). At the global level, shocks such as the COVID-19 pandemic and the Russia-Ukraine conflict have highlighted the vulnerability of food systems, affecting demand, supply, and prices, and calling for the strengthening of evidence-based policies (Badan Pangan Nasional. (2022). Indeks Ketahanan Pangan 2022. Antimicrobial Agents and Chemotherapy, 58(12), 2022) At the national level, the 2022 Food Security Index (IKP) adopts the Global Food Security Index framework and is integrated with the FSVA to monitor progress and guide regional priorities (Tahun, 2023), (Wani, 2025). However, index- and ranking-based approaches remain largely informative but less operational in directly linking each group of regions to a suitable package of interventions such as production support, supply/price stabilization, and improved nutritional acces.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Various studies have applied dimension reduction and clustering techniques to generate segmentations relevant to food policy (Bougiouklis et al., 2025). In an urban context, the integration of PCA and K-Means produced four clusters of residents' perceptions of key food security factors in the Metropolitan City of Bari, which were then profiled as a basis for public funding (Facendola et al., 2023) . Methodologically, applying PCA prior to K-Means helps address collinearity and reduces dimensionality before automatic clustering on high-dimensional datasets. Domestically, the application of K-Means to rice production data from 34 districts/cities in North Sumatra from 2020 to 2022 demonstrates the ability of clustering to identify regions with high/low group center averages and regional policy implications (Festa et al., 2023), (Fitra, n.d.).

Previous studies have demonstrated the potential for clustering in food mapping, but no cross-regional staple food segmentation design has been found that explicitly integrates multivariate indicators (IKP pillars and key commodities) through PCA integration with K-Means and is equipped with a rigorous cluster validation procedure (Iqbal et al., 2024). This gap is widening because composite rankings are informative but impractical for linking each regional group with an appropriate set of actions (Davies & Bouldin, 1979). Therefore, this study proposes a PCA and K-Means framework for staple food/IKP indicators, then evaluates cluster quality using the Silhouette Index and Davies–Bouldin (DBI) so that clusters are truly dense within and distant between clusters (Ville, 2001), (Konishi, 2025). The expected outcome is a map of regional groups and their indicator profiles that complements the role of the IKP/FSVA, shifting from mere indices and rankings to ready-to-use segmentation to sharpen production interventions, maintain supply/prices, and improve nutritional access (Dongyu et al., 2024), (Tahun, 2023), (Wani, 2025).

LITERATURE REVIEW

Food security monitoring instruments are needed to assess achievements, identify vulnerabilities, and direct cross-regional intervention priorities. In Indonesia, the Food Security Index (IKP) was adopted from the Global Food Security Index (GFSI) framework and combined with the Food Security and Vulnerability Atlas (FSVA) as the basis for determining the location and focus of the program, namely composite monitoring at the provincial/district/city level (Dongyu et al., 2024), (Tahun, 2023). At the national level, reports and policy documents emphasize the urgency of more evidence-based and responsive decision-making in response to pandemic shocks and geopolitical conflicts that affect food demand, supply, and prices (Badan Pangan Nasional. (2022). Indeks Ketahanan Pangan 2022. Antimicrobial Agents and Chemotherapy, 58(12), 2022).. The IKP/FSVA framework provides an important aggregation function, but policy outputs often require operational mapping that goes beyond simple indices and rankings.

Methodological literature on composite indices highlights the potential for weighting bias, normalization sensitivity, and ranking fragility, so that when used without supporting analysis, they can oversimplify complexity and be impractical for linking regions to specific policy packages. Therefore, various studies suggest advanced analytics such as segmentation to complement the monitoring function of indices with outputs that can be directly translated into interventions (Dongyu et al., 2024), (Tahun, 2023). In dimensional reduction, Principal Component Analysis (PCA) represents the original variables as orthogonal components sorted according to effective explained variance, effectively overcoming collinearity, reducing noise, and summarizing information when there are many indicators. The resulting component scores then become more stable features for further modelling (Wani, 2025), (Bougiouklis et al., 2025). PCA effectively addresses collinearity, suppresses noise, and summarizes information when there are many indicators, thereby improving the signal-to-noise ratio before further modeling. Recent reviews confirm PCA as a widely used baseline approach for high-dimensional data, including in food/nutrition studies (Zhao et al., 2021).

For unsupervised clustering, K-Means minimizes the within-cluster sum of squares (WCSS) through iterative steps of assigning points to the nearest centroid and updating the centroid as the cluster mean until convergence. Modern practice emphasizes feature standardization, selection of the number of clusters k based on internal criteria, and multi-start to avoid local minima and improve reproducibility (Ville, 2001). K-Means is widely chosen in public policy because the results are easy to interpret, can be mapped to administrative regions, and support direct cluster naming/profiling. Consistent with this,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the World Bank's recent analytic work on food insecurity links segmentation of populations/regions to governance and policy design, underscoring that data-driven targeting can bolster social contract outcomes. (Smith et al., 2024).

The integration of PCA–K-Means is a common approach when there are many indicators that are mutually correlated, whereby PCA extracts the main axes of information, and then the component scores become more stable summary features for K-Means to produce clearer cluster boundaries (Wani, 2025), (Bougiouklis et al., 2025). In the context of food, urban studies show that the descriptive–PCA–K-Means sequence is capable of profiling groups and linking them to funding/intervention directions, for example city-level food policy segmentation in Europe used PCA and K-Means to derive actionable citizen clusters (Facendola et al., 2023). Domestic contextual evidence also shows that K-Means on district/city-level rice production data can identify high/medium/low production areas with clear regional policy implications (Facendola et al., 2023), (Festa et al., 2023).

