ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

ENHANCING K-MEANS CLUSTERING MODEL TO IMPROVE RICE HARVEST PRODUCTIVITY AREAS IN ACEH UTARA USING PURITY

Sujacka Retno[1], Bustami[2], Rozzi Kesuma Dinata[3]

[1,2,3] Department of Informatics Engineering, Universitas Malikussaleh, Indonesia

email: sujacka@unimal.ac.id[1], bustami@unimal.ac.id[2], rozzi@unimal.ac.id[3]

Abstract

To optimize the performance of the clustering process using K-Means, an optimalization approach employing the Purity algorithm is needed. This research was tested on a dataset of rice harvest productivity areas in Aceh Utara Regency by comprehensively analyzing the number of iterations and the DBI values produced by K-Means and Purity K-Means in clustering priority and non-priority rice production areas. This is in line with the efforts of the Regional Government to implement rice production intensification programs in Aceh Utara Regency. From the testing of Purity K-Means, an average of 5, 2, 2, 5, and 3 iterations were obtained from all tested datasets sequentially from 2019 to 2023. Meanwhile, from the testing of conventional K-Means, the average number of iterations obtained was 5.4, 4.8, 4.2, 5.6, and 3.8 iterations, sequentially. This indicates that the clustering performance conducted by Purity K-Means is better than conventional K-Means. The DBI values obtained from Purity K-Means for the entire dataset sequentially are 0.6781, 0.4175, 0.4419, 0.6182, and 0.4973. This value is lower compared to the DBI values obtained from conventional K-Means, which are 0.7178, 0.6025, 0.4971, 0.7222, and 0.5519, respectively. This also indicates that the validity level of the clustering results performed by Purity K-Means is higher than conventional K-Means.

Keywords : Optimization, Purity, Clustering, K-Means, Davies-Bouldin Index

Received: 10-05-2024 | **Revised:** 10-07-2024 | **Accepted:** 21-07-2024
DOI: https://doi.org/10.23887/janapati.v13i2.78254

INTRODUCTION

The decreasing role of the agricultural sector is partly due to the conversion of agricultural land into non-agricultural land. Infrastructure development, housing, and industrial areas have reduced the amount of agricultural land. The continuous reduction of agricultural land raises concerns about future food availability [1]. Efforts to expand agricultural land will be in vain if not accompanied by an increase in the quality of rice produced. Intensification efforts in rice harvesting, through increasing land productivity, are one way to improve the quality of rice produced. A land survey is conducted using tools measuring 2.5 m x 2.5 m to calculate the productivity of harvested crops. Rice productivity is the rice production calculated per unit area of land. Rice productivity is calculated based on the amount of rice production in the form of Milled Dry Grain (MDG) per unit area of land, which is measured in quintals per hectare (kw/ha). Rice productivity is a value that indicates the average production yield per unit area per commodity of rice in a one-year reporting period [2].

Based on data obtained from the Department of Agriculture and Food of Aceh Utara Regency, a region is considered productive by reviewing various aspects including land area planted (ha), harvested area (ha), productivity (kw/ha), production (tons), and production percentage (%) [3]. There are a total of 26 districts in Aceh Utara Regency, where the majority of the population's occupation is rice farming. Therefore, to maximize efforts to intensify rice harvest productivity in these areas, a model needs to be developed to analyze the determination of areas in Aceh Utara Regency that have the potential for intensification by applying the development of an algorithm model that can optimize the clustering process of rice harvest productivity areas in Aceh Utara Regency. This research has a significant impact on the Aceh Utara Regency Local Government in determining policy directions for the management of resources such as increasing the amount of fertilizer subsidies for priority areas in the future.

Clustering is a data analysis technique aimed at grouping similar objects into clusters based on the similarity of their characteristics. In clustering, there are no predefined labels or categories given to the objects, so the main goal is to find patterns or structures hidden within the data [4]. By applying various algorithms such as

K-Means, Hierarchical Clustering, or DBSCAN, clustering helps identify relationships and patterns that may not be immediately apparent, allowing for a better understanding of complex data [5].

Previous research on K-Means clustering of rice harvest productivity areas in Jawa Timur with 3 clusters was able to cluster the areas in 5 iterations [6]. However, the attributes used in this reseach were only 2, namely Production Area and Harvest Yield, leading to less effective clustering results. Meanwhile, in a similar research conducted in Sumatra Utara in 2022 [7], K-Means successfully clustered the productivity areas, also using the same 2 attributes, namely harvested area and production. However, this research did not provide detailed information about the number of iterations generated. In another related research [8], K-Means was used to cluster rice harvest areas in Indonesia with a different number of attributes from the previous 2 researchers, using 3 attributes: harvested area, production, and rice productivity. The data was obtained from the Indonesian Central Statistics Agency (BPS) in 2022. The number of clusters formed was 3.

Based on the analysis of the previous research, several factors need to be added to the assessment of similar research. These include the addition of several other supporting attributes such as the number of districts/villages in the regency/city or the location to be studied, planted land area, harvested land area, harvest productivity per hectare, annual production quantity, and harvest yield percentage. Another factor to consider is the need for testing the evaluation results of the K-Means clustering process itself. The evaluation of the clustering results can be done by measuring the Davies-Bouldin Index value of the clustering process to determine how well the clustering algorithm performs [9].

The purpose of this research is to maximize the performance of the K-Means algorithm, particularly in clustering rice harvest productivity areas. This research will compare the testing conducted by the conventional K-Means method with the optimized K-Means method using Purity. The testing results of both models will then be evaluated for their performance using the Davies-Bouldin Index (DBI) to determine the performance outcome of both methods [10].

