JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 10, No. 1, JANUARY, 2025, Pp. 1 – 15
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Optimizing K-Means Algorithm Using the Purity Method for Clustering Oil Palm Producing Regions in North Aceh

Novia Hasdyna (1)*, Rozzi Kesuma Dinata (2), Balqis Yafis (3)
1 Informatika, Universitas Islam Kebangsaan Indonesia, Aceh, Indonesia
2 Teknik Informatika, Universitas Malikussaleh, Aceh, Indonesia
3 Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
e-mail : noviahasdyna@uniki.ac.id, rozzi@unimal.ac.id, balqisyafis@nycu.edu.tw.
* Corresponding author.

This article was submitted on 15 October 2024, revised on 2 November 2024, accepted on 3 November 2024, and published on 31 January 2025.

Abstract

The K-Means algorithm is a fundamental tool in machine learning, widely utilized for data clustering tasks. This research aims to improve the performance of the K-Means algorithm by integrating the Purity method, specifically focusing on clustering regions renowned for oil palm production in North Aceh. Oil palm cultivation is a vital agricultural sector in North Aceh, contributing significantly to the local economy and employment. This study examines two clustering techniques: the conventional K-Means algorithm and an optimized version, Purity K-Means. Integrating the Purity method increases K-Means' efficiency by decreasing the required convergence iteration. The data used for clustering analysis is sourced from the Department of Agriculture and Food in North Aceh Regency and pertains to oil palm production in 2023. The findings indicate that the Purity K-Means approach notably reduces the iteration count and improves cluster quality. The average Davies-Bouldin Index (DBI) for standard K-Means is 0.45, whereas the Purity K-Means method lowers it to 0.30. Furthermore, applying the Purity method reduced the number of K-Means iterations from 15 to just 3. These results highlight an enhancement in clustering performance and overall efficiency.

Keywords: K-Means Algorithm, Purity Method, Data Clustering, Oil Palm Production, North Aceh, Davies-Bouldin Index (DBI)

Abstrak

Algoritma K-Means merupakan alat dasar dalam pembelajaran mesin yang banyak digunakan untuk tugas pengelompokan data. Penelitian ini bertujuan untuk meningkatkan kinerja algoritma K-Means dengan mengintegrasikan metode Purity, yang secara khusus difokuskan pada pengelompokan wilayah-wilayah yang terkenal dengan produksi kelapa sawit di Aceh Utara. Budidaya kelapa sawit merupakan sektor pertanian yang vital di Aceh Utara, memberikan kontribusi signifikan terhadap perekonomian lokal dan penyerapan tenaga kerja. Studi ini membandingkan dua pendekatan pengelompokan, yaitu K-Means standar dan Purity K-Means yang telah dioptimalkan. Metode Purity digunakan untuk meningkatkan efisiensi algoritma K-Means dengan mengurangi jumlah iterasi yang diperlukan untuk konvergensi. Data yang digunakan dalam analisis pengelompokan bersumber dari Dinas Pertanian dan Pangan Kabupaten Aceh Utara dan berkaitan dengan produksi kelapa sawit pada tahun 2023. Hasil penelitian menunjukkan bahwa pendekatan Purity K-Means secara signifikan mengurangi jumlah iterasi dan meningkatkan kualitas cluster. Nilai rata-rata Davies-Bouldin Index (DBI) untuk K-Means standar adalah 0,45, sedangkan metode Purity K-Means menguranginya menjadi 0,30. Selain itu, jumlah iterasi K-Means berkurang dari 15 menjadi 3 saat menggunakan metode Purity. Temuan ini mengindikasikan peningkatan kinerja pengelompokan dan efisiensi secara keseluruhan.

Kata Kunci: Algoritma K-Means, Metode Purity, Pengelompokan Data, Produksi Kelapa Sawit, Aceh Utara, *Davies-Bouldin Index* (DBI)

1. INTRODUCTION

Clustering is an essential data analysis technique that groups similar objects based on specific attributes, providing valuable insights across various applications (Ezugwu et al., 2022). Among the many clustering algorithms available, the K-Means algorithm is particularly notable for its widespread use due to its simplicity, efficiency, and effectiveness in handling large datasets (Kouadio et al., 2024; Li et al., 2023). However, K-Means has its limitations; one significant challenge is the high number of iterations required for convergence, which can increase computational time and overall processing demands, especially with large datasets (Cebolla-Alemany et al., 2024). This study seeks to address these limitations by integrating the Purity method to enhance the efficiency of the K-Means algorithm.

The focus on North Aceh's oil palm production is grounded in the region's economic reliance on this sector, which plays a key role in local employment and economic growth. As demand for oil palm products increases, analyzing and understanding production data in regions like North Aceh has become essential. However, the large agricultural datasets generated in this sector pose significant challenges for traditional clustering methods, especially regarding scalability and meaningful data grouping. Therefore, this research aims to contribute to understanding and analyzing oil palm production in North Aceh. In this region, data-driven insights can substantially impact local and regional development.

In recent years, advances in clustering algorithms have shown significant potential for agricultural applications, where techniques such as hierarchical clustering, K-Medoids, and others have been employed to manage and analyze agricultural data effectively and efficiently. The Purity method, in particular, provides a valuable approach by enhancing cluster homogeneity, thus improving cluster interpretability and consistency. By integrating Purity with K-Means, this study aims to improve clustering quality and reduce the number of iterations, enabling more efficient processing of complex agricultural datasets.

