Check Plagiarism

by Jonson Manurung

Submission date: 12-Feb-2025 09:10PM (UTC+0700)
Submission ID: 2586455902
File name: IDSS_Ryan.docx (1.66M)
Word count: 4070
Character count: 24051

Journal of Intelligent Decision Support System (IDSS) x (x) (20xx), pp. x-xx

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Comparison of K-Means Clustering with Hierarchical Agglomerative Clustering for the Analysis of Food Security of Rice Sector in Indonesia

Ryan Fahlepy Sinaga[1], M Azhar Prabukusumo[2], Jonson Manurung[3]

[1,2,3] *Informatika, Universitas Pertahanan Republik Indonesia, Bogor, Indonesia* (9pt)

Article Info

Article history:

Received Jun 9, 2018
Revised Nov 20, 2018
Accepted Jan 11, 2019

Keywords:

First keyword
Second keyword
Third keyword
Fourth keyword
Fifth keyword

ABSTRACT (9 PT)

Food security in the rice sector is a critical issue in Indonesia, as rice is the main staple commodity. This study compares two clustering methods, K-Means Clustering and Hierarchical Agglomerative Clustering (HAC), to group provinces in Indonesia based on consumption patterns, production, price, and population. Data is obtained from the Central Bureau of Statistics (BPS) and covers all provinces in Indonesia. K-Means clusters the data based on the Euclidean distance to the centroid, while HAC uses a bottom-up hierarchical approach. As a result, both methods produce three similar clusters. HAC is more effective in distinguishing rice price patterns, especially in high-price regions such as Central Papua and Papua Mountains. Meanwhile, K-Means is superior in clustering provinces based on production and consumption, with West Java, Central Java, and East Java as the main producers.The findings provide data-driven policy recommendations to improve food security in the rice sector. Provinces with low production and high consumption require distribution interventions and productivity improvements, while provinces with high production can become national supply centers. This research highlights the importance of clustering analysis in formulating adaptive and sustainable food security strategies.

Corresponding Author:

Ryan Fahlepy Sinaga,
Program Studi Informatika,
Universitas Pertahanan Republik Indonesia,
Kawasan IPSC Sentul, Sukahati, Kec. Citeureup, Kabupaten Bogor, Jawa Barat 16810.
Email: humas@idu.ac.id

Introduction

Indonesia is an agricultural country where agricultural activities are very important for food security (Rozaki, 2021; Sintiya, 2023). One of the food crops that has an important role in supporting food security is the rice commodity (Marzani & Juliannisa, 2024). One of the staple foods, rice, is often consumed and supports two-thirds of the world's population (Kasote et al., 2022; Mohidem et al., 2022). Therefore, the sustainability of rice production and distribution is an important factor in maintaining the stability of food security in Indonesia. In Indonesia, food availability is synonymous with rice availability (Rusadi Akhmad, 2023). The balance between rice production and consumption is something that needs to be considered because there are differences in the fulfillment of consumption in each region. In maintaining this balance, the most influential factor is the population (Marzani & Juliannisa, 2024).

The imbalance in rice production and consumption may raise concerns about a potential food crisis in the future (Mavroeidis et al., 2022). If the population continues to increase while the availability of rice is insufficient, the short-term solution is imports. However, dependence on imports is not an ideal solution in the long run (Setiani et al., 2021). Therefore, the government is expected to formulate an effective strategy to increase rice production while ensuring its equitable distribution throughout the region. According to Saefudin from the Planning Bureau of the Secretariat General of the Ministry of Agriculture (2023), the government seeks to increase the area of cultivated land to boost rice production, with the target of reducing imports by 2024 and achieving rice self-sufficiency by 2025. In addition, there is a grand vision to make Indonesia the world's food barn in the next ten years.

Indonesia is predicted to continue to experience a demographic bonus in the next few years (Adriani & Yustini, 2021; Dwi Ariyanti et al., 2024). Based on data from the BPS (2024) Indonesia's population will be 281.6 million in 2024. This number increased by 1.04% compared to last year. According to BPS (2025), rice production in 2024 for food consumption by the population is estimated at 30.62 million tons, 480.04 thousand tons or 1.54 percent less than rice production in 2023 which amounted to 31.10 million tons. In addition, a report from United States Department of Agriculture (2024) that the total rice consumption of the Indonesian population in 2024 will be 36.5 million tons. If the trend of population growth and rice consumption continues to increase while rice production continues to decline, then national food security is threatened because the increase in population and demand is not matched by adequate availability.

In this context, the government needs to formulate appropriate food security policies to maintain a balance between rice availability and consumption across regions. This research implements K-Means Clustering and Hierarchical Agglomerative Clustering (HAC) methods to cluster regions based on rice availability and consumption patterns. K-Means works by grouping data into k clusters based on the distance to the nearest centroid, then updating the centroid until it converges or reaches the maximum iteration (Ay et al., 2023; Oti et al., 2021). Meanwhile, HAC works with a bottom-up approach, meaning that each piece of data is considered a separate cluster and gradually combined based on similarities to form the final cluster (Chhabra & Mohapatra, 2022; Yu & Hou, 2022). Both approaches allow the identification of regions with similar rice consumption and production patterns, which can be used to design more targeted distribution strategies and policy interventions. However, these two methods are not the sole basis for decision-making, but rather tools to evaluate food security conditions in a more objective and data-driven manner.