The innovation or novelty of this research lies in the development of the K-Means model with the assistance of the Purity algorithm in determining the best initial centroids to be used in this reserach. As we all know, the performance of the K-Means algorithm heavily relies on the initial centroids chosen. Therefore, in this research, besides adding several attributes in the data testing, modifications to the K-Means algorithm are also made in an effort to discover a new clustering model useful for future researchers investigating similar topics.

METHOD

This research employs various methods for its testing, including conventional K-Means and Purity K-Means, as well as performance testing of the algorithm using the Davies-Bouldin Index as its measure. The flowchart of the conducted research can be seen in Figure 1.

Figure 1 illustrates the steps taken in this research. The first step in this research is to collect and input the dataset, which will then be processed for clustering by two algorithm models, namely K-Means and Purity K-Means. The dataset used in this research is the rice harvest productivity dataset in Aceh Utara Regency from 2019 to 2023.

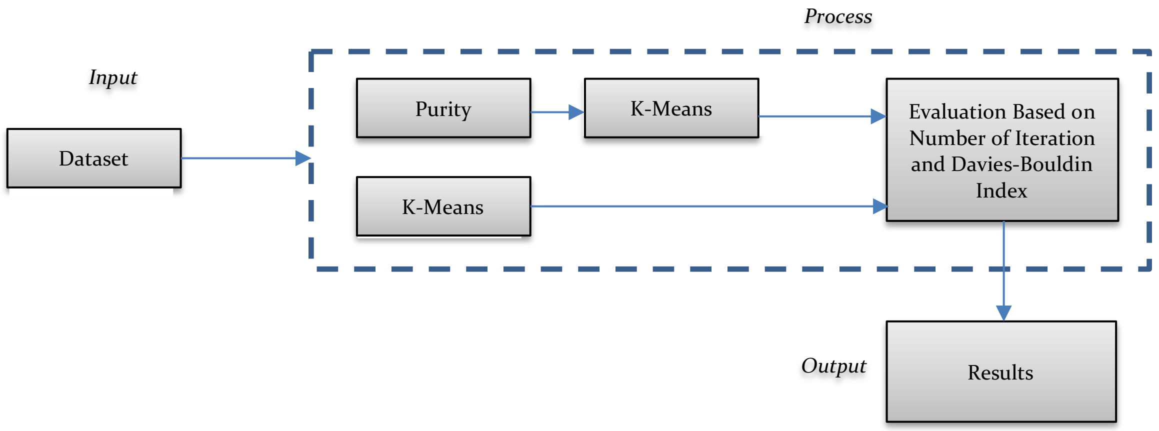
Process
Input
Dataset
Purity
K-Means
K-Means
Evaluation Based on Number of Iteration and Davies-Bouldin Index
Output
Results

Figure 1. Research Workflow Diagram

After the clustering process is completed with both methods, the number of iterations generated will be analyzed, and the testing will be evaluated using the DBI to determine which model is more effective and efficient in clustering the rice harvest productivity data [11].

Dataset

The dataset used in this study consists of rice harvest productivity data in Aceh Utara Regency over the past 5 years (2019 to 2023), 130 data in total, obtained from the Department of Agriculture and Food of Aceh Utara Regency. This dataset has 7 attributes, including district, number of villages, cultivated area (ha), harvested area (ha), harvest productivity (kw/ha), production quantity (tons), and production percentage (%). The details of the dataset can be seen in Table 1 to Table 5.

Table 1. Rice Harvest Productivity Dataset for the Year 2019

Districts	Num of Village	Planted Area (ha)	Harvested Area (ha)	Harvest Yield (kw/ha)	Total Production (tons)	Production Percentage (%)
Baktiya	57	9038	8726	52.18	45532.268	96.5
Baktiya Barat	26	3831	3548	50.74	18002.552	92.6
Banda Baro	9	1229	1033	43.73	4517.309	84.1
Cot Girek	24	475	493	40.07	1975.451	104
Dewantara	15	1130	1152	43.64	5027.328	102
Geureudong Pase	11	387	442	46.04	2034.968	114
Kuta Makmur	39	2136	2915	50.06	14592.49	136
Langkahan	23	1510	1514	50.67	7671.438	100
Lapang	11	554	769	43.37	3335.153	139
Lhoksukon	75	4722	4249	53.59	22770.391	90
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tanah Pasir	18	1435	1414	43.38	6133.932	98.5

Table 2. Rice Harvest Productivity Dataset for the Year 2020

Districts	Num of Village	Planted Area (ha)	Harvested Area (ha)	Harvest Yield (kw/ha)	Total Production (tons)	Production Percentage (%)
Baktiya	57	5243	5696	50.37	28689.74	109
Baktiya Barat	26	3256	3182	50.74	16143.95	97.7
Banda Baro	9	1302	1174	45.37	5324.17	90.2
Cot Girek	24	562	556	40.65	2259.73	98.9
Dewantara	15	872	858	45.78	3927.47	98.4
Geureudong Pase	11	416	416	46.04	1917.11	100
Kuta Makmur	39	3257	2875	50.06	14391.25	88.3
Langkahan	23	2651	2651	50.43	13367.98	100
Lapang	11	980	806	43.37	3496.06	82.2
Lhoksukon	75	7450	6665	51.63	34412.43	89.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tanah Pasir	18	906	928	43.38	4026.53	102

Table 3. Rice Harvest Productivity Dataset for the Year 2021

Districts	Num of Village	Planted Area (ha)	Harvested Area (ha)	Harvest Yield (kw/ha)	Total Production (tons)	Production Percentage (%)
Baktiya	57	10040	9750	51.05	49771.71	97.1
Baktiya Barat	26	4327	4427	51.47	22785.77	102
Banda Baro	9	2186	2148	48.52	10421.61	98.3
Cot Girek	24	1376	1351	48.05	6490.59	98.2
Dewantara	15	1480	1809	48.62	8797.30	122
Geureudong Pase	11	537	467	49	2288.30	87
Kuta Makmur	39	3969	3887	53.57	20824.27	97.9
Langkahan	23	2940	2884	52.04	15007.30	98.1
Lapang	11	1120	917	45.64	4182.91	81.9
Lhoksukon	75	6990	6666	52.19	34789.33	95.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tanah Pasir	18	1357	1476	48.04	7092.15	109

