ITEJ June-2026, Volume 11 No 1 Page 112 - 121

ITEJ
Information Technology Engineering Journals
eISSN : 2548-2157

Url : https://itej.uinssc.ac.id
Email : itej@syekhnurjati.ac.id

Clustering Analysis of Cocoa-Producing Areas Using the Gaussian Mixture Model Algorithm *(A Case Study in Southeast Aceh Regency)*

Pathia
Informatics Engineering
Malikussalaeh University
Lhokseumawe, Indonesia
pathia.210170128@mhs.unimal.ac.id

Taufiq
Informatics Engineering
Malikussalaeh University
Lhokseumawe, Indonesia
taufiq.te@unimal.ac.id

Sujacka Retno
Informatics Engineering
Malikussalaeh University
Lhokseumawe, Indonesia
sujacka@unimal.ac.id

Abstract—This study aims to identify and cluster agricultural areas in Southeast Aceh Regency using the Gaussian Mixture Model (GMM) algorithm. The dataset consists of village-level agricultural data, including land area, production volume, productivity, and the number of farmers. To ensure comparability across variables, Z-Score normalization was applied. The optimal number of clusters was determined using the Bayesian Information Criterion (BIC), resulting in three distinct groups: high, medium, and low production areas. Clustering performance was evaluated using the Silhouette Score (0.3893) and the Davies-Bouldin Index (0.8548), indicating moderate clustering quality with reasonable separation between clusters. To improve accessibility and practical use, a web-based information system was developed to visualize agricultural data, clustering outcomes, and evaluation metrics interactively. These findings highlight the value of GMM-based machine learning in supporting data-driven decision-making and prioritizing agricultural development efforts by local governments.

Keywords—Gaussian Mixture Model, clustering, agriculture, Southeast Aceh, information system.
Article info: Date Submitted: 23/02/2026 | Date Revised: 03/03/2026 | Date Accepted: 13/04/2026

I. INTRODUCTION

Cocoa is a nationally leading commodity that originates from South America and has spread widely across various tropical regions. [1].

Indonesia is one of the largest cocoa producers globally, ranking third after Côte d'Ivoire and Ghana, with an annual production of 593,331 tons—approximately 94.78% of which comes from smallholder plantations. [2]. Aceh Province has strong potential for cocoa development due to its favorable land and climate conditions.

However, there are various challenges in the upstream sector, such as low crop productivity and suboptimal bean quality due to inadequate fermentation processes. [3]. In Southeast Aceh Regency, there has been no comprehensive production mapping, resulting in agricultural policies

that are often misdirected. Therefore, a data mining approach is needed to cluster regions based on production levels—high, medium, and low—as a foundation for more accurate and targeted interventions.

The method used in this study is the Gaussian Mixture Model (GMM), a non-hierarchical clustering technique that models data as a combination of several Gaussian distributions [4]. GMM has also been successfully applied in Indonesia, for example, in clustering provinces based on poverty indicators [5], as well as in mapping provinces based on National Health Insurance (JKN) indicators from 2019 to 2021 [6]. The use of BIC and EM in both studies was able to produce representative clusters for national policy analysis.

The application of GMM, particularly in agricultural data, has also been growing: [7] using GMM in combination with K-Means to cluster cocoa production data in four provinces of Sumatra, with validation using the Silhouette Score. [7]. [8] applied GMM to hyperspectral images of wheat, resulting in accurate segmentation of seeds and embryos with a Jaccard Index of 0.745, demonstrating the advantages of GMM in handling unstructured multiview data [8].

Although this methodology has proven effective across various domains, the application of GMM for clustering cocoa-producing regions in Southeast Aceh at the local level has not yet been conducted. This study aims to fill that gap by applying GMM to identify production clusters based on village-level data. The results are expected to serve as a foundation for more accurate, objective, and data-driven evaluations and agricultural policy recommendations.

II. RELATED WORKS

A. Data Mining

Data mining is the process of discovering new patterns from large datasets using statistical methods and artificial intelligence, allowing users to efficiently access and analyze large volumes of data [9]. By employing various techniques and algorithms, data mining helps identify trends, classify data, and generate predictions that support decision-making processes. This process involves several stages, such as data preprocessing, modeling, evaluation, and result interpretation. One of the commonly used techniques in data mining is clustering [10], which functions to group data based on similarity.