One of the previous studies that became a reference was research conducted by Mawarni & Budi (2022) K-Means Clustering to assess student discipline based on attendance, neatness, and behavior. Analysis with Microsoft Excel and Orange divides students into three clusters with different discipline levels, the research helps schools in guidance and counseling. Another research using K-Means was conducted by Mirantika et al. (2021) to group the spread of COVID-19 in West Java into 3 clusters based on the level of cases, helping the government in its pandemic handling strategy. The results of this study are expected to support the government's strategic decision making in handling the spread of COVID-19 more effectively. One of the studies conducted using HAC is the research of Rahim et al. (2021) which categorizes data on the results of students' physical fitness scores in 2019/2020 at the State Police School. The methods used include preprocessing, normalization, selecting the number of clusters with the Elbow Method, and applying K-Means. The clustering results successfully grouped students based on their physical scores. Students with high scores are potentially placed in the Brimob Unit, while those with lower scores tend to go to the Polda or Polres units, this research helps the placement selection more objectively. Another study was conducted by Priambodo & Jananto (2022) which compared the K-Means and Hierarchical Agglomerative Clustering (HAC) algorithms in inventory planning at the manufacturing company PT Multi Lestari. The goal is to group goods based on past sales patterns in order to estimate the optimal amount of inventory to avoid excess or shortage of stock. The data used is sales data from February to June 2021, which is then grouped using both clustering algorithms. The results show that K-Means and AHC are equally capable of clustering goods based on the average amount sold, but produce a different number of clusters. Therefore, this research recommends further studies to

24
determine which algorithm is more accurate in predicting inventory needs based on real sales data in the past.

Method

14
K-Means Clustering Algorithm

The K-Means algorithm is an iterative clustering algorithm that partitions a dataset into k predefined clusters. The K-Mean Clustering algorithm is presented below (Ahmad & Khan, 2021; Zubair et al., 2024):22

1) Determine the desired number of clusters (k) in the dataset.
2) Determine the initial cluster center (centroid) by taking the smallest, average and largest values. 27 15
3) Calculating the closest distance between each data and the Centroid. Calculating the closest distance to the Centroid uses the Euclidean distance formula. The formula can be seen below:
   $\boldsymbol{d(xi, \mu j) = \sqrt{(xi - \mu j)^2}}$ ....(1)
   Description:
   $\boldsymbol{xi}$: Criteria data
   $\boldsymbol{\mu j}$: Centroid of the jth cluster
4) Recalculate the Cluster center with the current Cluster members. The formula can be seen below:
   $\boldsymbol{\mu j(t+1) = \frac{1}{Nsj} \sum j \in Sj\ xj}$ ....2)
   Description:
   $\boldsymbol{\mu j(t+1)}$: New centroid at the 1st iteration
   $\boldsymbol{Nsj}$: Number of data in cluster sj;

31
Hierarchical Agglomerative Clustering (HAC) Algorithm

Hierarchical Agglomerative Clustering is a clustering method that builds a hierarchy of data with a bottom-up approach, namely by combining data points one by one until they form one large cluster. The HAC algorithm is presented as follows (Chhabra & Mohapatra, 2022; Monath et al., 2021):

1) Calculating the Euclidean distance matrix (as in Formula 1).
2) Merge the two closest clusters. If the distance between objects a and b has the smallest distance value compared to the distance between other objects in the Euclidean distance matrix, the combined two clusters in the first stage is d_ab.
3) Update the distance matrix according to the Agglomerative method clustering technique If d_ab is the closest distance from the Euclidean distance matrix, then the formula for the agglomerative method is:
   a. Single linkage formula
   $\boldsymbol{d_{(ab)c} = min\{d_{a,c}; d_{b,c}\}}$
   b. Average linkage formula
   $\boldsymbol{d_{(ab)c} = average\{d_{a,c}; d_{b,c}\}}$
   c. Complete linkage formula
   $\boldsymbol{d_{(ab)c} = max\{d_{a,c}; d_{b,c}\}}$
4) Repeating steps 2 and 3 until only one cluster remains
5) Drawing up the Dendrogram

Research Stages

The research subjects in this study are provinces in Indonesia that are analyzed based on rice consumption and production data. The object of the study includes rice consumption and production patterns in each province, which are clustered using the K-Means Clustering and Hierarchical Agglomerative Clustering methods. The data used is obtained from the BPS and includes indicators such as rice consumption per capita per year, rice production, rice price per kilogram, and population in each province. The clustering results aim to identify regions with similar characteristics to support the formulation of national food security policies in a more targeted manner.

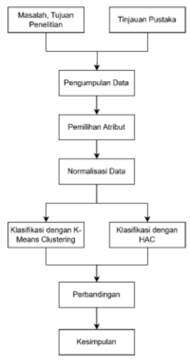

Figure 1 Flowchart of Research Phase

Problem and Research Objectives

This study begins with the identification of the problem and research objectives, which are to analyze food security based on rice consumption and production patterns in various provinces in Indonesia. The main focus of the research is to find rice distribution patterns to support more appropriate policy making.