Table 4. Rice Harvest Productivity Dataset for the Year 2022

Districts	Num of Village	Planted Area (ha)	Harvested Area (ha)	Harvest Yield (kw/ha)	Total Production (tons)	Production Percentage (%)
Baktiya	57	9142	9142	55.41	50655.27	100
Baktiya Barat	26	3995	3995	55.55	22193.89	100
Banda Baro	9	1548	1538	46.37	7131.24	99.3
Cot Girek	24	775	771	48.41	3734.35	99.5
Dewantara	15	713	713	50.03	3565.64	100
Geureudong Pase	11	197	123	47.63	586.33	62.5
Kuta Makmur	39	2925	3005	50.72	15239.33	103
Langkahan	23	2368	2418	55.17	13340.11	102
Lapang	11	902	902	50.11	4520.92	100
Lhoksukon	75	8649	8649	56.05	48474.84	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tanah Pasir	18	1038	1060	50.52	5354.61	102

Table 5. Rice Harvest Productivity Dataset for the Year 2023

Districts	Num of Village	Planted Area (ha)	Harvested Area (ha)	Harvest Yield (kw/ha)	Total Production (tons)	Production Percentage (%)
Baktiya	57	11341	9818	51.05	50120.89	86.6
Baktiya Barat	26	4740	3086	51.47	15883.64	65.1
Banda Baro	9	1568	1456	48.52	7064.51	92.9
Cot Girek	24	1195	1388	48.05	6669.34	116
Dewantara	15	1632	1008	48.62	4900.90	61.8
Geureudong Pase	11	636	504	49	2469.60	79.2
Kuta Makmur	39	3444	3051	53.57	16344.21	88.6

Langkahan	23	3005	2888	51.52	14878.98	96.1
Lapang	11	942	1050	45.64	4792.20	111
Lhoksukon	75	4260	4260	52.19	22232.94	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tanah Pasir	18	1245	1090	48.04	5236.36	87.6

The attached dataset in Table 1 to Table 5 represents rice harvest productivity data in Aceh Utara Regency over the past 5 years. This data will be clustered using the K-Means and Purity K-Means algorithms for each year. In addition to measuring algorithm performance, this is done with the aim of identifying areas consistently located in priority clusters. This information can be considered by the Aceh Utara Regency Local Government in determining the policy directions for resource management, such as increasing the amount of fertilizer subsidies for priority areas in the future.

K-Means

K-Means is a popular data clustering method in data analysis and machine learning. Its goal is to divide a dataset into several clusters based on attribute similarity [12]. The algorithm works by randomly initializing cluster centers, then iterating to update the cluster center positions until convergence. In each iteration, data points are labeled according to the cluster whose center is closest to them, and cluster centers are updated using the average of the data points within the cluster. This process continues until there are no more changes in the cluster center positions or the specified number of iterations has been reached [13].

K-Means algorithm is efficient in handling large datasets and suitable for data with clear structures in their attribute space. However, K-Means is sensitive to the initial cluster center initialization and can produce different solutions depending on that initialization. Therefore, multiple attempts with different initializations are often required to achieve optimal results [14].

As for the conventional K-Means scheme, it can be seen in Figure 2.

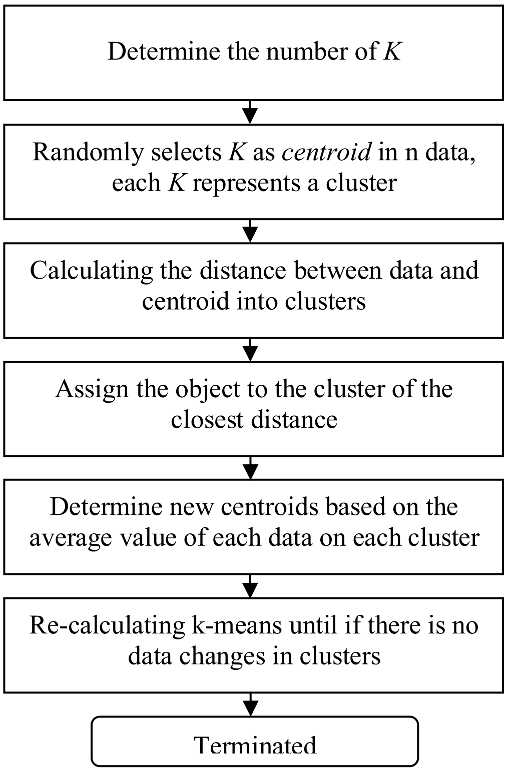
Determine the number of *K*
Randomly selects *K* as *centroid* in n data, each *K* represents a cluster
Calculating the distance between data and centroid into clusters
Assign the object to the cluster of the closest distance
Determine new centroids based on the average value of each data on each cluster
Re-calculating k-means until if there is no data changes in clusters
Terminated

Figure 2. Workflow of the K-Means algorithm

Based on Figure 2, the followings are procedure of the conventional K-Means [15]:

1. Determine the K value;
2. Randomly select K in n data for the initial centroid;
3. Calculate the distance of each data to the initial centroid by using a distance measure;

$$D_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(xi-yi)^2} \qquad (1)$$

   where, *yi* is the value of the centroid of *i-th* attribute.
4. Group the data to each cluster based on the minimum value of the distance;
5. Determine new centroids based on the average value of the data on each cluster;
6. Repeat steps 3-5 and stop the process until there is no data migrated from each cluster.