B. Clustering

Clustering is a technique or method used to group data [11]. This technique divides data into several groups that share similar characteristics and assigns data with differing characteristics to different groups [12]. In the process of grouping the data in machine learning, there are several clustering algorithms to use [13], such as the Gaussian Mixture Model.

The Gaussian Mixture Model (GMM) is a probability-based clustering model that utilizes weighted combinations of multiple normal distributions, which is why it is often referred to as a mixture component model. According to [5], the clustering process using the Gaussian Mixture Model is carried out through calculations using the following formula.

1. The determination of the optimal number of clusters using the Bayesian Information Criterion (BIC) is carried out using the following equation:

$$BIC = -2 \log f(x) + s \log N \quad (1)$$

Description:

Log f (x)	: *log-likelihood of the Gaussian distribution*
S	: number of parameters
N	: number of observations

2. The well-established paradigm of model-based clustering assumes that data are generated from a finite mixture model, where each component represents a cluster. The Gaussian Mixture Model (GMM), in particular, is widely used in various applications. Expectation Maximization (EM) and its variants are the most commonly used methods for estimating

parameters using the Maximum Likelihood approach [14]. In GMM, the probability density function for a data point x is defined as follows:

$$p(x) = \sum_{k=1}^{K} \pi_k \, .N\,(x\,|\mu_k, \sum_k) \qquad (2)$$

Description:

P(x)	: the probability that data x comes from the Gaussian mixture model.
K	: the number of Gaussian components (clusters).
π_k	: the weight of each cluster.
μ_k	: the mean of each cluster.
$\sum_k$	: the covariance of the data within the cluster.
N	: the k-th Gaussian distribution

C. Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm is one of the methods used in the process of classification or data clustering [15]. his algorithm works iteratively through two main steps: the Expectation step (E-step) and the Maximization step (M-step) [16]. The EM method is particularly suitable for Gaussian Mixture Models (GMM) because it can estimate the model parameters by iteratively constructing a lower bound of the likelihood function and maximizing it to improve the overall likelihood value. This process enables the model to handle hidden variables in the clustering structure and refine parameter estimates more accurately with each iteration [17].

1. E-Step (*Expectation Step*)

   The expectation process (E-step) is a function used to estimate the evaluation of the likelihood based on the existing parameters [18]. The E-step is calculated using the following formula:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \sum_j)} \qquad (3)$$

   Description:

γ_{ik}	: Responsibility or probability
π_k	: Weight of cluster *k*
μk	: Mean vector of the *k*-th cluster
$\mathcal{N}(x_i \mid \mu_k, \sum_k)$	: Gaussian probability density function
$\sum_\kappa$	: Covariance matrix of the *k*-th cluster

2. M-Step (Maximization)

   To maximize the likelihood value obtained in the E-step, the M-step is carried out by updating the model parameters such as the mean, covariance, and cluster weights based on the membership probabilities obtained during the E-step. The goal is to make the model better fit the data distribution, and this process is repeated until the model reaches convergence [18]. The calculations are performed using the following formulas:

   a. Mean

$$\mu_k = \frac{\sum_{i=1}^{N} \Upsilon_{ik} x_i}{\sum_{i=1}^{N} \Upsilon_{ik}} \qquad (4)$$

   Description:

γ_{ik}	: Probability that data point *i* belongs to cluster *k*
x_i	: The *i*-th data point
N	: Total number of data points

   b. Covariance

$$\sum_k = \frac{\sum_{i=1}^{N} \gamma_{ik} \, (x_i - \mu_k) \, (x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}} \qquad (5)$$

   Description:

$\sum_k$: Covariance matrix of cluster k
γ_{ik} : Probability that data point x_i belongs to cluster k
$(x_i - \mu_k)$: Difference between data point x_i and the mean of cluster k
$(x_i - \mu_k)(x_i - \mu_k)^T$: Matrix representing data dispersion, formed by the outer product of the deviation vector and its transpose

c. Weight

$$\pi_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{N} \quad (6)$$

Description:

π_k : Weight for cluster k
$\sum_{i=1}^{N} \gamma_{ik}$: Total responsibility for cluster k
N : Total number of data points

D. Normalisasi

Each attribute in agricultural data has different units and scales—for example: area in hectares, production in tons, and the number of farmers in people. Without normalization, attributes with larger values (e.g., production) may dominate those with smaller values (e.g., productivity), leading to biased clustering results. One commonly used normalization method is Z-Score normalization, which is based on the mean and standard deviation [19]. This method helps reduce the influence of outliers. The equation for calculating Z-Score normalization is shown below [20].

$$Z = \frac{X - \mu}{\sigma} \quad (7)$$

Description :

Z : Normalized value
X : Original data value
μ : Mean of the attribute
σ : Standar deviasi

III. METHODOLOGY

A. Research Flow

This study follows a structured sequence using the waterfall approach, starting from problem identification to data visualization and result interpretation, as illustrated in Figure 1.