Literature Review

Next, a literature review was conducted to examine previous research relevant to data classification in the context of food security. The methods used in this research, namely K-Means Clustering and Hierarchical Agglomerative Clustering, were chosen based on their effectiveness in grouping data based on certain patterns.

Data Collection

This stage involves collecting data from official sources, namely the Central Bureau of Statistics (BPS). The data collected included provincial variables, annual per capita rice consumption, rice production, rice price per kilogram, and population in each province.

1	Provinsi	Konsumsi Beras Setahun per kapita (Kg)	Harga Beras per Kg (Rp)	Produksi (Kg)	Jumlah Penduduk
2	Aceh	86,164	15670	947840746	5554800
3	Sumatera Utara	94,692	15230	1258983916	15588500
4	Sumatera Barat	78,572	17420	774543188	5836200
5	Riau	75,244	16220	126793810	6728100
6	Jambi	75,244	15830	160463591	2183300
7	Sumatera Selatan	84,188	15440	1661274064	3724300
8	Bengkulu	89,596	15070	155796522	8837300
9	Lampung	81,952	15990	1593859440	1531500
10	Kepulauan Riau	72,696	15890	44246670	2112200
11	Kepulauan Bangka Belitung	64,428	17310	174206	9419600
12	DKI Jakarta	68,588	16310	1317034	10684900
13	Jawa Barat	80,132	16650	4925948429	50345200
14	Jawa Tengah	68,796	16360	5076930616	12431400
15	DI Yogyakarta	62,66	16330	258566941	37892300
16	Jawa Timur	74,828	15760	5293418551	3759500
17	Banten	83,2	16570	885405996	41814500
18	Bali	96,148	16860	362855283	5695500
19	Nusa Tenggara Barat	99,112	16580	829896179	2809700
20	Nusa Tenggara Timur	107,432	16440	404149540	4273400

Figure 2 Pieces of Research Dataset

Attribute Selection

After the data was collected, the most relevant attributes for classification analysis were selected. The selected attributes include rice consumption per capita per year, rice production, rice price per kilogram, and population in each province.

Table 1 Selected Attributes

No	Selected Attributes
1	Annual Rice Consumption per capita (Kg)
2	Rice Price per Kg (Rp)
3	Production (Kg)
4	Total Population

Data Normalization

The data obtained has a different scale, so it is necessary to normalize it to ensure that there are no variables that dominate in the clustering process. Normalization is done so that the K-Means and GMM methods can provide optimal results. Examples of the top 5 (five) data from the research dataset are listed in table 2.

Table 2 Pieces of Data Normalization Results

Annual Rice Consumption per capita (Kg)	Price of Rice per Kg (Rp)	Production (Kg)	Total Population
0.519782	-0.527737	0.107964	-0.165741
1.159956	-0.709259	0.332960	0.731621
-0.050129	0.194224	-0.017352	-0.140574
-0.299953	-0.300835	-0.485757	-0.060807
-0.299953	-0.461729	-0.461410	-0.467270

Clustering Using K-Means Clustering and Hierarchical Agglomerative Clustering

In this stage, the data is clustered using the K-Means Clustering method based on the similarity of rice consumption and production patterns in each province. This method works by grouping data into a number of clusters based on the distance to the nearest centroid, then updating the centroid position until the clustering result is stable. In addition to K-Means, classification was also performed using Hierarchical Agglomerative Clustering (HAC), which groups data hierarchically based on similarities between provinces. This method works by gradually combining provinces that have the most similar characteristics until the final cluster is formed. HAC uses a distance and linkage-based approach to determine the relationship between data, resulting in a clearer and more interpretative cluster structure than K-Means.

Comparison of Results

After the clustering process is complete, the results of the two methods are compared by looking at the comparison graph of the clustering results as well as the interpretation obtained from the data visualization. From the resulting graphs, we can analyze the distribution of clusters formed, the distribution pattern of provinces within each cluster, and how the two methods group provinces based on rice consumption patterns, rice prices, rice production, and population. The conclusions from this comparison help determine which method is more suitable in describing food security patterns in each province.

Conclusion

The final stage is to conclude the research results. The conclusion includes an interpretation of the clustering patterns formed, as well as recommendations for food security policies based on areas with similar characteristics. The results of this study are expected to serve as a reference in determining a more effective rice distribution strategy.

Results and Discussions

Installing the Library

This research uses the Python programming language to implement the K-Means and HAC algorithms. Some of the libraries required for the algorithm to run properly on Kaggle Notebook must be installed first before use, including the following:

```
import pandas as pd
```
Used for data manipulation and analysis in tabular form.

```
import numpy as np
```
Used for numerical computation and operations on multidimensional arrays.

```
import matplotlib.pyplot as plt
```
Used to create data visualizations such as graphs and charts.

```
import seaborn as sns
```
Used for statistical data visualization with a more aesthetic appearance than Matplotlib.

```
from sklearn.preprocessing import StandardScaler
```
Used to normalize data

```
from sklearn.metrics import silhouette_score
```
Used to evaluate clustering quality with Silhouette Score

```
from sklearn.cluster import KMeans
```
Used to implement the K-Means Clustering algorithm.

```
from scipy.cluster.hierarchy import dendrogram, linkage
```
Used to create dendrogram and calculate linkage in Hierarchical Clustering.

```
from sklearn.cluster import AgglomerativeClustering
```
Used to implement Hierarchical Agglomerative Clustering (HAC).