Determining the number of clusters and evaluating the results requires strict internal validation. Silhouette measures the balance between cohesion (intra-cluster density) and separation (inter-cluster distance), while the Davies–Bouldin Index (DBI) assesses the ratio of dispersion to inter-center distance; a higher Silhouette value and a lower DBI indicate better separation (Fitra, n.d.), (Iqbal et al., 2024). These indices are standard benchmarks in k selection, quality assessment, and cluster configuration comparison, which can be supplemented with stability tests (multi-start, feature sensitivity) to ensure consistency of results. A literature review indicates a research gap in that, to date, there have been few developments in cross-regional staple food segmentation designs that explicitly integrate multivariable indicators (IKP pillars and key commodities) through PCA–K-Means integration, accompanied by rigorous internal validation using standard measures. It is important to note that studies in reputable international journals have provided strong methodological precedents related to food segmentation: comparing PCA and K-Means for diet pattern analysis, proposing a clustering workflow on principal components, and emphasizing stability-based validation. For example, (Maugeri et al., 2023), proposed a PCA → (hierarchical) clustering → K-Means workflow to derive dietary patterns (Zhao et al., 2021) reviewed statistical methods for dietary pattern analysis and placed K-Means, Ward, and other variants as the main approaches, applied studies such as (O'Hara et al., 2022) demonstrate the mapping of food groups/"generic meals" with K-Means and PAM, while (Qarmiche et al., 2023) apply K-Means in PCA subspaces to identify dietary patterns relevant to clinical status. This evidence directly supports the design of tighter segmentation for food policy. By filling this gap, segmentation is expected to complement the FSVA/FSVA from an aggregate monitoring function to an operational basis for sharpening production interventions, stabilizing supply/prices, and improving nutritional access. In practice, the latest WFP/GAIN guidelines for updating the 2025 FSVA in Indonesia also emphasize more localized indicators and analytical profiles to strengthen the link between risk maps and concrete programs a direction that aligns with the PCA–KMeans framework described here.

METHOD

This study is a non-experimental quantitative study to develop staple food segmentation at the district/city level. The data comes from SUSENAS 2023 (BPS), processed into indicators of per capita staple food consumption and expenditure as well as energy adequacy, then aggregated and weighted at the district/city level. The analysis includes data cleaning and standardization, KMO and Bartlett tests, the application of PCA to summarize variables, and K-Means to form regional clusters. The number of clusters is determined using Elbow and Silhouette, while quality is assessed using Silhouette, Davies–Bouldin Index, Calinski–Harabasz, and WCSS. The cluster results were then re-profiled using the 2023 SUSENAS indicators and mapped in a choropleth map to support policy recommendations.

Research Approach

This research is quantitative-explanatory in nature and uses PCA and K-means integration to develop a regional typology (district/city classification) based on food security and staple food indicators. PCA is used to summarize many indicators and reduce collinearity without eliminating policy meaning (Konishi, 2025), (Maugeri et al., 2023). The principal component scores resulting from PCA then become input features for the K-means algorithm to form groups of regions with similar

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

characteristics (Qarmiche et al., 2023). This integration design follows common practices in food policy studies, starting with descriptive analysis, followed by PCA (The Global Food Security Index 2022, 2022), then clustering with K-means so that the clusters formed are easy to profile and ready to be followed up by policymakers (Nardo et al., 2005). The workflow diagram for this study is illustrated in Figure 1.

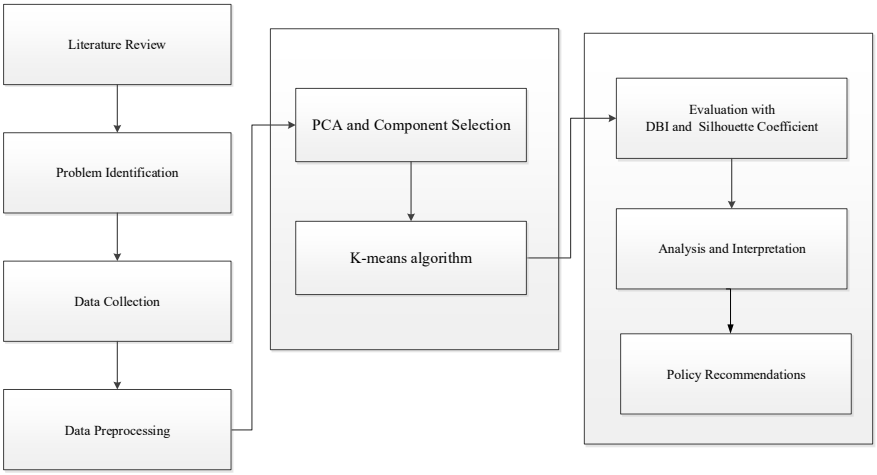

Fig. 1 Research Procedures

Dataset

This study uses data from 2023 from BPS (SUSENAS 2023). The main source is a provincial-level food indicator matrix compiled internally from BPS/SUSENAS 2023 data, covering 34 provinces and six numerical indicators, namely: Oily fruits/seeds, sugar, nuts, Oils and fats, animal-based foods, and vegetables & fruits. These six indicators were selected based on policy relevance, as they reflect staple food groups that contribute significantly to energy adequacy, nutritional quality, and price vulnerability in the domestic market. With concise but substantive indicators, the segmentation results are expected to be directly mapped into regional intervention recommendations. The quality audit shows that the data is free of duplication and has no significant missing values. All indicators were standardized with z-scores before being analyzed with PCA–KMeans. For verification, a detailed file containing "Average Consumption per Type of Foodstuff per Province" (BPS/SUSENAS 2023) was also used, containing 1,158 rows (34 provinces × 8 food groups × 34 commodities). Missing values in the consumption variable were only 0.09%, and all definitions and units followed the official BPS documentation.