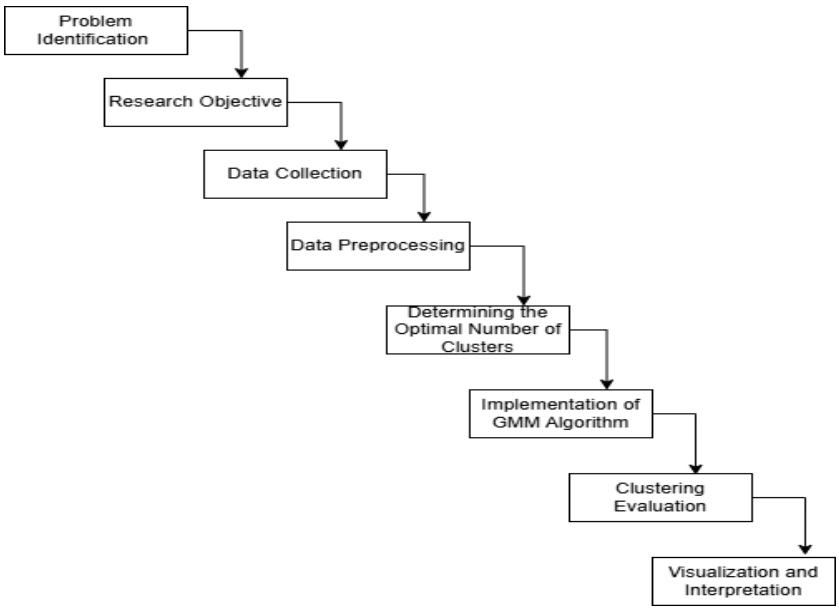

Figure 1 Research Flowchart

B. System Scheme

The system flow illustrates the clustering process using the Gaussian Mixture Model (GMM) algorithm, with the optimal number of clusters determined based on BIC calculation. The process begins with data input, followed by data preprocessing to ensure it is ready for analysis. Next, BIC is computed to determine the best number of clusters, and the GMM algorithm is applied using that number. The final stage includes model evaluation to assess the clustering results. This system flow is shown in Figure 2.

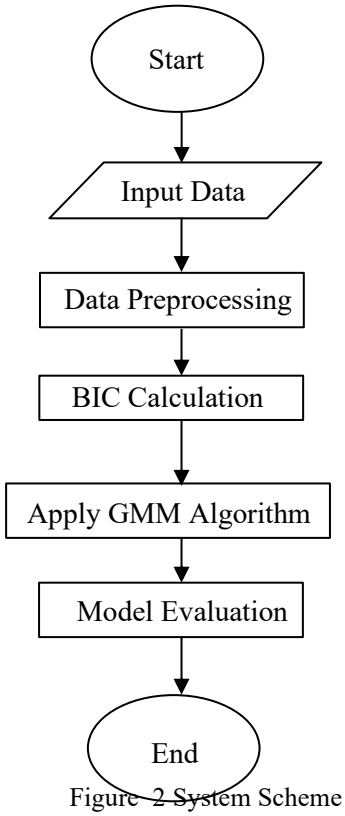

Figure 2 System Scheme

IV. RESULT AND DISCUSSION

This study produced a clustering system for cocoa-producing areas in Southeast Aceh using the GMM algorithm, developed as a web-based application with Python and Flask. The process applies Z-Score normalization to improve accuracy, enabling the clustering results to support decision-making in regional development and resource distribution.