Data Preparation

The first stage begins with displaying the data in csv form to be displayed in tabular form in a notebook as shown in Figure 3.1.

	Provinsi	Konsumsi Beras Setahun per kapita (Kg)	Harga Beras per Kg (Rp)	Produksi (Kg)	Jumlah Penduduk
0	Aceh	86,164	15670	1659966280	5554800
1	Sumatera Utara	94,692	15230	2204875510	15588500
2	Sumatera Barat	78,572	17420	1356467930	5836200
3	Riau	75,244	16220	222055710	6728100
4	Jambi	75,244	15830	281022050	2183300

Figure 3 Dataset Pieces in Dataframe Form

Data Transformation

Then change the columns that are not yet numeric into numeric form so that normalization can be done. After that, select numeric columns as attributes for normalization, namely Annual Rice Consumption per capita (Kg), Rice Price per Kg (Rp), Production (Kg), and Population.

	Konsumsi Beras Setahun per kapita (Kg)	Harga Beras per Kg (Rp)	Produksi (Kg)	Jumlah Penduduk
0	0.519782	-0.527737	0.107964	-0.165741
1	1.159956	-0.709259	0.332960	0.731621
2	-0.050129	0.194224	-0.017352	-0.140574
3	-0.299953	-0.300835	-0.485757	-0.060807
4	-0.299953	-0.461729	-0.461410	-0.467270

Figure 4 Transformed Dataframe Slice

Clustering Using K-Means Clustering

The stage starts from determining the number of clusters for the K-Means Clustering method using the Elbow Method whose results are as shown in Figure 5, and using Silhouette Score to assess the optimal cluster score based on the Elbow Method as shown in Figure 6. The results of the two graphs presented show that the optimal cluster used is as many as 3 clusters for further use in the K-Means Clustering method.

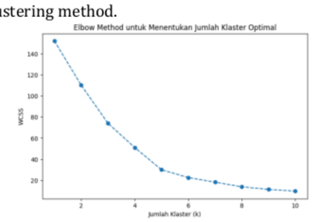

Figure 5 Elbow Method Results

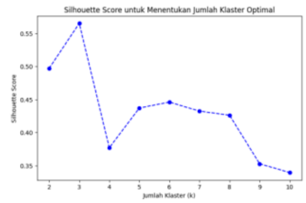

Figure 6 Silhouette Score Results

Clustering is done with the K-Means method, the results of which are as shown in Figure 7.

	Provinsi	Konsumsi Beras Setahun per kapita (Kg)	Harga Beras per Kg (Rp)	Produksi (Kg)	Jumlah Penduduk	Cluster_KMeans
0	Aceh	86.164	15670	947840746	5554800	1
1	Sumatera Utara	94.692	15230	1258983916	15588500	1
2	Sumatera Barat	78.572	17420	774543188	5836200	1
3	Riau	75.244	16220	126793810	6728100	1
4	Jambi	75.244	15830	160463591	2183300	1
5	Sumatera Selatan	84.188	15440	1661274064	3724300	1
6	Bengkulu	89.596	15070	155796522	8837300	1
7	Lampung	81.952	15990	1593859440	1531500	1
8	Kepulauan Riau	72.696	15890	44246670	2112200	1
9	Kepulauan Bangka Belitung	64.428	17310	174206	9419600	1
10	DKI Jakarta	68.588	16310	1317034	10684900	1
11	Jawa Barat	80.132	16650	4925948429	50345200	3
12	Jawa Tengah	68.796	16360	5076930616	12431400	3
13	DI Yogyakarta	62.660	16330	258566941	37892300	1
14	Jawa Timur	74.828	15760	5293418551	3759500	3
15	Banten	83.200	16570	885405996	41814500	1
16	Bali	96.148	16860	362855283	5695500	1
17	Nusa Tenggara Barat	99.112	16580	829896179	2809700	1
18	Nusa Tenggara Timur	107.432	16440	404149540	4273400	1
19	Kalimantan Barat	84.188	17140	436691750	4045900	1
20	Kalimantan Tengah	79.092	17250	209069834	739800	1
21	Kalimantan Selatan	76.232	16600	587883288	2701800	1
22	Kalimantan Timur	71.812	17520	142546096	1227800	1
23	Kalimantan Utara	75.400	17190	17175549	3121800	1
24	Sulawesi Utara	91.520	15370	155960051	9463400	1
25	Sulawesi Tengah	94.068	16100	435065679	1503200	1
26	Sulawesi Selatan	90.948	15720	2751323182	2793100	1
27	Sulawesi Tenggara	92.404	15890	317382402	4433300	1
28	Gorontalo	91.104	17000	134106705	5646000	1
29	Sulawesi Barat	100.412	14660	182078533	5656000	1
30	Maluku	81.848	17430	52032575	1945600	1
31	Maluku Utara	78.728	17860	17834014	1355600	1
32	Papua	68.120	15840	11836345	1090000	1
33	Papua Barat	66.040	16980	564513	569900	1
34	Papua Selatan	66.872	16000	2632281	545900	1
35	Papua Tengah	52.104	27810	124357873	1360000	2
36	Papua Pegunungan	60.268	25340	3467329	1466700	2
37	Papua Barat Daya	46.280	18440	24199	616100	1