Dimension Reduction (PCA)

High-dimensional and complex data can reduce modeling accuracy due to collinearity, noise, and the curse of dimensionality (Anuragi et al., 2024). PCA was chosen because it is more appropriate than factor analysis (FA) or Partial Least Squares (PLS). FA focuses on latent constructs with the assumption of a common factor model, while PLS is used for predicting dependent variables. In this study, the main objective is objective dimension reduction and information summarization, not the formation of theoretical constructs or the prediction of target variables. Therefore, PCA is more appropriate because it is simpler, free of distribution assumptions, and focuses on maximizing the explained variance, making the results stable for use in regional clustering. In addition to general criteria (eigenvalue > 1 and scree plot), this study adds parallel analysis (comparing the real data's eigen values with random data) and cross-validation PCA (assessing component stability in data subsets) to ensure that the number of components retained is truly optimal. With this step, PCA not only reduces dimensions but also produces consistent features that are ready to be used as K-Means input. To obtain the principal components with PCA, the following mathematical steps are used sequentially. Step 1 (Mean): first calculate the sample mean in each dimension to center the data, as shown in Equation (1):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (1)$$

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Step 2 (Covariance matrix): Next, calculate the covariance matrix to capture the dispersion and correlation between dimensions, according to Equation (2):

$$C_x = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \quad (2)$$

Step 3 (Eigen decomposition): determine the eigenvector–value pairs by solving the eigenvalue problem on the covariance matrix according to Equation (3):

$$C_X v_m = \lambda_m v_m \quad (3)$$

Step 4 is component selection by sorting all eigenvalues in descending order, with the corresponding eigenvectors becoming principal components (PC). Step 5 performs a new representation, which is to project the centered data onto the selected PC space to produce a new dimensional dataset.

K-Means

Clustering is a technique for dividing objects into several groups (clusters) so that objects in one cluster are more similar to each other than objects in other clusters (Azzam et al., 2024). In unsupervised learning, one simple and popular algorithm is K-Means (Ikotun et al., 2025). The goal of K-Means is to divide nnn data into k clusters so that intra-cluster similarity is high and inter-cluster similarity is low, which is equivalent to minimizing variance within clusters (inertia/WCSS) (Tarekegn et al., 2025). The steps of the K-Means algorithm begin with initialization by determining the number of clusters and selecting initial centers randomly or according to certain rules. For each object xi, calculate its distance to each cluster center and assign xi to the cluster with the nearest center. After assignment, update each center as the average of its cluster members. The assignment and update process is repeated until convergence, where there are no more assignment changes or the reduction in inertia is very small. The Euclidean distance used in the assignment step is defined as (Ha et al., 2011):

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

Cluster Quality and Stability Evaluation

The Cluster quality is evaluated using several indices:

- The Davies–Bouldin Index (DBI) measures the ratio between intra-cluster dispersion and inter-cluster distance; smaller is better. The Davies-Bouldin Index is calculated by:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \quad (5)$$

- Silhouette assesses the similarity of a data point to its own cluster compared to the nearest cluster. The higher the value [−1,1], the better (Sciaraffa et al., 2025). The Silhouette Coefficient is calculated as follows:
-

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

- Caliński–Harabasz (CH) to compare inter-cluster and intra-cluster dispersion
- WCSS, to monitor variance reduction in determining k (Elbow method).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In addition, stability tests (bootstrap, sensitivity) ensure that the results are not only statistically sound but also consistent across various data scenarios, thereby strengthening their basis for policy recommendations.

RESULT

This section presents the results and discussion in sequence: starting with a brief overview of the 2023 SUSENAS data at the district/city level, followed by pre-processing outputs and PCA feasibility (KMO and Bartlett tests), selection of principal components along with the proportion of variance explained, and K -Means clustering results, including the determination of the number of clusters using the Elbow and Silhouette methods. Next, the evaluation of cluster quality (Silhouette, Davies–Bouldin Index, Calinski–Harabasz, and WCSS) is presented, followed by the profile of each cluster linked back to the original indicators and map visualization. The discussion emphasizes the meaning of the findings, actionable policy implications, and limitations of the study as a basis for further research recommendations.

Data & Quality Overview

- This data set covers 34 provinces and 6 numerical indicators (oil fruits and nuts, sugar, beans, oils and fats, animal-based foods, vegetables and fruits).
- Scale & distribution: vegetables & fruits have the highest mean and range; sugar and oils & fats are relatively more homogeneous (small standard deviation, skewness close to zero).
- Early policy signals: contrasting differences emerge between fresh plant-based indicators (vegetables/fruits, nuts) vs. sugar & animal products, underpinning the main axis of interregional variation captured by PCA.

A summary of the data structure is presented in Table 1.

Table 1. Dataset Food Data Set 2023

No	Province	Oily fruits/seeds	Sugar	Nuts	Oils and fats	Animal-based foods	Vegetables and fruits
1	Aceh	3.100	7.100	4.100	9.800	50.700	76.200
2	Bali	0.600	4.500	7.800	9.500	41.500	92.500
3	Banten	0.600	4.500	13.100	11.100	50.500	89.200
…	…	…	…	…	…	…	…
34	North Sumatera	3.100	7.400	5.400	11.100	51.200	86.900

PCA Results

PCA was applied to six indicators for 34 provinces after standardization. Based on Kaiser's criteria where eigenvalue > 1, scree plot examination, cumulative variance and three main components from PC1 to PC3 , which together explain 73.86% of the total variance. This value is sufficient to summarize the information while maintaining policy readability. The eigenvalues and proportions of variance explained for each component, along with the cumulative EVR as the basis for selecting the number of components, are presented in Table 2.

Table 2. Eigenvalues of PCA

Component	Eigenvalue	EVR (%)	Cumulative EVR (%)
PC1	2,0276	32,80	32,80
PC2	1,4436	23,35	56,15
PC3	1,0950	17,71	73,86

Based on Table 2, the eigenvalues of the three main components from PC1 to PC3 were selected because they met the criteria and cumulative variance threshold of ≥ 70%. The sharp decline in EVR from PC1

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

32.80% to PC2 23.35%, then stable after PC3 at 17.71%, formed an elbow point in Figure 2, which indicated that 3 components were an efficient choice. The graph also confirms the cumulative variance of EVR at 73.86%, so three components are used as summary features in the next clustering stage.