Proposed Method (Purity K-Means)

To enhance K-Means clustering performance, it's important to establish the initial centroid. This study concentrates on determining them through the subsequent process: Initially, we compute the Purity value for the entire dataset. Then, we utilize the minimum and maximum purity values as the initial centroids for K-Means clustering. Thus, the formula below is used to compute the Purity value [16]:

$$Purity(p) = \frac{1}{N_j} max(n_{ij}) \qquad (2)$$

where, n_{ij} is the amount of data in j-th cluster; and j is the index of the cluster.

As for the proposed method scheme in this research, it can be seen in Figure 3.

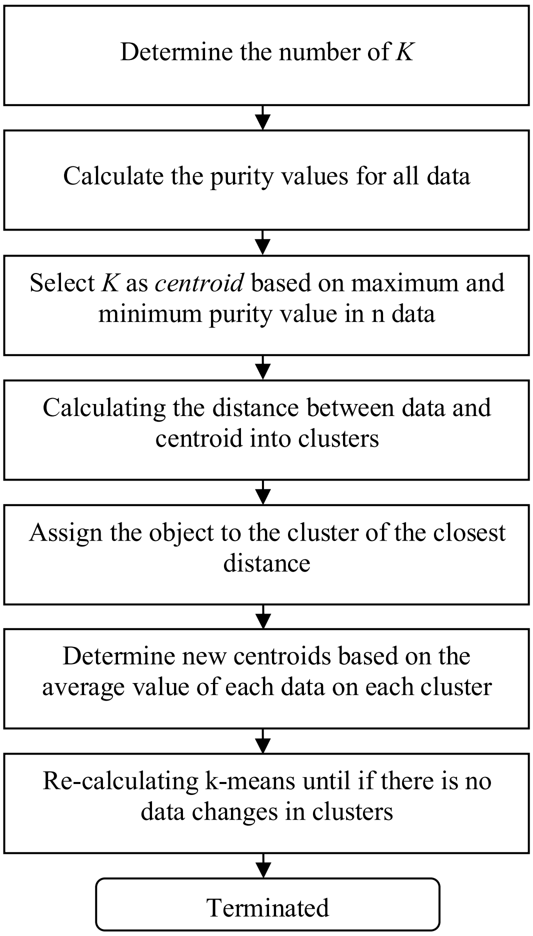
Determine the number of *K*
Calculate the purity values for all data
Select *K* as *centroid* based on maximum and minimum purity value in n data
Calculating the distance between data and centroid into clusters
Assign the object to the cluster of the closest distance
Determine new centroids based on the average value of each data on each cluster
Re-calculating k-means until if there is no data changes in clusters
Terminated

Figure 3. Workflow of the Purity K-Means algorithm

Based on Figure 3, the followings are steps of the proposed Purity K-Means algorithm [17]:

1. Determine the K value, where K is the number of clusters;
2. Calculate the purity value of each data by formula (2);
3. Initialize K based on the maximum and minimum Purity value as the initial centroid in n data;
4. Calculate the distance of each data to the initial centroid by using formula (1);
5. Group the data to each cluster based on the minimum value of the distance;
6. Determine new centroids based on the average value of the data on each cluster;
7. Repeat steps 4-6 and stop the process until there is no data migrated from each cluster [18].

Davies-Bouldin Index (DBI)

In this research, the Davies-Bouldin Index (DBI) was used to assess the cluster validity of both original K-Means and Purity K-Means. A smaller DBI value indicates better clustering validity [19]. The following steps were utilized to calculate the DBI [20]:

1. Calculate SSW (Sum of Squares Within cluster) by using the formula:

$$SSWi = \frac{1}{mi}\sum_{j=i}^{mi} d(xj, ci) \qquad (3)$$

where, d(xj,ci) is the distance of each data to the centroid.

2. Calculate SSB (Sum of Squaes Between cluster) by using the formula:

$$SSBi,j = d(ci, cj) \qquad (4)$$

3. Calculate the Ratio by using the formula:

$$Rij = \frac{SSWi + SSWj}{SSBij} \qquad (5)$$

4. Determine the Davies Bouldin Index by using the formula:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i\neq j}(R_{i,j}) \qquad (6)$$

where, *k* is the number of clusters [21].

RESULT AND DISCUSSION

Measure the Purity

Formula (2) was applied to calculate the Purity value. The Purity calculations for the data used in this research are presented below:

1. $Purity(1{,}2019) = \frac{1}{N_j} max(n_{ij})$
$= \frac{1}{(63599)}(45532.268)$
$= 0.715932555$
2. $Purity(2{,}2019) = \frac{1}{N_j} max(n_{ij})$
$= \frac{1}{(25644)}(18002.552)$
$= 0.702031295$
3. $Purity(3{,}2019) = \frac{1}{N_j} max(n_{ij})$
$= \frac{1}{(7000)}(4517.309)$
$= 0.64531666$
4. $Purity(4{,}2019) = \frac{1}{N_j} max(n_{ij})$
$= \frac{1}{(3215)}(1975.451)$
$= 0.61442911$
5. $Purity(5{,}2019) = \frac{1}{N_j} max(n_{ij})$
$= \frac{1}{(7572)}(5027.328)$
$= 0.663948726$

The comprehensive Purity value results for all datasets are presented in Table 6, 7, 8, 9, and 10, respectively.