Figure 7 Clustering Results with K-Means Clustering

The visualization of the relationship between attributes can be seen in Figure 8.

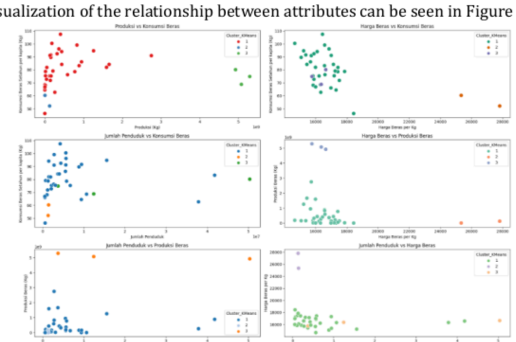

Figure 8 Visualization of Relationship Between Attributes with K-Means Clustering

Journal of Intelligent Decision Support System (IDSS) ISSN 2721-5792 (Online)
Vol. x, No. x, Mount 20xx, pp. xx-xx

Clustering Using Hierarchical Agglomerative Clustering

The stage starts from determining the number of clusters for the HAC method by using the linkage = single method which results in a dendrogram as shown in Figure 9. The dendrogram was evaluated with the Cophenetic Correlation Coefficient (CCC) to assess the optimal cluster score. The results of the dendrogram presented show that the optimal cluster used is 3 clusters and the CCC score obtained is 0.9167 which indicates that by using 3 clusters based on the dendrogram, HAC is very good and can be used to determine the optimal cluster with high confidence.

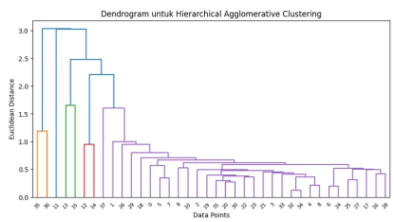

Figure 9 Dendrogram for Hierarchical Agglomerative Clustering

Furthermore, clustering is carried out with the HAC method, the results of which are as shown in Figure 10.

	Provinsi	Konsumsi Beras Setahun per kapita (Kg)	Harga Beras per Kg (Rp)	Produksi (Kg)	Jumlah Penduduk	Cluster_HAC
0	Aceh	86.164	15670	1659966280	5554800	1
1	Sumatera Utara	94.692	15230	2204875510	15588500	1
2	Sumatera Barat	78.572	17420	1356467930	5836200	1
3	Riau	75.244	16220	222055710	6728100	1
4	Jambi	75.244	15830	281022050	2183300	1
5	Sumatera Selatan	84.188	15440	2909411670	3724300	1
6	Bengkulu	89.596	15070	272848550	8837300	1
7	Lampung	81.952	15990	2791347530	1531500	1
8	Kepulauan Riau	72.696	15890	77489790	2112200	1
9	Kepulauan Bangka Belitung	64.428	17310	305090	9419600	1
10	DKI Jakarta	68.588	16310	2306540	10684900	1
11	Jawa Barat	80.132	16650	8626879910	50345200	3
12	Jawa Tengah	68.796	16360	8891297050	12431400	1
13	DI Yogyakarta	62.660	16330	452831770	37892300	1
14	Jawa Timur	74.828	15760	9270435290	3759500	1
15	Banten	83.200	16570	1550623460	41814500	1
16	Bali	96.148	16860	635473350	5695500	1
17	Nusa Tenggara Barat	99.112	16580	1453408370	2809700	1
18	Nusa Tenggara Timur	107.432	16440	707792540	4273400	1
19	Kalimantan Barat	84.188	17140	764784150	4045900	1
20	Kalimantan Tengah	79.092	17250	366146820	739800	1
21	Kalimantan Selatan	76.232	16600	1029567930	2701800	1
22	Kalimantan Timur	71.812	17520	249642900	1227800	1
23	Kalimantan Utara	75.400	17190	30079770	3121800	1
24	Sulawesi Utara	91.520	15370	273134940	9463400	1
25	Sulawesi Tengah	94.068	16100	761936390	1503200	1
26	Sulawesi Selatan	90.948	15720	4818429390	2793100	1
27	Sulawesi Tenggara	92.404	15890	555836080	4433300	1
28	Gorontalo	91.104	17000	234862880	5646000	1
29	Sulawesi Barat	100.412	14660	318876590	5656000	1
30	Maluku	81.848	17430	91125350	1945600	1
31	Maluku Utara	78.728	17860	31232950	1355600	1
32	Papua	68.120	15840	20729150	1090000	1
33	Papua Barat	66.040	16980	988640	569900	1
34	Papua Selatan	66.872	16000	4609950	545900	1
35	Papua Tengah	52.104	27810	217789620	1360000	2
36	Papua Pegunungan	60.268	25340	6072380	1466700	2
37	Papua Barat Daya	46.280	18440	42380	616100	1

Figure 10 Clustering Results with Hierarchical Agglomerative Clustering

The visualization of the relationship between attributes can be seen in Figure 11.