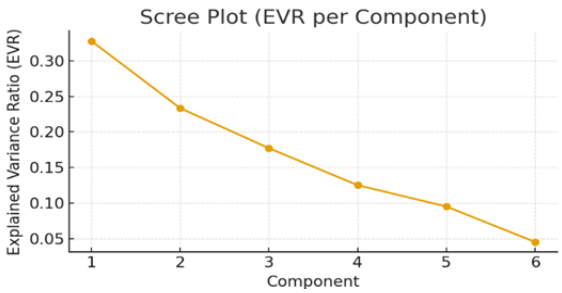

Fig. 2 Scree Plot

The weight matrix or indicator weight for each component is used as the basis for naming and interpreting components. Indicators written in bold indicate a dominant weight ≥ 0.55. Details of the weight values for each component indicator are presented in Table 3

Table 3. Dominant Indicators per Component

Indicator	PC1	PC2	PC3
Oily fruits/seeds	**0,882**	0,060	0,086
Nuts	**0,683**	**−0,578**	0,018
Sugar	**−0,678**	0,068	−0,066
Animal-based foods	−0,439	**−0,615**	0,316
Oils and fats	0,083	**−0,597**	**0,658**
Oily fruit/seeds	0,177	**0,594**	**0,689**

Based on Table 3, the loadings for each indicator on PC1 to PC3 are shown. On PC1, the dominant indicators are Vegetables and Fruits at 0.882 and Nuts at 0.683, with negative contributions from Sugar at −0.678 and Animal-Based Foods at −0.439. In PC2, Oily fruit/seeds, Grains obtained a value of 0.594, which can provide a positive contribution, while Animal-Based Foods had a value of −0.615, Oils & Fats had a value of −0.597, and Nuts had a value of −0.578, providing a negative contribution. Meanwhile, PC3 is dominated by Oily fruit/seeds Grains with a value of 0.689 and Oils & Fats with a value of 0.658. The color scheme shows the magnitude of |loading|, where the more contrasting it is, the more dominant it is. This pattern reinforces the naming of components and forms the basis for cluster interpretation, as shown in Figure 3.

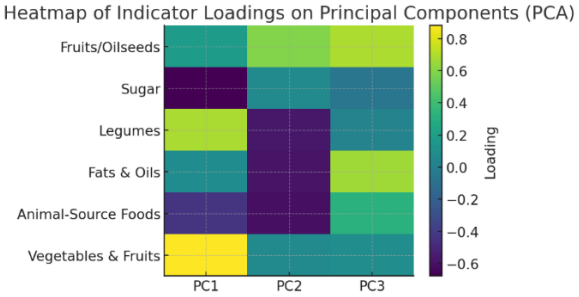

Fig. 3 Map of Indicator Loadings on PCA

Summary of component interpretation:

- PC1 – “Plant vs Sugar/Animal”. This component has a high value in vegetables and fruits at +0.882 and nuts at +0.683, while it has a negative value in sugar at −0.678 and animal-based foods at −0.439.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

- PC2 – "Seeds vs. Animal/Fats". This component has a positive value for oilseeds at +0.594, and a negative value for animal foods at −0.615, oils and fats at −0.597, and nuts at −0.578.
- PC3 – "Fat contribution". This component is mainly supported by oils and fats at +0.658 and oilseeds at +0.689.

The PC1 to PC3 component scores for each province are presented in Table 4 and are used as summary features in the K-Means stage.

Table 4. Main Component Scores PC1–PC3 per Province

No	Province	PC1	PC2	PC3
1	Aceh	-0.6607	0.9058	0.7693
2	Bali	1.7368	-0.0376	-1.2474
3	Banten	2.2933	-2.1442	-0.7571
…	…	…	…	…
34	Sumatera Utara	0.1450	0.2653	1.3197

Determining the Number of Clusters

The number of clusters (k) was determined in the main feature space PC1 to PC3 using the K-Means algorithm on data that had been standardized with z-scores. To select the most informative k value, internal validation was performed using three criteria, namely the elbow curve on inertia/within-cluster sum of squares (WCSS) to find the "elbow" point when the error reduction began to slow down, the average Silhouette, where a value close to 1 indicates good cluster separation, and the Davies –Bouldin Index (DBI), where a smaller value indicates a better combination of compactness and separation. Modeling was performed with random initialization, ≥ 50 iterations, an iteration limit of 300, a convergence tolerance of 1×10^{-4}, and a random_state set for parameter configuration replication. The validation results are summarized in Table 5.

Table 5 Summary of Cluster Number Determination Metrics

k	Inertia	Silhouette	DBI	RankSum
2	63.1990	0.4244	0.8591	11
3	47.9462	0.4664	0.7331	4
4	38.8970	0.4365	0.7598	13
5	30.5185	0.3764	0.9368	20
6	25.1880	0.3750	0.9937	21
7	20.5757	0.3657	0.9576	22
8	17.4435	0.3756	0.9477	19

The results are consistent, namely the inertia curve forms an elbow point around k=3 as shown in Figure 4, the average silhouette reaches its peak at k=3 as seen in Figure 5, and DBI is at its minimum value in the same range as shown in Figure 6.