A literature review shows various studies that have explored the application of clustering algorithms in agricultural contexts. For example, Majumdar et al. (2023) used K-Means to optimize irrigation management in rice production, leading to better yield predictions. Similarly, Rezaee et al. (2023) applied K-Means to classify soil types based on diverse attributes, highlighting their effectiveness in agricultural land management. Naz et al. (2024) utilized K-Means to analyze crop yield data, identifying patterns that improved resource allocation. Thakur & Kaur (2024) used K-Means to identify potential areas for organic farming, demonstrating its value in promoting sustainable practices. Finally, Bhatti et al. (2024) combined K-Means with other machine learning techniques to improve crop disease prediction, showing the algorithm's adaptability in various agricultural applications. Despite these advancements, limited focus has been on optimizing K-Means using methods like Purity. This study seeks to address this gap by examining the potential of the Purity method to enhance K-Means performance, particularly for oil palm production data.

The objectives of this study are as follows:

a) To optimize the K-Means algorithm by integrating the Purity method, specifically focusing on clustering regions known for oil palm production in North Aceh.
b) To evaluate the performance of the standard K-Means algorithm compared to the optimized Purity K-Means approach, using data from the Department of Agriculture and Food in North Aceh Regency from 2023.
c) The effects of the purity method on the number of iterations required for convergence and overall clustering quality will be analyzed using the Davies-Bouldin Index (DBI).

This study hypothesizes that integrating the Purity method with the K-Means algorithm will significantly reduce the number of iterations required for convergence and enhance clustering quality, as indicated by a lower Davies-Bouldin Index (DBI) compared to the standard K-Means approach.

2. METHODS

This research adopts a quantitative approach to analyze the clustering of oil palm production regions in North Aceh. The study compares two clustering methodologies: the conventional K-Means algorithm and the enhanced Purity K-Means algorithm. The research design is structured to facilitate the assessment of the performance of these methods using agricultural data.

2.1 Dataset Preparation

The dataset utilized for this research consists of data related to oil palm production in North Aceh for 2023, as presented in Table 1. This data was sourced from the Department of Agriculture and Food in North Aceh Regency. It encompasses several key features that are essential for analyzing the factors influencing oil palm production, including:

a) Production Volume (X1): This feature represents the total volume of oil palm produced in each region, quantified in metric tons.
b) Land Area (X2): This denotes the area allocated for oil palm cultivation, measured in hectares.
c) Yield per Hectare (X3): This variable reflects the average oil palm yield per hectare, offering valuable insights into agricultural productivity.

Table 1 Research Dataset

No.	District Name	Production Volume	Land Area	Yield per Hectare
1	Sawang	0,851	11,388	15,600
2	Nisam	0,727	10,189	15,700
3	Nisam Antara	0,465	6,726	16,900
4	Kuta Makmur	2,388	39,603	17,500
5	Syamtalira Bayu	0,454	7,012	15,900
6	Geureudong Pase	0,952	13,675	15,540
7	Samudera	0,018	0,252	14,000
8	Meurah Mulia	0,461	5,277	15,800
9	Tanah Luas	0,441	5,379	16,500
10	Matang Kuli	0,358	1,766	16,500
11	Pirak Timu	0,380	4,051	16,400
12	Lhoksukon	2,170	35,055	16,520
13	Baktiya	1,047	16,286	16,500
14	Tanah Jambo Aye	1,629	18,431	16,500
15	Cot Girek	2,597	40,340	17,188
16	Langkahan	2,188	34,122	16,500
17	Baktiya Barat	0,100	1,504	15,500
18	Paya Bakong	0,423	3,185	16,500
19	Nibong	0,043	0,375	15,000
20	Simpang Kramat	0,410	4,603	16,800

Table 1 outlines the dataset utilized in this study, encompassing data from 20 districts in North Aceh related to oil palm production for 2023. Each entry provides distinct characteristics for the districts, offering a snapshot of agricultural dynamics in the region. The dataset is vital for evaluating the varying oil palm output levels and understanding the land distribution dedicated to this crucial crop. By analyzing these parameters, researchers can derive insights into regional agricultural practices and identify potential areas for improvement and intervention. This diverse dataset facilitates a comprehensive examination of factors that may influence oil palm production across different districts in North Aceh.

2.2 Proposed Model

This research introduces two distinct models for clustering oil palm-producing regions in North Aceh: the Purity K-Means model and the conventional K-Means model. Both models aim to enhance the clustering process but differ in their methodologies and implementation.

2.2.1 Purity K-Means Model

The Purity K-Means model integrates the standard K-Means algorithm with the Purity method to optimize the clustering process, as shown in Figure 1.

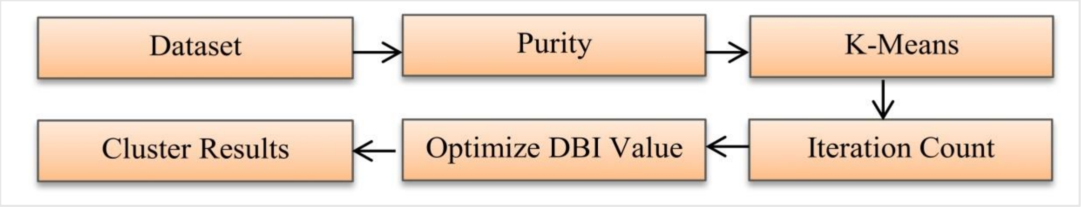

Figure 1 Purity K-Means Model

In Figure 1, the stages of the Purity K-Means process are as follows:

a) Data Preparation
The dataset is normalized to ensure equal contribution from each feature during the clustering process. This step is essential for minimizing bias associated with varying feature scales.

b) Purity Calculation
After each iteration, the model calculates the Purity score for the generated clusters. This score evaluates cluster homogeneity, providing insights into the effectiveness of centroid adjustments and data point assignments. Purity is utilized to assess a cluster's purity value, identifying the most suitable cluster member within a class (Dinata et al., 2023). The formula for calculating Purity is presented in Equation (1). Purity (y) reflects the purity level for the y-variable, where N_y represents the total data points within the y-cluster, and y signifies the cluster index (Hasdyna & Dinata, 2024).