Table 6. Purity Values for the 2019 Dataset

Data No-	X1	X2	X3	X4	X5	X6	Purity Value
1	57	9038	8726	52.18	45532.268	96.5	0.715932555
2	26	3831	3548	50.74	18002.552	92.6	0.702031295
3	9	1229	1033	43.73	4517.309	84.1	0.64531666
4	24	475	493	40.07	1975.451	104	0.61442911
5	15	1130	1152	43.64	5027.328	102	0.663948726
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	18	1435	1414	43.38	6133.932	98.5	0.663745952

Table 7. Purity Values for the 2020 Dataset

Data No-	X1	X2	X3	X4	X5	X6	Purity Value
1	57	5243	5696	50.37	28689.74	109	0.718080309
2	26	3256	3182	50.74	16143.95	97.7	0.706390758
3	9	1302	1174	45.37	5324.17	90.2	0.66263551
4	24	562	556	40.65	2259.73	98.9	0.620766263
5	15	872	858	45.78	3927.47	98.4	0.663981684
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	18	906	928	43.38	4026.53	102	0.657200745

Table 8. Purity Values for the 2021 Dataset

Data No-	X1	X2	X3	X4	X5	X6	Purity Value
1	57	10040	9750	51.05	49771.71	97.1	0.712408749
2	26	4327	4427	51.47	22785.77	102	0.716041377
3	9	2186	2148	48.52	10421.61	98.3	0.694327374
4	24	1376	1351	48.05	6490.59	98.2	0.684228936
5	15	1480	1809	48.62	8797.30	122	0.709779854
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	18	1357	1476	48.04	7092.15	109	0.69471281

Table 9. Purity Values for the 2022 Dataset

Data No-	X1	X2	X3	X4	X5	X6	Purity Value
1	57	9142	9142	55.41	50655.27	100	0.73146631
2	26	3995	3995	55.55	22193.89	100	0.728493812
3	9	1548	1538	46.37	7131.24	99.3	0.681026439
4	24	775	771	48.41	3734.35	99.5	0.672640196
5	15	713	713	50.03	3565.64	100	0.678313172
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	18	1038	1060	50.52	5354.61	102	0.693121169

Table 10. Purity Values for the 2023 Dataset

Data No-	X1	X2	X3	X4	X5	X6	Purity Value
1	57	11341	9818	51.05	50120.89	86.6	0.70039313
2	26	4740	3086	51.47	15883.64	65.1	0.664106152
3	9	1568	1456	48.52	7064.51	92.9	0.683767409
4	24	1195	1388	48.05	6669.34	116	0.697871242
5	15	1632	1008	48.62	4900.90	61.8	0.634170372
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	18	1038	1060	50.52	5354.61	102	0.670254046

As indicated in Tables 6 to 10, the highest purity value in the 2019 dataset was 0.744839407 for the 13th data (Muara Batu), while the lowest was 0.61442911 for the 4th data (Cot Girek). In the 2020 dataset, the maximum value was 0.731896917 for the 19th data (Sawang), and the minimum was 0.620766263 for the 4th data (Cot Girek). For the 2021 dataset, the highest purity value was 0.732278252 for the 19th data (Sawang), and the lowest was 0.648936824 for the 6th data (Geureudong Pase). In the 2022 dataset, the maximum value reached 0.7343035 for the 25th data (Tanah Luas), whereas the minimum was 0.53797451 for the 6th data (Geureudong Pase). Lastly, in the latest dataset, the top purity value recorded was 0.74959294 for the 23rd data (Syamtalira Bayu), with the lowest being 0.51273326 for the 20th data (Seunuddon).

With the proposed model, the data with the maximum and minimum Purity values are utilized as the initial centroids for K-Means clustering. The clustering process is conducted 10 times. Subsequently, the clustering results are analyzed by comparing the average number of iterations and the DBI values between the original K-Means and Purity K-Means.

Clustering Process

The initial centroids utilized for the clustering process are presented in detail in Table 11.

Table 11. Initial Centroid for Each Dataset

Dataset Year-	Data No-	X1	X2	X3	X4	X5	X6
2019	13	24	2612	2641	60.12	15877.692	101
	4	24	475	493	40.07	1975.451	104
2020	19	39	6520	6436	56.2	36167.51	98.7
	4	24	562	556	40.65	2259.73	98.9
2021	19	39	6463	6344	56.48	35828.09	98.2
	6	11	537	467	49	2288.30	87
2022	25	57	4488	4488	57.21	25675.28	100
	6	11	197	123	47.63	586.33	62.5
2023	23	38	2816	4041	53.6	21659.76	144
	20	33	5310	1393	51.67	7197.63	26.2