A. Training Dataset Analysis

Before the clustering process is carried out, the data is first collected and organized in tabular form to facilitate processing and analysis. The dataset used in this study is shown in the following table 1:

Table 1. Research Dataset

No.	Kecamatan	Desa	TBM	TM	TTM	Luas Lahan (Ha)	Produksi (Ton)	Produktivitas (Kg/ha)	Petani (KK)
1.	Lawe Alas	Muara Baru	108	34	4	146	31	912	50
2.	Lawe Alas	Pasir Bangun	100	34	2	136	31	912	49
3.	Lawe Alas	Ngkeran	90	35	4	129	31	886	50
4.	Lawe Alas	Lawe Kongkir	80	32	3	115	31	969	55
5.	Lawe Alas	Kubu	110	30	4	144	31	1033	49
…..	…….	……	……	…..	……	……	……	……..	……..
381.	Deleng Pokhisen	Tading Niulihi	22	36	6	64	33	917	59
382.	Deleng Pokhisen	Peseluk Pesimbe	30	34	6	70	31	912	60
383.	Deleng Pokhisen	Kati Jeroh	27	33	7	67	31	939	59
384.	Deleng Pokhisen	Kane Lot	34	34	6	74	33	917	59
385.	Deleng Pokhisen	Lawe Hakhum	28	36	8	72	31	861	60

Table 1 displays data from 385 cocoa-producing villages in Southeast Aceh Regency, covering variables such as sub-district, village name, land area, number of immature (TBM), mature (TM), and damaged (TTM) cocoa plants, production volume, productivity, and number of farmers.

These attributes reflect the agricultural capacity and productivity of each village and serve as the basis for clustering analysis in this study.

1. Z-Score Normalization

Normalization is necessary because each attribute in the dataset has different scales and units, such as hectares, tons, and number of individuals. These scale differences can cause bias in the clustering process, where attributes with larger values may dominate the results. Therefore, Z-Score normalization is applied to standardize all features, allowing the Gaussian Mixture Model (GMM) algorithm to perform clustering in a fair and balanced manner.

2. BIC Calculation

To determine the optimal number of clusters, the Bayesian Information Criterion (BIC) was used as a model selection metric. Several clustering models with different numbers of clusters were evaluated, and the model with the lowest BIC value was selected. Based on this evaluation,

the GMM model with three clusters produced the lowest BIC score, indicating the best model fit for the data.

3. Calculation Using the Expectation-Maximization (EM) Algorithm

The clustering process using the Gaussian Mixture Model (GMM) was carried out through the Expectation-Maximization (EM) algorithm, which iteratively updated cluster parameters to maximize the likelihood function. After 6 iterations, the model successfully converged and produced three clusters with distinct characteristics. Evaluation of the clustering results yielded a Silhouette Score of 0.3893 and a Davies-Bouldin Index of 0.8548, indicating that the model was able to form well-separated and reasonably compact clusters suitable for further analysis. The final clustering results show that Cluster 1 (high production) consists of 251 villages, Cluster 2 (medium production) consists of 116 villages, and Cluster 3 (low production) consists of 18 villages.

B. Website Implementation

The Plantation Data Page displays information on land and production outputs in Southeast Aceh Regency, including sub-district and village names, land area (immature, mature, and damaged plants), total area, production, productivity, and the number of farmers. The page interface is designed to be simple and responsive, equipped with a search feature and an "Add Data" button for adding new entries. This data serves as the basis for clustering analysis using the Gaussian Mixture Model (GMM) algorithm, aiming to group regions based on similar agricultural characteristics to support more targeted policy-making.