$$Purity\ (y) = \frac{1}{N_y} \max\ (n_{xy}) \tag{1}$$

c) Initial Centroid Selection
Instead of random initialization, this model uses the Purity method to select initial centroids. This approach identifies centroids representing the data's inherent structure, improving the chances of forming meaningful clusters from the outset.

d) Clustering Process using k-means
The algorithm iteratively assigns data points to the nearest centroid, recalculates centroids based on current assignments, and repeats this process until convergence. The integration of the Purity method allows for continuous assessment of cluster quality during iterations. The K-means algorithm is applied for data clustering. The clustering process with K-means follows these steps: In Step 1, the desired number of clusters, denoted by 'k,' is determined. In Step 2, initial random values are assigned to the centroids of each of the 'k' clusters. The Euclidean distance formula is then used to calculate the distance between each data point and the centroids, shown in Equation (2) (Ariyanto et al., 2024).

$$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2} \tag{2}$$

Here, d represents a data point, xi denotes the data criteria, and μj indicates the cluster j's centroid. In Step 3, each data point is assigned to the cluster of the nearest centroid. Step 4 involves updating the centroids by calculating the mean of the data points within each cluster using the formula in Equation (3) (Retno et al., 2024).

$$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \in sj} xj \tag{3}$$

In this context, the symbol $\mu j(t+1)$ represents the centroid updated at iteration $t+1$, indicating the evolving center of a specific cluster. The term Nsj corresponds to the dataset contained within the Sj cluster, signifying the collection of data points grouped in that particular cluster. Additionally, xj represents the cumulative values within cluster Sj, effectively summarizing the overall attributes of the clustered data points. Finally, Step 5 marks the completion of the process. Steps 2 through 4 are repeated until no further changes occur in cluster membership, confirming convergence and consistent cluster assignments.

e) Performance Evaluation using DBI (Davies-Bouldin Index)

The model's performance is evaluated using metrics such as the Davies-Bouldin Index (DBI) and the number of iterations needed for convergence. These metrics facilitate clustering quality and efficiency assessment with the traditional K-Means method. The initial step in calculating the DBI involves determining the Sum of Squares Within the Cluster (SSW), which indicates the cohesion value. The DBI is then computed using the formula presented in Equation (4) (Ros et al., 2023).

$$SSW_i = \frac{1}{mi} \sum_{j=i}^{mi} d(xj, ci) \tag{4}$$

Once SSW has been computed, the subsequent step involves calculating the Sum of Squares Between Clusters (SSB), which reflects the cluster separation value. This is achieved using the formula presented in Equation (5) (Henderi et al., 2024).

$$SSB_{i,j} = d(c_i, c_j) \tag{5}$$

The following step involves calculating the Ratio to compare the i-cluster with the j-cluster, utilizing the formula in Equation (6).

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{ij}} \tag{6}$$

After deriving the ratio value, the final step is to compute the DBI value using the formula provided in Equation (7).

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \tag{7}$$

In the context of DBI, a smaller value signifies better clustering results, indicating that the clusters are more internally cohesive and distinct, which is ideal for clustering tasks.

2.2.2 Conventional K-Means Model

The conventional K-Means model serves as a benchmark for evaluating the effectiveness of the Purity K-Means model. The steps involved in this model are outlined as follows:

a) Data Preparation
Like the Purity K-Means model, the dataset undergoes normalization to ensure that all features contribute equally to the clustering process.

b) Initial Centroid Selection
In this model, initial centroids are selected randomly from the dataset. This randomness can lead to varying clustering outcomes across different runs.

c) Clustering Process
The K-Means algorithm iteratively assigns data points to the nearest centroid and recalculates centroids based on the assigned points. This process continues until convergence, which may take multiple iterations.

d) Performance Evaluation
The performance of the conventional K-Means model is measured using the Davies-Bouldin Index (DBI) and the total number of iterations required for convergence. These metrics serve as indicators of clustering quality and efficiency.

3. RESULTS AND DISCUSSION

3.1 Purity calculation results

Table 2 Purity Calculation Results

No.	District Name	Production Volume	Land Area	Yield per Hectare	∑	Purity
1	Sawang	0,851	11,388	15,600	27,839	0,560364956
2	Nisam	0,727	10,189	15,700	26,616	0,589870754
3	Nisam Antara	0,465	6,726	16,900	24,091	0,701506787
4	Kuta Makmur	2,388	39,603	17,500	59,491	0,665697332
5	Syamtalira Bayu	0,454	7,012	15,900	23,366	0,680475905
6	Geureudong Pase	0,952	13,675	15,540	30,167	0,515132429
7	Samudera	0,018	0,252	14,000	14,27	0,981079187
8	Meurah Mulia	0,461	5,277	15,800	21,538	0,733587148
9	Tanah Luas	0,441	5,379	16,500	22,32	0,739247312
10	Matang Kuli	0,358	1,766	16,500	18,624	0,885953608
11	Pirak Timu	0,380	4,051	16,400	20,831	0,787288176
12	Lhoksukon	2,170	35,055	16,520	53,745	0,652246721
13	Baktiya	1,047	16,286	16,500	33,833	0,487689534
14	Tanah Jambo Aye	1,629	18,431	16,500	36,56	0,504130197
15	Cot Girek	2,597	40,340	17,188	60,125	0,670935551
16	Langkahan	2,188	34,122	16,500	52,81	0,646127627
17	Baktiya Barat	0,100	1,504	15,500	17,104	0,906220767
18	Paya Bakong	0,423	3,185	16,500	20,108	0,820568928
19	Nibong	0,043	0,375	15,000	15,418	0,972888831
20	Simpang Kramat	0,410	4,603	16,800	21,813	0,770182918

Table 2 and Figure 2 present the purity calculation results. To initiate K-Means clustering using purity values as centroids, we selected three representative subdistricts based on their purity scores: Samudera (high Purity, 0.9811), Langkahan (medium Purity, 0.6461), and Baktiya (low Purity, 0.4877). Samudera exhibits highly consistent attributes with the highest Purity (X1, X2, X3), making it an ideal centroid for clustering subdistricts with similar stability. Langkahan, with a moderate purity value and a balanced attribute sum of 52.81, serves as a centroid that captures subdistricts with average consistency. In contrast, Baktiya, having one of the lowest purity scores, indicates a high degree of attribute variability, making it suitable for clustering subdistricts with less consistency or greater diversity in attributes. These centroids will allow us to analyze groupings based on stability and variance within subdistrict attributes.