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.metrics import accuracy_score
file_path = '/content/drive/MyDrive/CSV/2019.csv'
data = pd.read_csv(file_path)
selected_features =
data[['Jumlah_Desa','Luas_Tanam','Luas_Panen','Produktivitas','Produksi','Persentase']]
num_clusters = 2
def calculate_purity(data):
    purity_values = []
    for index, row in data.iterrows():
        purity = row['Jumlah_Desa'] / (row['Luas_Tanam'] + row['Luas_Panen'] + 1)
        purity_values.append(purity)
    return np.array(purity_values)
data['Purity'] = calculate_purity(selected_features)
max_purity_index = data['Purity'].idxmax()
min_purity_index = data['Purity'].idxmin()
initial_centroids = [selected_features.iloc[max_purity_index].values,
selected_features.iloc[min_purity_index].values]
kmeans = KMeans(n_clusters=num_clusters, init=np.array(initial_centroids), n_init=1)
kmeans.fit(selected_features)
data['cluster_label'] = kmeans.labels_
print(data)
print("Number of iterations:", kmeans.n_iter_)
data.info()
plt.figure(figsize=(8,4))
for cluster in range(num_clusters):
    cluster_data = data[data['cluster_label'] == cluster]
    plt.scatter(cluster_data['Jumlah_Desa'], cluster_data['Produktivitas'], label=f'Cluster
{cluster + 1}')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 3], marker='x',
color='black', label='Centroids')
plt.xlabel('Jumlah Desa')
plt.ylabel('Produktivitas')
plt.title('Purity K-Means Clustering on 2019 Dataset')
```

Figure 4. Procedure of Purity K-Means in Python

Figure 4 depicts the K-Means clustering procedure in Python applied to data that has been computed for its purity using the Purity method. This process is conducted to cluster the data present in the entire dataset that will be tested from 2019 to 2023. The clustering results performed with Purity K-Means on the 2019 dataset are shown in Table 12 and visualized in Figure 5. The comprehensive clustering results of the dataset tested with Purity K-Means are shown in the subsequent Figure 6.

Table 12. Purity K-Means Clustering Results for the 2019 Dataset

No	District	Cluster
1	Baktiya	Priority
2	Baktiya Barat	Priority
3	Banda Baro	Non Priority
4	Cot Girek	Non Priority
5	Dewantara	Non Priority
6	Geureudong Pase	Non Priority
7	Kuta Makmur	Non Priority
8	Langkahan	Non Priority
9	Lapang	Non Priority
10	Lhoksukon	Priority
11	Matang Kuli	Priority
12	Meurah Mulia	Priority
13	Muara Batu	Non Priority
14	Nibong	Non Priority
15	Nisam	Non Priority
16	Paya Bakong	Non Priority
17	Pirak Timu	Non Priority
18	Samudera	Non Priority
19	Sawang	Priority
20	Seunuddon	Priority
21	Simpang Kramat	Non Priority
22	Syamtalira Aron	Non Priority
23	Syamtalira Bayu	Non Priority
24	Tanah Jambo Aye	Priority
25	Tanah Luas	Priority
26	Tanah Pasir	Non Priority

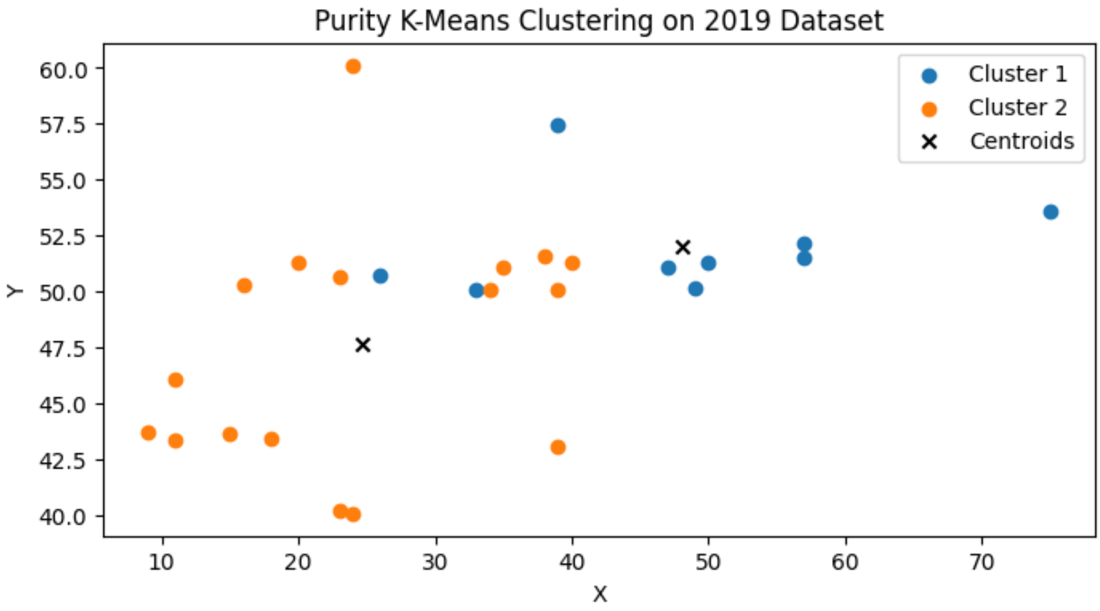

Figure 5. Visualization of Purity K-Means Clustering Results for the 2019 Dataset

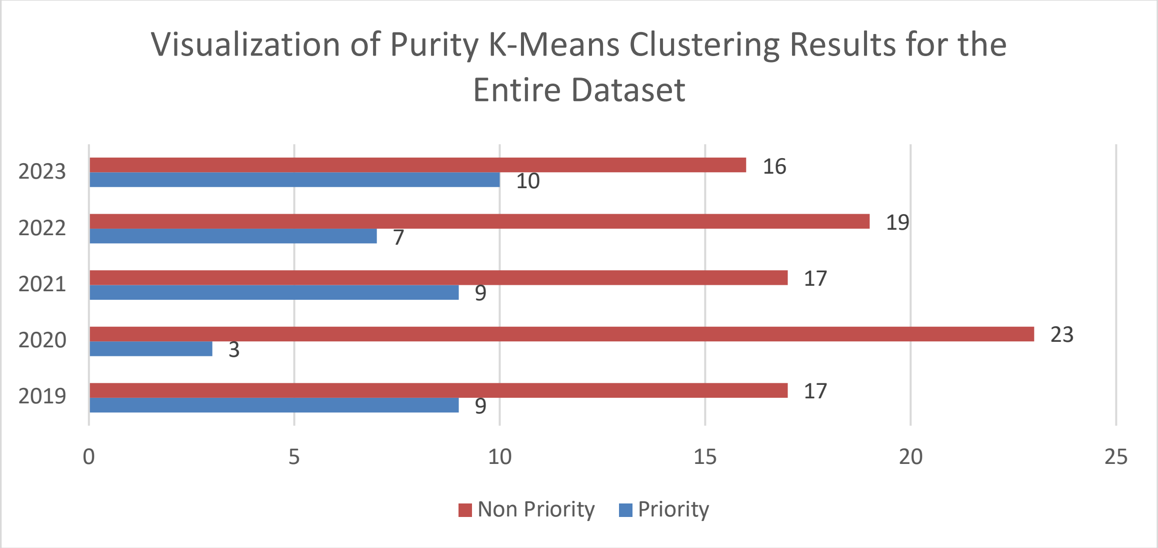

Figure 6. Visualization of Purity K-Means Clustering Results for the Entire Dataset

Clustering Evaluation

We conducted the clustering process with both models (the conventional model and the proposed model) 10 times on each dataset. As for the evaluation of the results, the number of iterations produced for each test is displayed in Table 13 and visualized in the following Figure 7.

Table 13. Comparison of Clustering Results on the 2019 Dataset

No	Num of Iters	
	Conventional K-Means	Purity K=Means
1	7	5
2	7	5
3	5	5
4	3	5
5	5	5
6	7	5
7	7	5
8	5	5
9	5	5
10	3	5