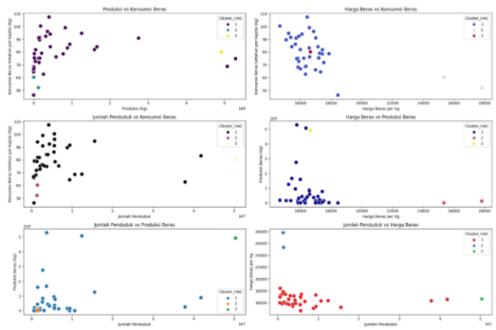

Figure 11 Visualization of Relationships Between Attributes with Hierarchical Agglomerative Clustering

Comparison of Results

Based on the clustering results using 2 (two) different methods, K-Means Clustering and Hierarchical Agglomerative Clustering recommend the same number of clusters, namely 3 (three) clusters. is Province data by cluster listed in table 3 and analysis of each cluster are listed in table 4.

Table 3 Province Data by Cluster

	K-Means	HAC
Number of Clusters	3	3
Cluster 1	Other than Clusters 2 and 3	Other than Clusters 2 and 3
Cluster 2	• Papua Tengah • Papua Pegunungan	• Papua Tengah • Papua Pegunungan
Cluster 3	• Jawa Barat • Jawa Tengah • Jawa Timur	• Jawa Barat

Table 4 Analysis of Attributes in Each Cluster

Mean of	Cluster	K-Means	HAC	Analysis
Rice Consumption (Kg per capita per year)	Cluster 1	81.06	80.53	Moderate consumption, most provinces
	Cluster 2	56.19	56.19	Lowest consumption
	Cluster 3	74.59	80.13	Lower consumption than Cluster 1, possible availability of food other than rice
Rice Price (IDR per kg)	Cluster 1	16,429	16,407	Medium price, reflecting national average price
	Cluster 2	26,575	26,575	Most expensive price
	Cluster 3	16,257	16,650	Medium price, reflecting national average price
Rice Production (Tons)	Cluster 1	452,133	722,592	Moderate production, most provinces
	Cluster 2	63,913	63,913	Very low production
	Cluster 3	5,098,766	4,925,948	Highest production
Population (people)	Cluster 1	6.5 million	6.5 million	Medium population
	Cluster 2	2.1 million	1.4 million	Small population
	Cluster 3	37 million	50 million	Largest population

Based on the previously mentioned data, the USDA states that total rice consumption in Indonesia reaches 36.5 million tons, while BPS estimates national rice production at 30.62 million tons.

From this comparison, it can be concluded that Indonesia has a rice deficit of 5.88 million tons, which may have to be met through imports or other food security policies.

The clustering results show that both K-Means and HAC successfully form 3 (three) clusters, with relatively similar patterns, but have differences in the data segmentation approach:

1) In terms of rice consumption, both methods show that Cluster 2 has the lowest consumption, which includes provinces such as Central Papua and Papua Mountains, the conclusion can be drawn for cluster 2 areas either there are many other food sources besides rice or it is difficult to get rice.
2) In terms of rice prices, the K-Means and HAC methods show similar price differences. The price of rice in cluster 2 areas is very expensive, which may refer to the previous point on rice consumption that there is difficulty in obtaining rice, resulting in high demand and low availability.
3) In terms of rice production, Cluster 3 in both methods reflects the provinces with the highest rice production (West Java, Central Java, and East Java), while Cluster 2 has the lowest production, indicating areas that have little agricultural land and depend on supplies from other areas.
4) In terms of population, K-Means is more accurate in reflecting the population distribution, where Cluster 3 includes the provinces with the largest population (West Java, Central Java, and East Java), while HAC only includes one province, West Java.

K-Means more accurately reflects the population distribution, where Cluster 3 includes provinces with the largest population such as West Java, Central Java and East Java. Meanwhile, HAC only includes West Java in Cluster 3, indicating a lack of precision in grouping provinces with large populations. In addition, HAC tends to produce clusters with higher variation in rice production, as seen in Cluster 1 which has greater production than K-Means. Overall, K-Means is more effective in describing the distribution of population and rice production, while HAC shows weaknesses in segmenting high-population provinces.

Conclusions

Both K-Means Clustering and Hierarchical Agglomerative Clustering (HAC) are able to cluster Indonesian provinces based on consumption, price, production, and population. HAC is more effective in distinguishing rice prices, especially in high-price areas, while K-Means more accurately reflects the distribution of rice population and production. This clustering provides a scientific basis for optimizing rice distribution and data-driven food security policies to improve distribution efficiency, price stabilization, and production management. Future research could consider weather, infrastructure, and logistics factors to improve clustering accuracy and more adaptive and sustainable policies.