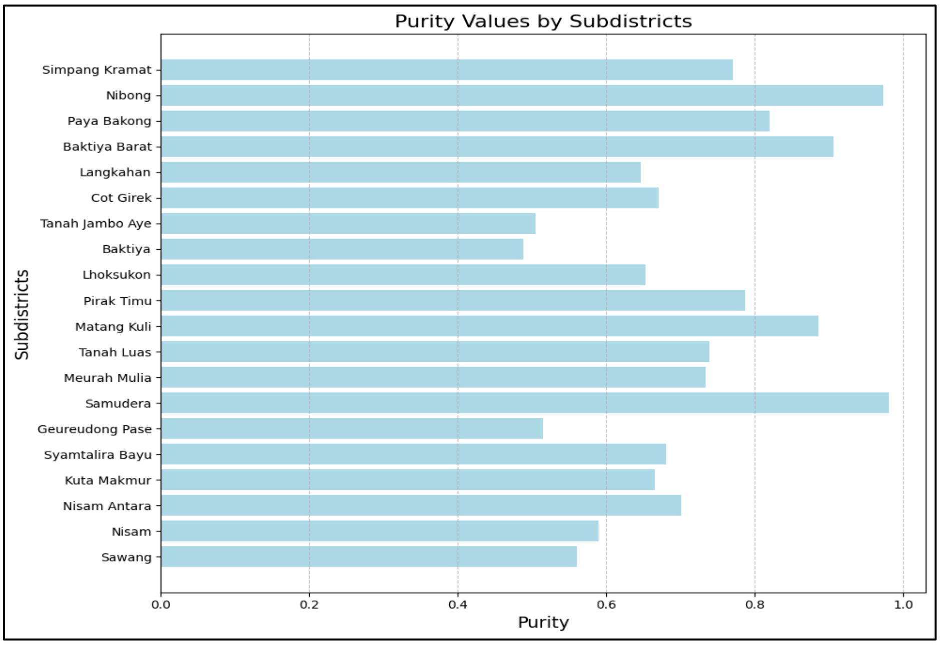

Figure 2 Purity Values in the Oil Palm Dataset

Table 3 outlines the initial centroids selected for K-Means clustering based on purity levels of three subdistricts. Samudera, with the highest purity value of 0.9811, signifies a cluster characterized by highly uniform attribute distributions (X1, X2, X3). This suggests that subdistricts grouped around Samudera are likely to share similar socioeconomic or environmental conditions, making it an effective point for clustering those with stable characteristics. In contrast, Langkahan, exhibiting a medium purity score of 0.6461, reflects a blend of consistency and variability in its attributes. This makes it an appropriate centroid for clustering subdistricts with average attributes, thereby capturing a wider range of subdistrict profiles without leaning too heavily towards extremely high or low consistency.

Table 3 Initial Centroids for K-Means Clustering Based on Purity Levels

Purity Level	Subdistrict	Purity Value	Characteristics
High Purity	Samudera	0.9811	Highly consistent attributes (X1, X2, X3), suitable as a centroid for stable subdistricts
Medium Purity	Langkahan	0.6461	Moderate consistency, with a balanced attribute sum of 52.81, ideal for capturing average subdistricts
Low Purity	Baktiya	0.4877	High variability in attributes, suitable for clustering subdistricts with less consistency or greater diversity

On the other hand, Baktiya, with the lowest purity score of 0.4877, indicates considerable diversity within its attributes. By choosing Baktiya as a low-purity centroid, the clustering process can effectively encompass subdistricts with more pronounced variations in their characteristics. This allows for identifying groups that may experience diverse conditions, which could be critical for targeted interventions or resource allocation. Overall, selecting these three subdistricts as centroids based on their purity scores allows for a nuanced approach in clustering, capturing varying levels of consistency and diversity across the dataset. This stratified methodology is beneficial for understanding the different dynamics present within the region.

3.2 Clustering process using Purity K-Means

To manually do K-Means clustering using the initial centroids provided in Table 4., the following steps will be undertaken:

a) Selection of Initial Centroids

The subdistricts Samudera, Langkahan, and Baktiya will be designated as the initial centroids, as shown in Table 4.

Table 4 Initial Centroids for K-Means Clustering

Centroid	Subdistrict	X1	X2	X3
C1	Samudera	0.018	0.252	14.000
C2	Langkahan	2.188	34.122	16.500
C3	Baktiya	1.047	16.286	16.500

b) Cluster Assignment

For each subdistrict, we will compute the Euclidean distance to each centroid and assign the subdistrict to the nearest centroid based on the calculated distances, as shown in Table 5.