Average	**5.4**	**5**

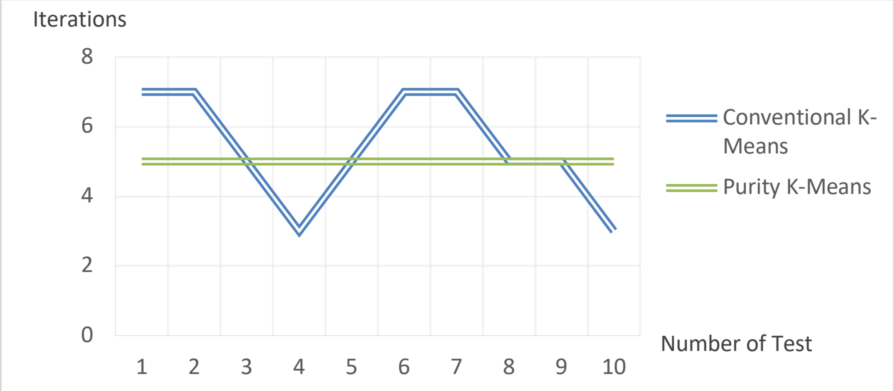

Figure 7. Visualization of number of Iteration for the Clustering of the 2019 Dataset

Based on Table 13 and Figure 7, conventional K-Means shows an unpredictable number of iterations in 10 tests conducted on the 2019 dataset. The highest was 7 iterations in the 1st, 2nd, 6th, and 7th tests, while the lowest was 3 iterations in the 4th and 10th tests. Purity K-Means, on the other hand, completed the clustering process in only 5 iterations. On average, conventional K-Means required 5.4 iterations, slightly more than Purity K-Means.

As for the overall evaluation of the number of iterations produced for each clustering process, it is displayed in Figure 8. As shown in Figure 8, it can be concluded that Purity has an impact in reducing the number of iterations in the clustering process conducted with the K-Means method, where the number of iterations obtained with Purity K-Means to achieve convergence is fewer compared to conventional K-Means. Also, the performance measurement of the clustering process can be evaluated and analyzed using DBI. Formulas (3), (4), (5), and (6) were used to calculate the Davies-Bouldin Index value. Therefore, the DBI value for the datasets clustered by both conventional K-Means and Purity K-Means is presented in Table 14 and visualized in Figure 9.

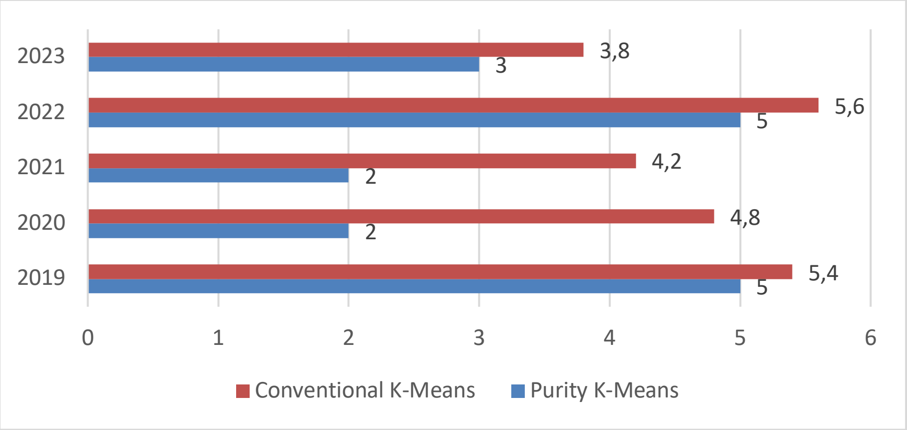

Figure 8. Comparison Graph of number of Iteration for the Clustering Proces

Table 14. Davies-Bouldin Index (DBI) Value for the Clustering Results

Dataset	Davies Bouldin Index Value	
	Conventional K-Means	Purity K=Means
2019	0.7178	0.6781
2020	0.6025	0.4175
2021	0.4971	0.4419
2022	0.7222	0.6182
2023	0.5519	0.4973

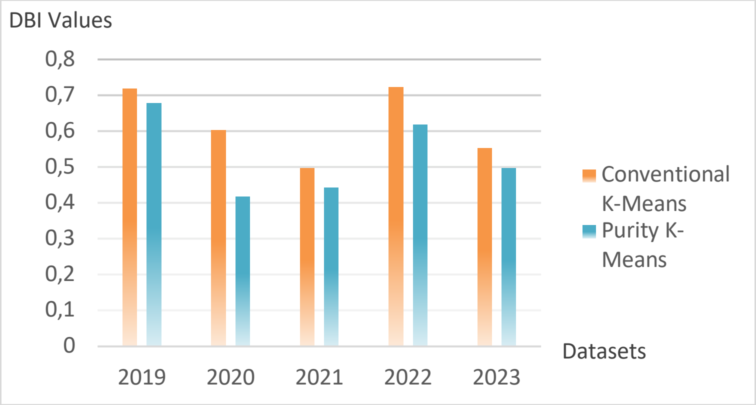

Figure 9. Visualization of the DBI values

Based on Figure 9, it is concluded that Purity K-Means produces a lower DBI value compared to conventional K-Means. This is consistent with the principle of DBI, where a value approaching zero indicates a more valid clustering process.

CONCLUSION

From the research conducted using the rice harvest productivity dataset in Aceh Utara Regency, the use of the Purity algorithm in the K-Means clustering process showed a higher level of effectiveness and efficiency compared to conventional K-Means clustering. The results of the Purity K-Means testing obtained an average of 5, 2, 2, 5, and 3 iterations from the entire dataset tested sequentially from 2019 to 2023. Meanwhile, conventional K-Means testing obtained an average number of iterations of 5.4, 4.8, 4.2, 5.6, and 3.8 iterations, respectively. The DBI values obtained from Purity K-Means for the entire dataset sequentially are 0.6781, 0.4175, 0.4419, 0.6182, and 0.4973. These values are smaller compared to the DBI values obtained from conventional K-Means. The results of this study also provide useful and more valid references for the Aceh Utara Regency Government in making further policies for the management of rice harvest intensification programs in Aceh Utara Regency, such as increasing the distribution of fertilizer subsidies to areas included in priority clusters. Furthermore, this research contributes significantly to the development of algorithms, especially in evaluating clustering models in similar cases.

REFERENCES

[1] E. Suryani, R. A. Hendrawan, I. Muhandhis, and R. Indraswari. A simulation model to improve the value of rice supply chain (A case study in East Java-Indonesia). *Journal of Simulation.* 2022; 16(4): 392-414..