References

Adriani, D., & Yustini, T. (2021). Anticipating the demographic bonus from the perspective of human capital in Indonesia. *International Journal of Research in Business and Social Science (2147-4478)*, *10*(6), 141–152.

Ahmad, A., & Khan, S. S. (2021). initKmix-A novel initial partition generation algorithm for clustering mixed data using k-means-based clustering. *Expert Systems with Applications*, *167*, 114149. https://doi.org/https://doi.org/10.1016/j.eswa.2020.114149

Ay, M., Özbakır, L., Kulluk, S., Gülmez, B., Öztürk, G., & Özer, S. (2023). FC-Kmeans: Fixed-centered K-means algorithm. *Expert Systems with Applications*, *211*, 118656.

BPS. (2024, June 28). *Jumlah Penduduk Pertengahan Tahun (Ribu Jiwa), 2022-2024*. BPS. https://www.bps.go.id/id/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun--ribu-jiwa-.html

BPS. (2025, February 3). *Pada 2024, luas panen padi mencapai sekitar 10,05 juta hektare dengan produksi padi sebanyak 53,14 juta ton gabah kering giling (GKG).* BPS. https://www.bps.go.id/id/pressrelease/2025/02/03/2414/pada-2024--luas-panen-padi-mencapai-sekitar-10-05-juta-hektare-dengan-produksi-padi-sebanyak-53-14-juta-ton-gabah-kering-giling--gkg--.html

Chhabra, A., & Mohapatra, P. (2022). Fair Algorithms for Hierarchical Agglomerative Clustering. *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 206–211. https://doi.org/10.1109/ICMLA55696.2022.00036

Dwi Ariyanti, S., Nabila, U., & Rahmawati, L. (2024). Pemenuhan Kebutuhan Produksi Beras Nasional Dalam Meningkatkan Kesejahteraan Masyarakat Menurut Perspektif Ekonomi Islam. *Jurnal Ekonomi Syariah Dan Bisnis*, *7*(1). https://doi.org/10.31949/maro.v7i1.9121

Kasote, D., Sreenivasulu, N., Acuin, C., & Regina, A. (2022). Enhancing health benefits of milled rice: current status and future perspectives. *Critical Reviews in Food Science and Nutrition*, *62*(29), 8099–8119.

Marzani, Y., & Juliannisa, I. A. (2024). *ASSESSMENT KETERSEDIAAN BERAS PADA 34 PROVINSI DI INDONESIA Assessment Of Rice Availability In 34 Provinces In Indonesia*. *8*(2), 700–711. https://doi.org/10.21776/ub.jepa.2024.008.02.25

Mavroeidis, A., Roussis, I., & Kakabouki, I. (2022). The role of alternative crops in an upcoming global food crisis: A concise review. *Foods*, *11*(22), 3584.

Mawarni, Q. I., & Budi, E. S. (2022). Implementasi Algoritma K-Means Clustering Dalam Penilaian Kedisiplinan Siswa. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *3*(4), 522. https://doi.org/10.30865/json.v3i4.4242

Mirantika, N., Tsamratul'ain, A., & Diviana Agnia, F. (2021). *PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK PENGELOMPOKAN PENYEBARAN COVID-19 DI PROVINSI JAWA BARAT*. *15*. https://journal.uniku.ac.id/index.php/ilkom

Mohidem, N. A., Hashim, N., Shamsudin, R., & Che Man, H. (2022). Rice for Food Security: Revisiting Its Production, Diversity, Rice Milling Process and Nutrient Content. *Agriculture*, *12*(6), 741. https://doi.org/10.3390/agriculture12060741

Monath, N., Dubey, K. A., Guruganesh, G., Zaheer, M., Ahmed, A., McCallum, A., Mergen, G., Najork, M., Terzihan, M., & Tjanaka, B. (2021). Scalable hierarchical agglomerative clustering. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1245–1255.

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive Review of K-Means Clustering Algorithms. *International Journal of Advances in Scientific Research and Engineering*, *07*(08), 64–69. https://doi.org/10.31695/ijasre.2021.34050

Priambodo, E. P., & Jananto, A. (2022). Perbandingan Analisis Cluster Algoritma K-Means Dan AHC Dalam Perencanaan Persediaan Barang Pada Perusahaan Manufaktur. *Progresif: Jurnal Ilmiah Komputer*, *18*(2), 257. https://doi.org/10.35889/progresif.v18i2.868

Rahim, M. S., Nguyen, K. A., Stewart, R. A., Ahmed, T., Giurco, D., & Blumenstein, M. (2021). A clustering solution for analyzing residential water consumption patterns. *Knowledge-Based Systems*, *233*, 107522. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107522

Rozaki, Z. (2021). Food security challenges and opportunities in Indonesia post COVID-19. *Advances in Food Security and Sustainability*, *6*, 119–168.