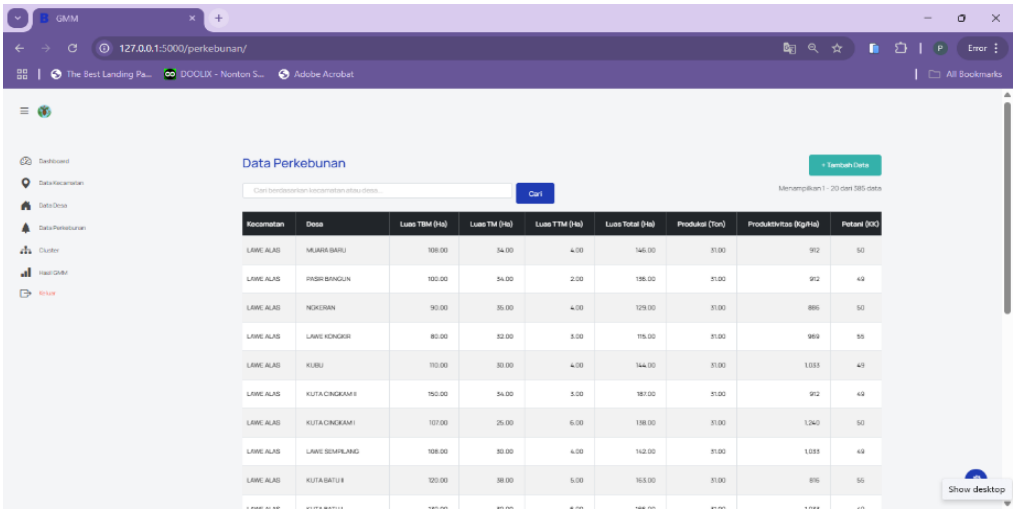

Figure 3 Display of Plantation Data for Each Village in Southeast Aceh Regency

The GMM Results Page displays the output of the clustering process using the Gaussian Mixture Model algorithm based on village-level data from Southeast Aceh Regency. On this page, users can view the total number of data points analyzed, the number of iterations required until convergence, and the number of clusters formed. Each village is classified into a specific cluster, such as High Production, along with its membership probability, iteration status, and options to view details or delete data. This interface helps users understand the regional grouping results based on production characteristics in a visual and systematic manner.

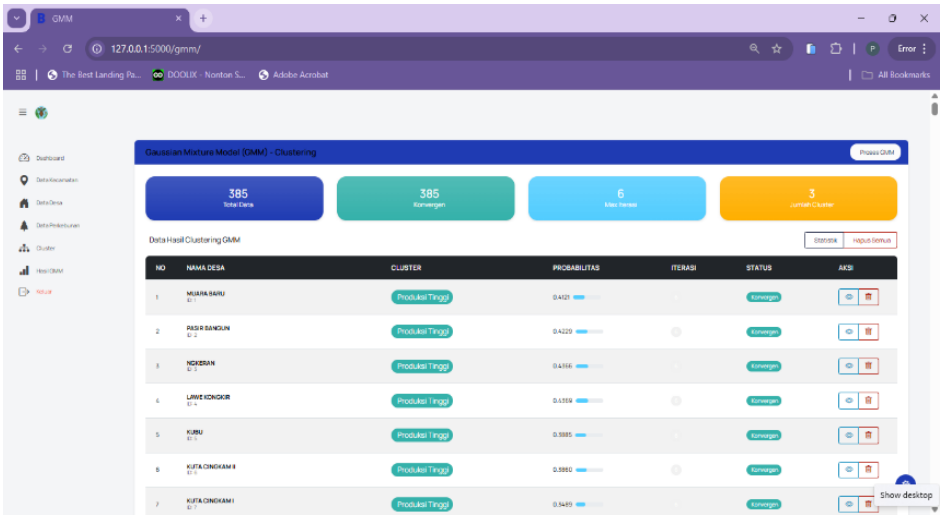

Figure 4 GMM Results Page Displaying the Output of the Clustering Process

The GMM Evaluation Page presents a visualization of the clustering results in the form of a scatter plot based on two main variables: productive planting area (Ha) and production (Tons). Each point represents a village, with colors indicating the cluster: High Production (green), Medium Production (yellow), and Low Production (red). On the right side, a summary is displayed showing the number of villages per cluster, the average membership probability, as well as the minimum and maximum probabilities within each cluster. The overall average probability of the model is 0.53, indicating that the GMM successfully formed fairly consistent clusters in separating villages based on production and planting area characteristics.