[2] N. Liundi, A. W. Darma, R. Gunarso, and H. L. H. S. Warnars. Improving rice productivity in Indonesia with artificial intelligence. *7th International Conference on Cyber and IT Service Management (CITSM).* 2019; 1-5.

[3] Kantor Dinas Pertanian dan Pangan Kabupaten Aceh Utara. 2024.

[4] K. P. Sinaga, and M. S. Yang. Unsupervised K-Means clustering algorithm. *IEEE Access 8.* 2020; 80716-80727.

[5] T. M. Ghazal. Performances of k-means clustering algorithm with different distance metrics. *Intelligent Automation & Soft Computing.* 2021; 30(2): 735-742.

[6] F. Marissa, A. Zahma, A. M. Bau, E. Noviansa, A. S. Neno, A. Lidya, and Maukar. Digitasi Produktivitas Panen Padi Berbasis K-Means Clustering. *SMARTICS Journal*. 2021; 7(1): 21-26.

[7] C. J. Silalahi, A. Situmorang, and J. F. Naibaho. Implementasi Metode K-Means Clustering Untuk Memetakan Daerah Potensial Penghasil Padi di Provinsi Sumatera Utara. *Methotika: Jurnal Ilmiah Teknik Informatika.* 2022; 2(2): 49-57.

[8] E. A. P. Putri. K-Means Clustering untuk pengelompokan daerah penghasil padi di Indonesia berdasarkan luas panen, produksi, dan produktivitas padi tahun 2022. *Maliki Interdisciplinary Journal (MIJ)*. 2024; 2(1): 128-137.

[9] B. Kristanto, A. T. Zy, and M. Fatchan. Analisis Penentuan Karyawan Tetap Dengan Algoritma K-Means dan Davies Bouldin Index. *Bulletin of Information Technology (BIT).* 2023; 4(1): 112-120.

[10] M. Mughnyanti, S. Efendi, and M. Zarlis. Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. *IOP Conference Series: Materials Science and Engineering.* 2020; 725(1): 012128.

[11] F. Ros, R. Riad, and S. Guillaume. PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*. 2023; 528: 178-199.

[12] M. Ahmed, R. Seraj, and S. M. S. Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics.* 2020; 9(8): 1295.

[13] L. Nigro. Performance of parallel K-Means algorithms in Java. *Algorithms.* 2022; 15(4): 117.

[14] L. K. Choi, K. B. Rii, and H. W. Park. K-Means and J48 Algorithms to Categorize Student Research Abstracts. *International Journal of Cyber and IT Service.* 2023; 3(1): 61-64.

[15] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng. A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots. *Applied Sciences.* 2021; 11(23): 11202.

[16] S. Retno, N. Hasdyna, and B. Yafis. K-NN with Purity Algorithm to Enhance the Classification of the Air Quality Dataset. *Journal of Advanced Computer Knowledge and Algorithms.* 2024; 1(2): 42-46.

[17] R. K. Dinata, N. Hasdyna, S. Retno, and M. Nurfahmi. K-means algorithm for clustering system of plant seeds specialization areas in east Aceh. *ILKOM Jurnal Ilmiah*. 2021; 13(3): 235-243.

[18] S. Retno. Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai

Titik Pusat Cluster Awal (Centroid). *Thesis*. 2019.

[19] J. D. Sitompul, O. S. Sitompul, and P. Sihombing. Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. *Journal of Physics: Conference Series.* 2019; 1235(1): 012015.

[20] N. Suarna, Y. A. Wijaya, T. Hartati, and T. Suprapti. Comparison K-Medoids algorithm and K-Means algorithm for clustering fish cooking menu from fish dataset. *IOP Conference Series: Materials Science and Engineering.* 2021; 1088(1): 012034.

[21] F. B. Ashraf, A. Martin, M. S. R. Shafi, and M. U. Islam. An improved k-means clustering algorithm for multi-dimensional multi-cluster data using meta-heuristics. *24th International Conference on Computer and Information Technology (ICCIT).* 2021; 1-6.