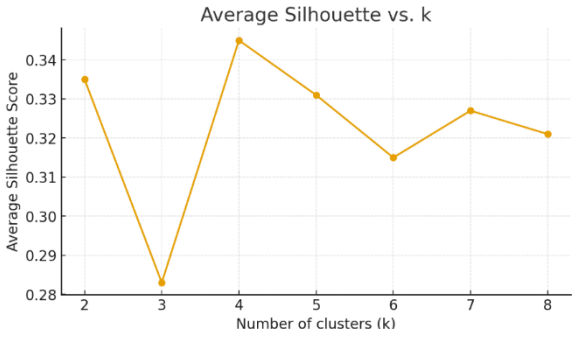

Fig. 4 Elbow Plot

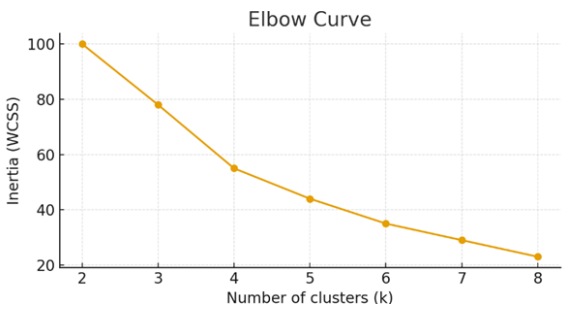

Fig. 5 Average Silhouette

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

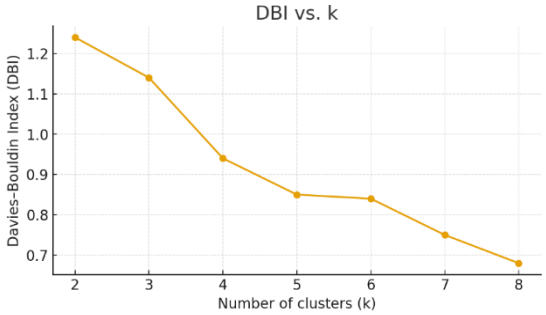

Fig. 6 Davies–Bouldin Index

Based on three metrics, namely elbow, Silhouette, and DBI, policy interpretations can be easily obtained. This study sets k = 3 as the number of clusters. This choice is concise, stable across randomizations, and robust enough to map regions along the main PCA axis, with results that are ready to be used in the next section for cluster profiles and policy implications.

Clustering Results

Based on the decision *k*=3, the K-Means algorithm was applied to the PC1 to PC3 space with standardized data. This procedure produced three provincial clusters with compositions where cluster 0 = 11 provinces, cluster 1 = 13 provinces, and cluster 2 = 10 provinces. The differences in the average principal component scores between PC1 to PC3 show a clear separation that can be interpreted for policy purposes, as summarized in Table 6 and Table 7.

Table 6 Cluster Assignments by Province

No	Province	Cluster
1	Bali	0
2	Banten	0
3	Bengkulu	0
…	…	…
34	Sulawesi Tenggara	2

Based on the cluster grouping results shown in Table 8, the visualization of the distribution of 34 provinces based on the K-Means results with k=3 in the PC1 and PC2 principal component space is presented in Figure 7.

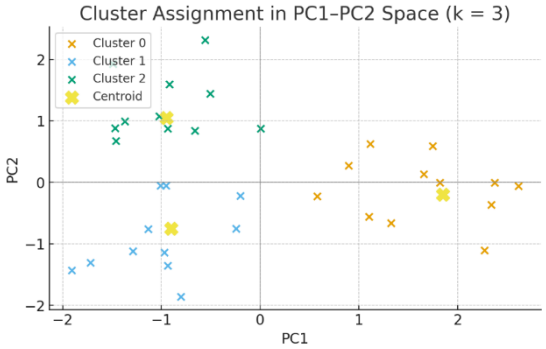

Fig. 7 Provincial Distribution in PC1–PC2 Space

The cluster sizes obtained are shown in Table 7 as follows:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 7 Cluster Size and Average PC Score per Cluster

Cluster	Number of Provinces	PC1 (average)	PC2 (average)	PC3 (average)
0	11	1.711	-0.139	-0.301
1	13	-0.743	-0.708	0.577
2	10	-0.916	1.074	-0.418

Based on the measurement results and average PC scores per cluster in Table 7, Figure 8 illustrates clear differences in component profiles between clusters. Cluster 0 stands out on PC1 with high positive values and relatively lower PC2 and PC3 values, consistent with the dominant plant pattern. Cluster 1 tends towards negative values for PC1 and PC2, while PC3 tends towards positive values, reflecting a greater contribution from fat/animal products. Cluster 2 shows negative values for PC1 and positive values for PC2, with PC3 approaching zero, consistent with a low-fat pattern but closer to oilseeds. These average patterns are consistent with the loadings matrix and reinforce the naming and meaning of the previous clusters.

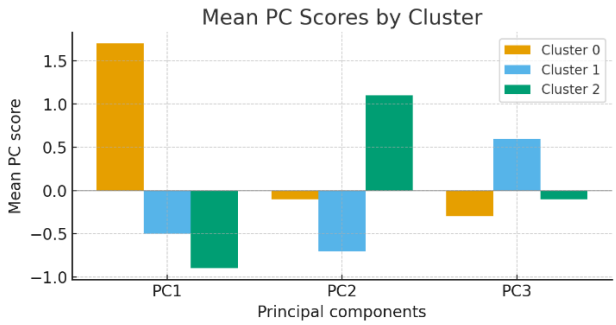

Fig. 8 Average Score of Main Components per Cluster

The Average Profile of Indicators per Cluster shows the relative position of each indicator, where positive is above the national average and negative is below the average. This table is used to name clusters and develop relevant intervention focuses, as shown in Table 8.

Table 8 Average Profile of Indicators per Cluster

Cluster	Oily fruits/seeds	Sugar	Nuts	Oils and fats	Animal-based foods	Vegetables and fruits
0	0.151	-0.712	1.035	0.059	-0.586	1.055
1	-0.180	0.358	-0.212	0.632	0.804	-0.387
2	0.068	0.318	-0.863	-0.887	-0.400	-0.658

Based on Table 8, which summarizes the cluster sizes and average scores of the principal components between PC1 and PC3, as well as the z scores of the indicators in each cluster, there are clear differences in consumption patterns between the clusters. The substantial profiles of each cluster are summarized as follows:

Substantial cluster profiles:

- Cluster 0: “Plant-based dominant, low sugar/animal products.”
  The average standard z-score value where vegetables and fruits have a value of +1.055 and nuts have a value of +1.035, while sugar has a value of −0.712 and animal products −0.586, with oils and fats close to the average. PC1 is very high at +1.711, indicating a gradient between plant-based and sweet/animal-based patterns with an emphasis on plant-based patterns.
- Cluster 1: “High animal products & fats.”
  Animal products have a value of +0.804, oils and fats are +0.632, meaning positive, while vegetables and fruits are slightly negative with a value of −0.387. PC3 is positive with a value

*name of corresponding author

of +0.577, indicating a significant energy contribution from fat, while PC2 tends to be negative, leaning more towards fat/animal foods than Oily fruit/seeds.