Table 5 Euclidean Distances to Initial Centroids

No.	Subdistrict	Distance to C1 (Samudera)	Distance to C2 (Langkahan)	Distance to C3 (Baktiya)
1	Sawang	28.837	29.356	24.098
2	Nisam	27.769	27.054	25.274
3	Nisam Antara	17.172	19.058	16.185
4	Kuta Makmur	41.266	43.942	40.351
5	Syamtalira Bayu	23.097	23.719	22.005
6	Geureudong Pase	29.205	31.187	27.602
7	Samudera	0.000	30.250	14.000
8	Meurah Mulia	21.721	23.169	18.071
9	Tanah Luas	24.640	25.712	22.825
10	Matang Kuli	18.448	20.226	16.070
11	Pirak Timu	20.096	21.798	18.000
12	Lhoksukon	86.424	86.956	53.245
13	Baktiya	31.135	32.530	0.000
14	Tanah Jambo Aye	36.435	36.564	36.198
15	Cot Girek	60.125	61.892	29.382
16	Langkahan	52.610	52.013	52.241
17	Baktiya Barat	17.104	19.073	16.817
18	Paya Bakong	20.108	22.760	19.907
19	Nibong	15.417	16.300	14.628
20	Simpang Kramat	21.813	23.013	20.218

c) Centroid Update

After assigning the clusters, we will determine the new centroids by calculating the mean attribute values of the subdistricts within each cluster, as shown in Table 6.

Table 6 Updated Centroids After First Iteration

Centroid	Coordinates (X1, X2, X3)
C1	(0.0305, 0.3135, 14.500)
C2	(2.33575, 37.029, 17.077)
C3	(0.553, 7.928, 16.444)

d) Repetition

The assignment and update steps will be repeated until the centroids converge, when they no longer change significantly, or when the cluster assignments remain constant. The purity K-Means clustering process reached convergence after three iterations, where the centroids stabilized, indicating that further adjustments in cluster assignments were no longer necessary. In the first iteration, the initial centroids, Samudera, Langkahan, and Baktiya, were assigned to clusters based on the proximity of subdistricts, resulting in new centroid calculations. Subsequent iterations demonstrated a gradual refinement of cluster assignments and centroid positions, reflecting the algorithm's effectiveness in identifying distinct groupings among the subdistricts based on their attributes. Ultimately, the stability achieved in the third iteration suggests that the clustering solution accurately captures the underlying patterns in the data, enabling meaningful insights into the characteristics of each subdistrict based on their calculated purity values.

Table 7 Purity K-Means Clustering Results for Subdistricts

Subdistricts	Cluster
Samudera	C1
Nibong	C1
Langkahan	C2
Cot Girek	C2
Baktiya	C3
Nisam Antara	C3
Sawang	C1
Nisam	C2
Kuta Makmur	C2
Syamtalira Bayu	C2
Geureudong Pase	C2
Meurah Mulia	C2
Tanah Luas	C2
Matang Kuli	C3
Pirak Timu	C2
Lhoksukon	C2
Tanah Jambo Aye	C3
Langkahan	C2
Paya Bakong	C2
Simpang Kramat	C2

In the context of clustering oil palm-producing regions in North Aceh, the Purity K-Means analysis identified three distinct clusters that highlight varying regional characteristics.

Cluster C1, which includes subdistricts like Samudera and Nibong, demonstrates high purity values, suggesting these areas have similar environmental and economic conditions conducive to oil palm production. This consistency may indicate effective agricultural practices or favorable land conditions, warranting the implementation of targeted agricultural policies to enhance production efficiency. Cluster C2 consists of larger subdistricts, such as Langkahan and Cot Girek, showcasing moderate attribute consistency.

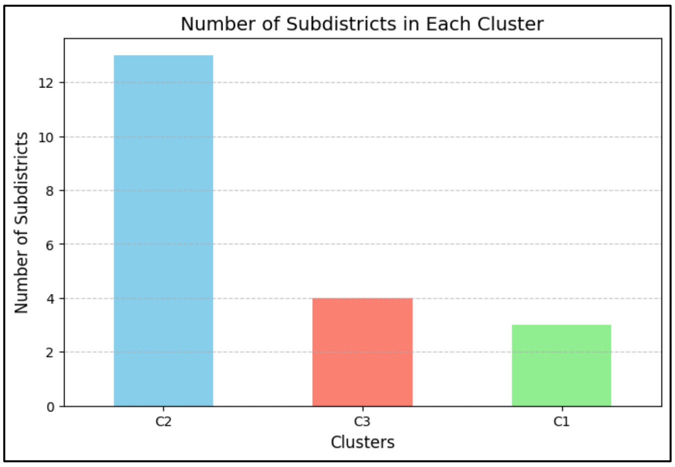

Figure 3 Distribution of Clustering Results with Purity K-Means

This cluster likely represents regions with diverse agricultural practices and varying production levels, indicating a need for tailored support and resource allocation to optimize their oil palm outputs. Conversely, Cluster C3, which encompasses Baktiya and Nisam Antara, exhibits lower purity values, highlighting greater variability in production conditions and practices. The unique challenges faced by these regions may require specialized interventions or research to improve their oil palm production capabilities. Overall, these clustering results provide critical insights that can inform strategic decision-making and development efforts in the oil palm sector of North Aceh. The visualization of the clustering results using Purity K-Means is presented in Figure 4.

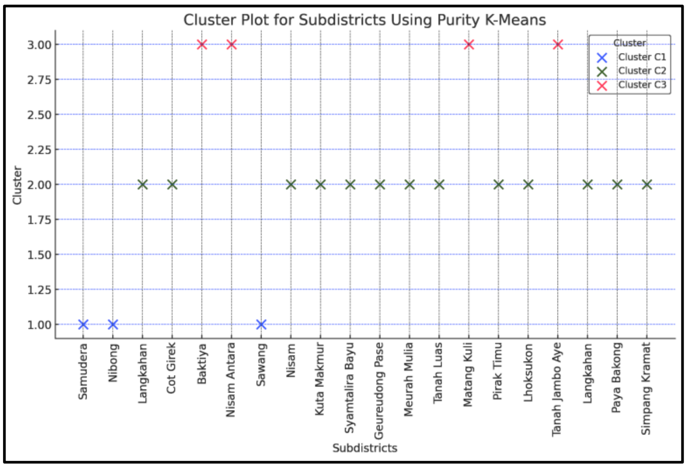

Figure 4 Visualization of the clustering results using Conventional K-Means

3.3 Results of the Conventional K-Means Model

In conventional K-Means, the initial centroids are selected randomly, unlike in Purity K-Means, where the initial centroids are chosen based on purity results.