Rusadi Akhmad, G. (2023). Proyeksi Kebutuhan dan Ketersediaan Beras di Provinsi DIY Tahun 2045. *Jurnal Pendidikan Geografi Undiksha*, *11*(2), 94–104. https://doi.org/10.23887/jjpg.v11i2.66394

Saefudin. (2023). STRATEGI PERENCANAAN MENGHADAPI KRISIS PANGAN DAN EL NINO. *WARTA BSIP PERKEBUNAN*, *1*(3).

Setiani, S. Y., Pratiwi, T., & Fitrianto, A. R. (2021). Tenaga Muda Pertanian dan Ketahanan Pangan di Indonesia. *CAKRAWALA*, *15*(2), 95–108. https://doi.org/10.32781/cakrawala.v15i2.386

Sintiya, E. S. (2023). Analisis Ketersediaan Beras Menggunakan Sistem Dinamik Sebagai Pendukung Kebijakan Ketahanan Pangan Endah Septa Sintiya. *TECNOSCIENZA*, *7*(2).

United States Department of Agriculture. (2024). *Grain and Feed Update*.

Yu, H., & Hou, X. (2022). Hierarchical clustering in astronomy. *Astronomy and Computing*, *41*, 100662.

Zubair, M., Iqbal, M. D. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2024). An improved K-means clustering algorithm towards an efficient data-driven modeling. *Annals of Data Science*, *11*(5), 1525–1544.

Check Plagiarism

ORIGINALITY REPORT

19%	15%	8%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.untag-sby.ac.id Internet Source	3%
2	api.uinjkt.ac.id Internet Source	2%
3	www.idss.iocspublisher.org Internet Source	1%
4	Submitted to University of Anbar Student Paper	1%
5	journal.uniku.ac.id Internet Source	1%
6	Submitted to Brookdale Community College Student Paper	1%
7	www.catalyzex.com Internet Source	1%
8	Bhawna Singh. "Chapter 1 Introduction to Large Language Models", Springer Science and Business Media LLC, 2024 Publication	1%
9	www.essays.se Internet Source	1%
10	idss.iocspublisher.org Internet Source	<1%
11	journal.ugm.ac.id Internet Source	<1%

12	"Climate Crisis, Social Responses and Sustainability", Springer Science and Business Media LLC, 2024 Publication	<1%
13	Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming. "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data", Information Sciences, 2023 Publication	<1%
14	www.mdpi.com Internet Source	<1%
15	Muh. Hizbul Zainul Muttaqim, Ruliana Ruliana, Zulkifli Rais. "Application of K-Medoids Algorithm in Provincial Grouping in Indonesia Based On Case of Environmental Pollution", SAINSMAT: Journal of Applied Sciences, Mathematics, and Its Education, 2023 Publication	<1%
16	trytechweb.wordpress.com Internet Source	<1%
17	Submitted to Erasmus University of Rotterdam Student Paper	<1%
18	Submitted to unibuc Student Paper	<1%
19	Marco M. Vlajnic, Steven J. Simske. "Accuracy and Performance of Machine Learning Methodologies: Novel Assessments of Country Pandemic Vulnerability based on Non-Pandemic Predictors", IEEE Access, 2023 Publication	<1%

20	www.docstoc.com Internet Source	<1 %
21	fidelity.nusaputra.ac.id Internet Source	<1 %
22	docs.mipro-proceedings.com Internet Source	<1 %
23	ejournal.unma.ac.id Internet Source	<1 %
24	ojs.stmik-banjarbaru.ac.id Internet Source	<1 %
25	www.researchgate.net Internet Source	<1 %
26	www.scielo.br Internet Source	<1 %
27	Dodit Ardiatma, Puji Lestari, Mochammad Chaerul. "Real data mapping of DKI Jakarta waste generation using the K-mean Clustering method at final disposal Bantargebang", E3S Web of Conferences, 2024 Publication	<1 %
28	docnum.univ-lorraine.fr Internet Source	<1 %
29	ejurnal.bppt.go.id Internet Source	<1 %
30	ijsshr.in Internet Source	<1 %
31	www.coursehero.com Internet Source	<1 %
32	www.notulaebotanicae.ro Internet Source	<1 %

33	Glykeria Kyrou, Vasileios Charilogis, Ioannis G. Tsoulos. "Refining the Eel and Grouper Optimizer with Intelligent Modifications for Global Optimization", Computation, 2024 Publication	<1 %
34	H Kurniawan, N Hidayatun, Kristamtini, S Widyayanti, A Risliawati. "SSR diversity on rice landraces collected from Yogyakarta Province", IOP Conference Series: Earth and Environmental Science, 2023 Publication	<1 %
35	Harold Kevin Alfredo. "Does inflation provide a more accurate expected return than sharia bonds?", JPPI (Jurnal Penelitian Pendidikan Indonesia), 2024 Publication	<1 %
36	Iqbal Ahmad Dahlan, Muhammad Bryan Gutomo Putra, Suhono Harso Supangkat, Fadhil Hidayat, Fetty Fitriyanti Lubis, Faqih Hamami. "Real-time passenger social distance monitoring with video analytics using deep learning in railway station", Indonesian Journal of Electrical Engineering and Computer Science, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		