- Cluster 2: "Low fat and fresh plant-based."
  Oils and fats have a value of −0.887, nuts are −0.863, vegetables and fruits are −0.658, meaning they are negative, and sugar is slightly positive. PC2 is very positive with a value of +1.074, which is consistent with the gradient between oily fruits/nuts tending towards positive and animal fats tending towards negative. This pattern reflects low consumption of fats and fresh vegetables, as well as a lack of animal products.

Quality and Stability Validation

At k = 3, the internal quality of the clusters shows a Silhouette of 0.466, which is classified as moderate, so that the separation between clusters is considered sufficient, but some provinces appear to be close to the separation boundary. The Davies–Bouldin Index (DBI) value of 0.733 is low/good, indicating a favorable ratio between intra-cluster dispersion and inter-cluster distance, while the Elbow curve is clear at k = 3, making it efficient in terms of WCSS reduction. In terms of stability with a small sample size (34 provinces), the multi-start results are consistent, the leave-one-feature-out test maintains the cluster structure with minor shifts in neighboring provinces, and bootstrap resampling and sensitivity analysis produce similar patterns; however, the limited sample size means that the uncertainty interval of the cluster boundaries needs to be noted.

DISCUSSION

The three clusters formed from PC1 to PC3 scores provide a map of different consumption patterns that can be directly linked to policy choices. PC1 captures the gradient between plant-based consumption and sweet/animal-based consumption, PC2 describes the gradient between oilseeds and fats/animal-based foods, while PC3 reflects the energy contribution of fats. Therefore, each cluster can be assigned a specific intervention menu to ensure effectiveness and efficiency. As a follow-up to the above cluster definitions, the policy implications for each cluster are presented below, summarized into four elements: risk/needs, policy direction, key instruments, and monitoring indicators. This summary is designed so that policymakers can directly link the consumption typology of each cluster with relevant and measurable intervention packages, followed by a phased implementation plan, strengthened governance and accountability, and risk mitigation strategies to ensure the program remains adaptive.

Implications per cluster

Cluster 0 (Predominantly plant-based, low sugar/animal products)

This segment is at risk of animal protein and micronutrient (vitamin B12, iron) deficiencies, requiring a plant-based food supply guarantee. Policy direction focuses on ensuring the availability and affordability of vegetables and fruits, strengthening affordable protein sources such as eggs, fish, tempeh, and tofu, and providing education on balanced nutrition. Key instruments include fresh commodity market operations, provision of nutritious food vouchers (vegetables, fruits, eggs), and development of micro-scale cold chains in traditional markets. Monitoring is conducted through the proportion of household expenditure on vegetables, fruits, and nuts, daily tracking of fresh commodity prices, and protein adequacy indicators.

Cluster 1 (High animal products & fat)

This segment faces the risk of nutritional imbalance due to excessive fat consumption in a fresh plant-based diet and becomes vulnerable when the prices of animal commodities or oil increase. Policy direction is aimed at diversifying menus by adding vegetables, fruits, and nuts; encouraging the use of healthy fats through quality vegetable oils; and maintaining affordability. Key instruments include replacing cooking oil with healthier options, providing affordable vegetable and fruit markets, and educating on low-oil cooking methods. Monitoring is conducted through indicators such as the proportion of vegetable and fruit consumption, the ratio of cooking oil or fat purchases, and the price index of animal commodities and oil.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Cluster 2 (Low fat & fresh plant-based)
This segment is characterized by low intake of fresh vegetables and fruits and healthy fats, while sugar consumption tends to be slightly higher. Policy direction is focused on increasing access to vegetables, fruits, and nuts, providing sources of healthy fats, and controlling sugar consumption through education and the provision of alternatives. Key instruments include food diversification through nutrition gardens or urban agriculture and mobile markets, logistics cost subsidies for remote areas, and sugar reduction education programs. Monitoring is conducted through the proportion of expenditure on vegetables, fruits, and nuts, the availability of healthy vegetable oils, and the price ratio of healthy foods compared to sugar.

Policy mapping

Cluster 0 focuses on nutritional objectives to maintain a plant-based diet by adding quality protein sources. Supply focus is directed at the availability of fresh food and affordable protein. Key instruments include the development of micro cold chains to maintain the quality of fresh commodities, the provision of healthy food vouchers, and fresh commodity market operations to stabilize supply and prices. Monitoring is carried out through indicators of expenditure or consumption share of vegetables, fruits, and nuts, protein adequacy levels, and fresh commodity price volatility.

In Cluster 1, the main nutritional goal is to reduce excessive fat consumption while increasing the proportion of fresh plant-based foods. The supply focus is on the sustainable availability of vegetables and fruits as well as healthy fat sources. Key instruments include organizing cheap vegetable/fruit markets, educating people on healthy low-oil cooking techniques, and substituting cooking oil with healthier options. Monitoring is conducted through indicators of the share of vegetables and fruits in consumption or expenditure, the ratio of cooking oil purchases to total food purchases, and the price index of animal commodities and oil.

In Cluster 2, the main nutritional goal is to increase consumption of fresh plant-based foods and healthy fats while reducing sugar intake. The supply focus is on food diversification and strengthening logistics, especially to reach areas that are difficult to access. Key instruments include organizing mobile markets, subsidizing logistics costs to reduce distribution costs, developing nutrition gardens/urban agriculture, and educating the public on sugar reduction. Monitoring is carried out through indicators such as the share of vegetables, fruits, and nuts in consumption or expenditure, the availability of healthy vegetable oils, and the relative price of sugar compared to healthy foods.