Table 8 Initial Centroids for K-Means Clustering

Centroid	Subdistrict	X1	X2	X3
C1	Nisam Antara	0,103	1,656	16,900
C2	Samudera	0,018	0,252	14,000
C3	Pirak Timu	0,285	4,051	16,400

Table 8 provides an overview of the initial centroids assigned randomly for three regions within the conventional K-Means clustering model: Nisam Antara, Samudera, and Pirak Timu. These centroids represent starting points for each cluster, specifically in terms of three selected variables (Value 1, Value 2, and Value 3) relevant to the clustering analysis. The distance calculation results are presented in Table 6.

Table 9 Euclidean Distances to Initial Centroids in Conventional K-means

No.	Subdistrict	Distance to C1 (Samudera)	Distance to C2 (Langkahan)	Distance to C3 (Baktiya)
1	Sawang	9,847	11,281	7,402
2	Nisam	8,640	10,106	6,194
3	Nisam Antara	5,083	7,108	2,727
4	Kuta Makmur	38,020	39,577	35,631
5	Syamtalira Bayu	5,460	7,035	3,008
6	Geureudong Pase	12,125	13,543	9,685
7	Samudera	3,223	0,000	4,502
8	Meurah Mulia	3,801	5,356	1,376
9	Tanah Luas	3,760	5,720	1,341
10	Matang Kuli	0,487	2,942	2,288
11	Pirak Timu	2,462	4,508	0,095
12	Lhoksukon	33,465	34,960	31,061
13	Baktiya	14,666	16,260	12,259
14	Tanah Jambo Aye	16,849	18,421	14,443
15	Cot Girek	38,765	40,297	36,371
16	Langkahan	32,535	34,031	30,131
17	Baktiya Barat	1,408	1,956	2,708
18	Paya Bakong	1,613	3,875	0,883
19	Nibong	2,292	1,008	3,941
20	Simpang Kramat	2,965	5,189	0,693

Table 9 presents the Euclidean distances calculated between each subdistrict and the initial centroids for three clusters in the conventional K-Means model, with Samudera, Langkahan, and Baktiya serving as initial centroids (C1, C2, and C3, respectively). The table reveals how closely each subdistrict aligns with these centroids, where smaller distances indicate a higher likelihood of a subdistrict belonging to that particular cluster. For instance, the subdistrict Samudera has a distance of 0.000 to C1, affirming it as the initial centroid for that cluster. Similarly, Baktiya shows relatively small distances to its designated centroid, C3, while Langkahan displays minimal distance to C2, anchoring each as central points in their respective clusters. Some subdistricts, such as Pirak Timu and Baktiya Barat, show low distances to multiple centroids, suggesting they may lie near the boundaries of these clusters and may shift in subsequent iterations. This table provides insight into the initial grouping structure. K-Means will iteratively adjust centroids based on these calculated distances to minimize within-cluster variance, ultimately forming clusters with greater homogeneity. The initial distances guide the model's iterative process, influencing cluster composition and convergence in the final clustering result. The clustering results are presented in Table 10.

Table 10 shows the clustering results for each subdistrict in North Aceh, based on the conventional K-Means model, with each subdistrict assigned to one of three clusters (Cluster 1, Cluster 2, and Cluster 3). These clustering results can support targeted strategies for developing the oil palm sector in North Aceh, as shown in Figure 5.

Table 10 Conventional K-Means Clustering Results for Subdistricts

Subdistricts	Cluster
Samudera	1
Nibong	1
Langkahan	2
Cot Girek	3
Baktiya	2
Nisam Antara	1
Sawang	2
Nisam	2
Kuta Makmur	2
Syamtalira Bayu	2
Geureudong Pase	2
Meurah Mulia	3
Tanah Luas	1
Matang Kuli	1
Pirak Timu	3
Lhoksukon	3
Tanah Jambo Aye	2
Langkahan	2
Paya Bakong	2
Simpang Kramat	2

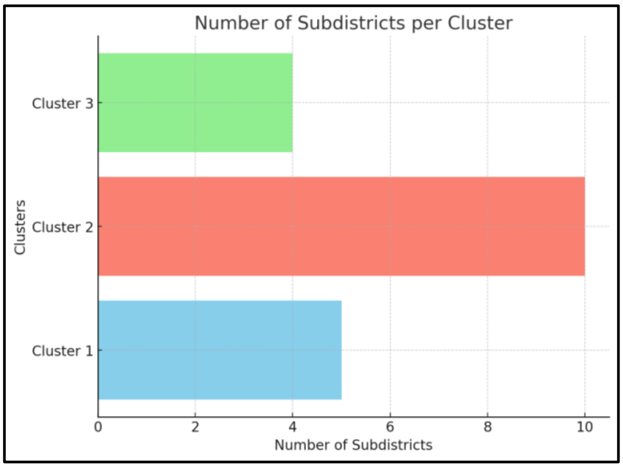

Figure 5 Distribution of Clustering Results with Conventional K-Means

These clusters provide insight into patterns within the region's oil palm sector. Cluster 1 includes subdistricts such as Sawang, Nisam, Geureudong Pase, Baktiya, and Tanah Jambo Aye, suggesting that these areas may share specific characteristics in oil palm productivity or resources that distinguish them from other clusters. Cluster 2, as the largest group, includes subdistricts like Nisam Antara, Samudera, and Meurah Mulia, indicating that this cluster represents the dominant pattern across the data, potentially covering regions with average productivity or typical oil palm-related features. Cluster 3, consisting of Kuta Makmur, Lhoksukon, Cot Girek, and Langkahan, likely represents subdistricts with unique attributes that set them apart from Clusters 1 and 2, possibly due to distinct environmental or infrastructural factors affecting oil palm production. The size of Cluster 2 suggests that it may capture the most prevalent characteristics across North Aceh's oil palm sector.

In contrast, Clusters 1 and 3 may represent more specialized or unique patterns within the industry. For instance, interventions or policies could be tailored to address each cluster's specific needs or strengths, likely more homogeneous within groups than across them. By leveraging these insights, decision-makers can apply targeted approaches to improve productivity, resource allocation, and sustainable practices within the sector, ensuring each cluster receives appropriate support based on its shared characteristics. The visualization of the clustering results using Conventional K-Means is presented in Figure 6.

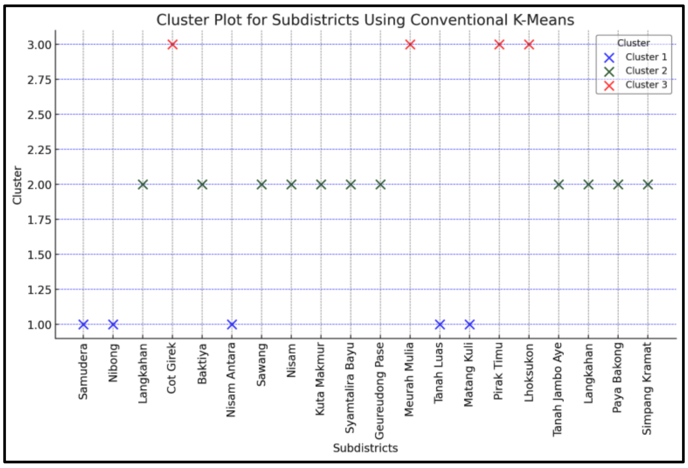

Figure 6 Visualization of the clustering results using Conventional K-Means

3.4 DBI Values and Iterations in Purity K-Means and Conventional K-Means

In evaluating the performance of clustering algorithms, the Davies-Bouldin Index (DBI) is a crucial metric for assessing the quality of the clusters formed. A lower DBI value indicates better cluster separation and cohesion. This section compares the iterations and DBI values of the Purity K-Means and Conventional K-Means algorithms. The analysis highlights the effectiveness of the Purity K-Means approach, demonstrating its superior performance in achieving lower DBI values with fewer iterations, thereby suggesting more efficient clustering, as shown in Table 11.

Table 11 Comparison of Iterations and DBI Values for Purity K-Means and Conventional K-Means

Method	Iterations	DBI Value
Purity K-Means	3	0.30
Conventional K-Means	15	0.45

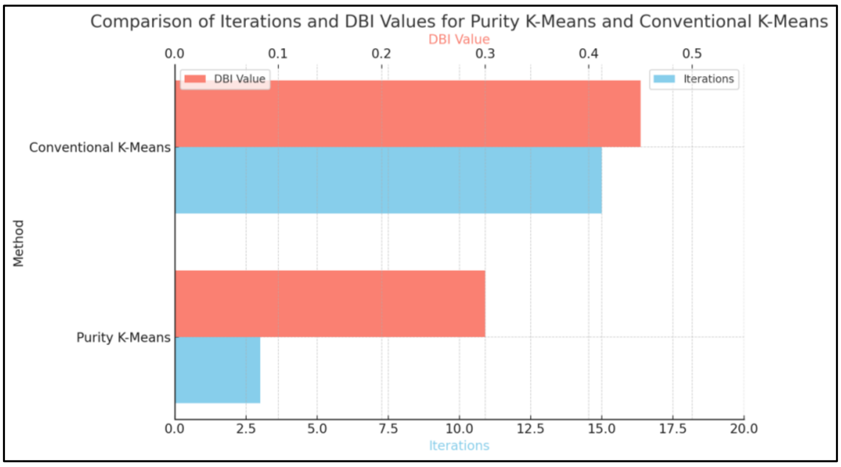

Figure 7 Comparison of Iterations and DBI Values for Purity K-Means and Conventional K-Means

The data presented in Table 8 highlights a significant improvement in the performance of the Purity K-Means algorithm compared to Conventional K-Means. Specifically, the Purity K-Means achieved its clustering results with only three iterations, while Conventional K-Means required 15 iterations. This reduction in iterations demonstrates the efficiency of the Purity K-Means approach and suggests that it can optimize the K-Means algorithm's performance. Additionally, the Dunn Index (DBI) values further substantiate these findings, with the Purity K-Means recording a lower DBI value of 0.30 compared to 0.45 for the Conventional K-Means. A lower DBI value indicates better clustering quality, confirming that the Purity K-Means method effectively minimizes the clustering overlap while enhancing the distinctiveness of clusters. Overall, these results illustrate that the Purity K-Means algorithm not only streamlines the clustering process but also enhances the overall quality of the results.

The horizontal bar chart illustrates the superior performance of the Purity K-Means algorithm compared to Conventional K-Means in terms of iterations and Davies-Bouldin Index (DBI) values. The Purity K-Means method achieves a remarkable reduction in iterations, requiring only three compared to 15 for the Conventional K-Means, indicating a more efficient clustering process and faster convergence to optimal solutions. Furthermore, the Purity K-Means demonstrates a lower DBI value of 0.30, in contrast to 0.45 for the Conventional K-Means. This lower DBI signifies better clustering performance, highlighting greater cluster separation and reduced intra-cluster variance. These findings emphasize that the Purity K-Means algorithm optimizes computational efficiency and enhances clustering quality, making it a valuable approach for effective data clustering.