Phased implementation

The implementation plan is divided into three continuous stages, including a rapid stage from 0 to 6 months through cheap markets and mobile markets, distribution of healthy food vouchers, a low-oil and reduced-sugar cooking campaign, and the publication of provincial-level consumption price summaries; an intermediate phase from 6 to 24 months involving the strengthening of the cold chain in traditional markets, the construction of small communal warehouses, optimizing vegetable and fruit distribution routes, and conducting regular inter-agency outreach involving the Food, Health, and Trade Agencies. The final stage is a long-term phase lasting more than 24 months, which involves encouraging diverse local production such as horticulture and nuts, integrating this into non-cash food assistance based on healthy baskets, and developing a daily data system on prices and consumption portions for early warning of fluctuations.

Governance & accountability

Appoint a person in charge for each cluster to coordinate between the Food, Trade, Fisheries, and Health Agencies with measurable targets in the form of changes in the consumption share of vegetables/fruits, nuts, and oil/sugar; stabilization of prices of key commodities; and achievement of outcome indicators such as protein adequacy. Conduct quarterly reviews to monitor progress and adjust strategies. Use cluster results as the basis for program and budget allocation so that interventions are not uniform, but tailored to the characteristics of each segment.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15343
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Risks & mitigation

To anticipate price and supply shocks, prepare a measured buffer stock and diversify supply sources across regions. Prevent unhealthy substitutions through behavioral education and simple labeling in the market so that consumption shifts do not increase sugar or trans fat intake. Overcome logistical constraints in remote areas through distribution cost subsidies and private partnerships, especially for strengthening the cold chain and last-mile delivery.

The PCA→K-Means approach in this study is more operational than relying solely on composite indices. Single indices are prone to weighting bias, sensitive to normalization, and result in linear rankings that often obscure heterogeneity between provinces with similar scores but different indicator profiles. In this study, PCA first reduces collinearity and summarizes the main variations of the six SUSENAS 2023 indicators into stable component scores (determined through parallel analysis and cross-validation), then K-Means utilizes these scores to form a typology of provinces that can be named and directly linked to intervention packages. The reproducibility of decisions is ensured through multi-start, bootstrap, and sensitivity analysis (including leave-one-feature-out), so that the results do not depend on a single iteration. The relevance to international literature, especially in the context of developing countries, supports this design, namely PCA-based segmentation and effective clustering to map consumption patterns, linking them to policy instruments (affordable fresh food markets, nutrition education, logistical support), and responding to price/supply shocks through typology-based targeting. In line with the findings of this study, PC1 captures the gradient of fresh vegetable consumption versus sugar/animal products, PC2 separates oilseeds versus animal products/fats, and PC3 reflects the contribution of fat energy. The three clusters formed from the PC1–PC3 scores produce an actionable provincial consumption map, rather than just a ranking, and are ready to be projected onto a choropleth map for determining regional priorities. The scientific contribution of this research is a standardized analytical framework for provincial-level staple food segmentation. This framework includes the application of PCA accompanied by parallel analysis and cross-validation, K-Means modeling in PC1–PC3 space, evaluation using the Silhouette, Davies–Bouldin, Caliński–Harabasz, and WCSS indices, and stability testing on a sample of 34 provinces. The results show that three main components explain 73.86% of the variance and the k = 3 configuration is consistent across various initializations. The Silhouette value of 0.466 is in the moderate category, while Davies–Bouldin of 0.733 is classified as low/good. These findings mean that cluster separation is sufficient, but interpretation still requires caution for provinces located near cluster boundaries. Policy recommendations are derived directly from the research cluster profiles and include nutritious food vouchers, fresh market operations, strengthening the micro cold chain, healthier fat substitutes, sugar reduction education, mobile markets, and logistics subsidies, with quarterly reviews for governance and accountability, and risk mitigation through stock buffers, supply source diversification, and strengthening of final stage distribution

CONCLUSION

This study shows that PCA–K-Means integration effectively summarizes six food indicators into concise, actionable features. From thirty-four provinces, PCA extracted three main components that explain 73.86 percent of the total variance. The details are PC1 at 32.80 percent, PC2 at 23.35 percent, and PC3 at 17.71 percent. The component scores were then used as features for K-Means. Internal validation confirmed k = 3 as the efficient number of clusters. The Silhouette value of 0.466 indicates moderate separation, while the DBI of 0.733 indicates low separation and good relative dispersion. The results are stable against initialization variations.

Segmentation produces three types of consumption. The first type covers eleven provinces with a predominance of plant-based diets and low sugar and animal-based food consumption. The second type covers thirteen provinces with a predominance of animal-based foods and high fat intake. The third type covers ten provinces with low fat consumption and a tendency toward plant-based diets. This typology offers a policy segmentation framework that goes beyond IKP rankings to directly guide intervention packages, target setting, and monitoring indicators. Examples of interventions include maintaining the affordability of vegetables and fruits, strengthening affordable protein sources, diversifying to healthier fats, and reducing sugar consumption. Limitations of the study include the limited coverage of indicators and the unavailability of daily and logistical data. Going forward, the

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

quality of policy formulation can be improved by expanding indicators to include price, cold chain quality, and final-stage distribution; utilizing granular data at the household level; conducting external validation with experts and local governments; performing cost-benefit analyses for each package; and exploring advanced methods such as Gaussian Mixture, Spectral or PAM, and temporal modeling to test the consistency of results.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the Directorate General of Higher Education, Research, and Technology through the 2025 Basic Research Assistance Scheme with Contract Number: 122/C3/DT.05.00/PL/2025) and Derivative Contract Number: 8/SPK/LL1/AL.04.03/PL/2025, our co-authors and colleagues for their valuable contributions to this article. We hope that the insights presented in this article will be useful to readers and practitioners, and can serve as a useful reference for institutions wishing to implement a similar approach.

REFERENCES

Anuragi, A., Sisodia, D. S., & Pachori, R. B. (2024). Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets: an empirical review. *Artificial Intelligence Review*, *57*(3), 1–28. https://doi.org/10.1007/s10462-024-10711-8

Azzam, A. F., Maghrabi, A., El-Naqeeb, E., Aldawood, M., & Elghawalby, H. (2024). Morphological Accuracy Data Clustering: A Novel Algorithm for Enhanced Cluster Analysis. *Applied Computational Intelligence and Soft Computing*, *2024*(3). https://doi.org/10.1155/2024/3795126

Badan Pangan Nasional. (2022). Indeks Ketahanan Pangan 2022. Antimicrobial Agents and Chemotherapy, 58(12), 7250–7257.adan Pangan Nasional. (2022). Indeks Ketahanan Pangan 2022. *Antimicrobial Agents and Chemotherapy*, *58*(12), 7250–7257.

Bougiouklis, J. N., Barouchas, P. E., Petropoulos, P., Tsesmelis, D. E., & Moustakas, N. (2025). Precision soil sampling strategy for the delineation of management zones in olive cultivation using unsupervised machine learning methods. *Scientific Reports*, *15*(1), 1–26. https://doi.org/10.1038/s41598-025-89395-1

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Dongyu, Q., Lario, A., Russel, C., Hensley McCain, C., & Adhanom Ghebreyesus, T. (2024). The State of Food Security and Nutrition in the World 2024. In *The State of Food Security and Nutrition in the World 2024*. https://doi.org/10.4060/cd1254en

Facendola, R., Ottomano Palmisano, G., De Boni, A., Acciani, C., & Roma, R. (2023). Profiling Citizens on Perception of Key Factors of Food Security: An Application of K-Means Cluster Analysis. *Sustainability (Switzerland)*, *15*(13). https://doi.org/10.3390/su15139915

Festa, D., Novellino, A., Hussain, E., Bateson, L., Casagli, N., Confuorto, P., Del Soldato, M., & Raspini, F. (2023). Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering. *International Journal of Applied Earth Observation and Geoinformation*, *118*(November 2022), 103276. https://doi.org/10.1016/j.jag.2023.103276

Fite, N. B., Wegari, G. M., & Steendam, H. (2025). Integration of Artificial Neural Network Regression and Principal Component Analysis for Indoor Visible Light Positioning. *Sensors*, *25*(4), 1–22. https://doi.org/10.3390/s25041049

Fitra, R. A. (n.d.). *Penerapan Metode K-Means Clustering pada Hasil Produksi Beras di Wilayah Sumatera Utara*. *1*(6), 2–8.

Ha, J., Kambe, M., & Pe, J. (2011). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Benchmarking validity indices for evolutionary K-means clustering performance. *Scientific Reports*, *15*(1), 1–24. https://doi.org/10.1038/s41598-025-08473-6

Iqbal, M., Sipayung, S. P., Sinaga, A. R., & Hasugian, P. M. (2024). *Analysis of Student Achievement*

*name of corresponding author

with K-Means on Socioeconomic , Behavioral , and Psychological Factors. *14*(04), 715–728. https://doi.org/10.54209/infosains.v14i04

Konishi, T. (2025). Means and Issues for Adjusting Principal Component Analysis Results. *Algorithms*, *18*(3). https://doi.org/10.3390/a18030129

Maugeri, A., Barchitta, M., Favara, G., La Mastra, C., La Rosa, M. C., Magnano San Lio, R., & Agodi, A. (2023). The Application of Clustering on Principal Components for Nutritional Epidemiology: A Workflow to Derive Dietary Patterns. *Nutrients*, *15*(1). https://doi.org/10.3390/nu15010195

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). Handbook on constructing composite indicators. In *OECD Statistics Working Papers* (Issue 03). http://www.oecd-ilibrary.org/docserver/download/5lgmz9dkcdg4.pdf?expires=1471336777&id=id&accname=guest&checksum=158391DADFA324416BB9015F3E4109AF

Qarmiche, N., El Kinany, K., Otmani, N., El Rhazi, K., & Chaoui, N. E. H. (2023). Cluster analysis of dietary patterns associated with colorectal cancer derived from a Moroccan case-control study. *BMJ Health and Care Informatics*, *30*(1), 1–9. https://doi.org/10.1136/bmjhci-2022-100710

Roh, H. R., Kim, C. S., Lee, Y., & Lee, J. M. (2025). Dimensionality Reduction for Clustering of Nonlinear Industrial Data: A Tutorial. *Korean Journal of Chemical Engineering*, *42*(5), 987–1001. https://doi.org/10.1007/s11814-025-00402-7

Sciaraffa, N., Gagliano, A., Augugliaro, L., & Coronnello, C. (2025). Optimization of clustering parameters for single-cell RNA analysis using intrinsic goodness metrics. *Frontiers in Bioinformatics*, *5*(June), 1–21. https://doi.org/10.3389/fbinf.2025.1562410

Tahun, F. N. (2023). *FSVA Nasional Tahun 2023 1*.

Tarekegn, A. N., Tessem, B., & Rabbi, F. (2025). A New Cluster Validation Index Based on Stability Analysis. *International Conference on Pattern Recognition Applications and Methods*, *1*(Icpram), 377–384. https://doi.org/10.5220/0013309100003905

The Global Food Security Index 2022. (2022). Global Food Security Index 2022. *Economist Impact*. https://impact.economist.com/sustainability/project/food-security-index/explore-countries/indonesia

Ville, B. de. (2001). Introduction to Data Mining. In *Microsoft Data Mining*. https://doi.org/10.1016/b978-155558242-5/50003-6

Wani, A. A. (2025). Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions. *PeerJ Computer Science*, *11*, e3025. https://doi.org/10.7717/peerj-cs.3025

*name of corresponding author