4. CONCLUSIONS

This study successfully enhanced the performance of the K-Means algorithm by integrating the Purity method, focusing on oil palm production regions in North Aceh. The results demonstrated a significant improvement in clustering efficiency, as the Purity K-Means approach reduced the number of iterations required for convergence from 15 in conventional K-Means to just 3. Additionally, the Davies-Bouldin Index (DBI) value indicated a notable enhancement in cluster quality, decreasing from 0.45 in conventional K-Means to 0.30 in the Purity K-Means method.

The clustering analysis identified three distinct clusters within the subdistricts of North Aceh. Cluster 1 included subdistricts such as Sawang, Nisam, Geureudong Pase, Baktiya, and Tanah Jambo Aye, indicating shared characteristics in oil palm productivity. Cluster 2, the largest, comprised subdistricts such as Nisam Antara, Samudera, and Meurah Mulia, representing the region’s predominant production patterns. Finally, Cluster 3 included Kuta Makmur, Lhoksukon, Cot Girek, and Langkahan, likely reflecting distinctive attributes shaped by specific environmental or infrastructural factors.

These findings provide valuable insights for developing targeted strategies to enhance the oil palm sector in North Aceh and offer potential applications in other regions with similar conditions. By understanding each cluster's unique characteristics and needs, policymakers and stakeholders can implement tailored interventions to optimize productivity, resource allocation, and sustainability. Integrating the Purity method with the K-Means algorithm demonstrates significant potential for improving clustering outcomes in agricultural data analysis and related fields.

REFERENCES

Ariyanto, Y., Sabilla, W. I., & As Sidiq, Z. S. (2024). Recommendation System for Clustering to Allocate Classes for New Students Using The K-Means Method. *Compiler*, *13*(1), 27. https://doi.org/10.28989/compiler.v13i1.1962

Bhatti, M. A., Zeeshan, Z., M.S., S., Bhatti, U. A., Khan, A., Ghadi, Y. Y., Alsenan, S., Li, Y., Asif, M., & Afzal, T. (2024). Advanced Plant Disease Segmentation in Precision Agriculture Using Optimal Dimensionality Reduction With Fuzzy C-Means Clustering and Deep Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 18264–18277. https://doi.org/10.1109/JSTARS.2024.3437469

Cebolla-Alemany, J., Macarulla Martí, M., Viana, M., Moreno-Martín, V., San Félix, V., & Bou, D. (2024). Optimizing indoor air models through k-means clustering of nanoparticle size distribution data. *Building and Environment*, *266*, 112091. https://doi.org/10.1016/j.buildenv.2024.112091

Dinata, R. K., Adek, R. T., Hasdyna, N., & Retno, S. (2023). K-Nearest Neighbor Classifier Optimization Using Purity. *AIP Conference Proceedings*, *2431*(1). https://doi.org/10.1063/5.0117058/2906121

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743

Hasdyna, N., & Dinata, R. K. (2024). Comparative Analysis of K-Medoids and Purity K-Medoids Methods for Identifying Accident-Prone Areas in North Aceh Regency. *Scientific Journal of Informatics*, *11*(2), 263–272. https://doi.org/10.15294/SJI.V11I2.3433

Henderi, H., Fitriana, L., Iskandar, I., Astuti, R., Arifandy, M. I., Hayadi, B. H., Mesran, M., Chin, J., & Kurniawan, A. (2024). Optimization of Davies-Bouldin Index with k-medoids algorithm. *Science and Technology Research Symposium 2022*, *3065*(1), 030002. https://doi.org/10.1063/5.0225220/3311944

Kouadio, K. L., Liu, J., Liu, R., Wang, Y., & Liu, W. (2024). K-Means Featurizer: A booster for intricate datasets. *Earth Science Informatics*, *17*(2), 1203–1228. https://doi.org/10.1007/S12145-024-01236-3/METRICS

Li, M., Frank, E., & Pfahringer, B. (2023). Large scale K-means clustering using GPUs. *Data Mining and Knowledge Discovery*, *37*(1), 67–109. https://doi.org/10.1007/S10618-022-00869-6/TABLES/22

Majumdar, P., Bhattacharya, D., Mitra, S., Solgi, R., Oliva, D., & Bhusan, B. (2023). Demand prediction of rice growth stage-wise irrigation water requirement and fertilizer using Bayesian genetic algorithm and random forest for yield enhancement. *Paddy and Water Environment*, *21*(2), 275–293. https://doi.org/10.1007/S10333-023-00930-0/METRICS

Naz, H., Saba, T., Alamri, F. S., Almasoud, A. S., & Rehman, A. (2024). An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection. *IEEE Access*, *12*, 78611–78623. https://doi.org/10.1109/ACCESS.2024.3392032

Retno, S., Hasdyna, N., & Yafis, B. (2024). K-NN with Purity Algorithm to Enhance the Classification of the Air Quality Dataset. *Journal of Advanced Computer Knowledge and Algorithms*, *1*(2), 42–46. https://doi.org/10.29103/jacka.v1i2.15890

Rezaee, L., Davatgar, N., Moosavi, A. A., & Sepaskhah, A. R. (2023). Implications of Spatial Variability of Soil Physical Attributes in Delineating Site-Specific Irrigation Management Zones for Rice Crop. *Journal of Soil Science and Plant Nutrition*, *23*(4), 6596–6611. https://doi.org/10.1007/S42729-023-01513-Y/METRICS

Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, *528*, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043

Thakur, B., & Kaur, S. (2024). The Role of Artificial Intelligence in Biofertilizer Development. In *Metabolomics, Proteomics and Gene Editing Approaches in Biofertilizer Industry* (pp. 157–176). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-2910-4_9