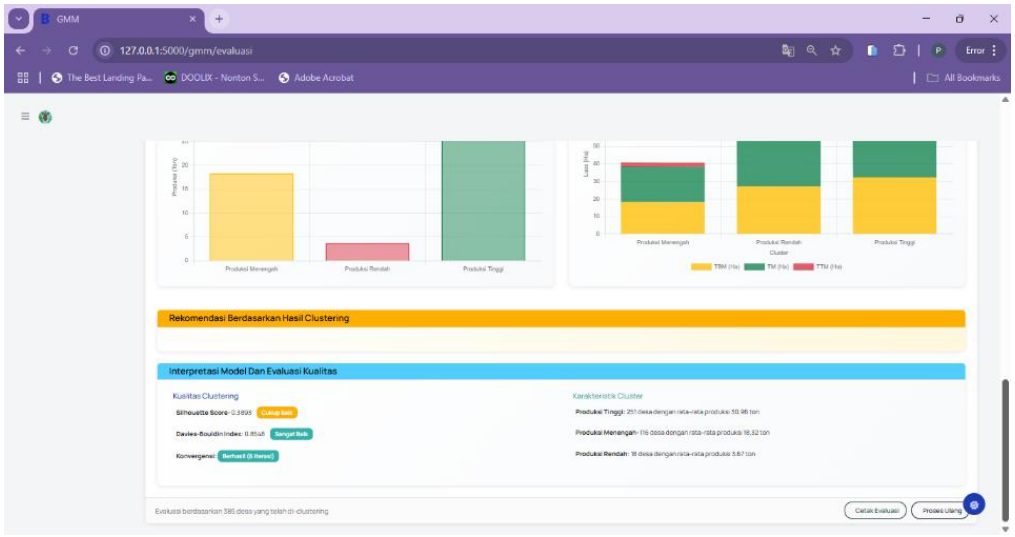

Figure 5 GMM Evaluation Page with Cluster and Performance Overview

This page presents the evaluation results of GMM clustering quality both visually and numerically. Two bar charts at the top compare the average production and the composition of land area (immature, mature, and damaged plants) for each cluster. Below, model evaluation metrics are displayed: a Silhouette Score of 0.3893 (fairly good) and a Davies-Bouldin Index of 0.8548 (very good), along with a model convergence status indicating successful convergence in 6 iterations. The right section provides a summary of each cluster's characteristics, where the

High Production cluster includes 251 villages with the highest average production, while the Low Production cluster includes only 18 villages. This information supports the validity of the clustering results and serves as a basis for formulating policy recommendations.

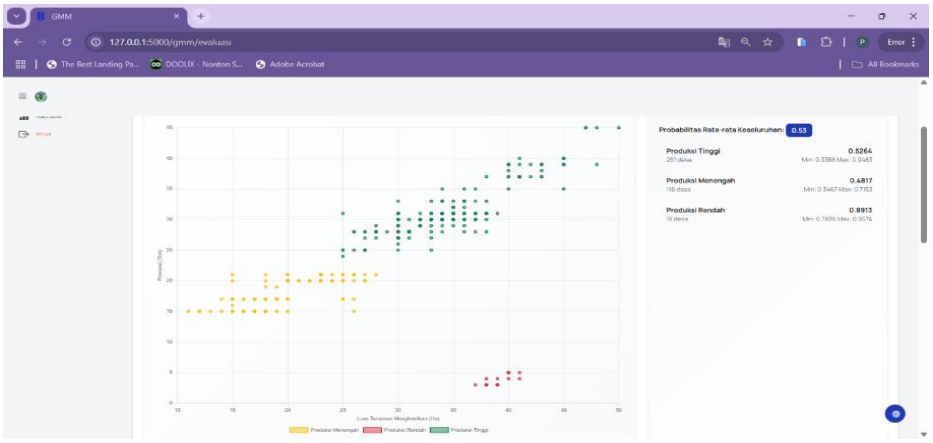

Figure 6 Model Evaluation Page for GMM Clustering

The implementation of this website serves as a visual representation of the clustering output for agricultural data in Southeast Aceh Regency, developed using the Gaussian Mixture Model (GMM) algorithm. Pages such as clustering results, model evaluation, and visualization of probabilities and cluster distribution are designed to help users better understand the analysis outcomes. This implementation also functions as an evaluation tool to assess whether the system is performing effectively or still requires improvements or additional features to support more accurate and data-driven decision-making.

C. Discussion

In the development of an agricultural data clustering system using the Gaussian Mixture Model (GMM) algorithm, ethical considerations must also be taken into account, particularly in data-driven decision-making. The data used in the system must be kept confidential to prevent misuse by irresponsible parties. Moreover, the process of regional clustering must be carried out objectively and fairly, without causing discrimination against any specific region or group.

This system is not intended to fully replace the role of analysts or policymakers, but rather to serve as a supporting tool in facilitating faster and data-based decision-making. Therefore, the clustering results should still be carefully reviewed by the appropriate authorities before being used as a basis for policy formulation or field interventions.

V. Conclusion

This study successfully clustered agricultural regions in Southeast Aceh Regency using the Gaussian Mixture Model (GMM) algorithm, based on key attributes such as land area, production volume, productivity, and the number of farmers. The process involved data preprocessing, normalization with Z-Score, and selection of the optimal number of clusters using the Bayesian Information Criterion (BIC). The analysis resulted in three distinct clusters—High, Medium, and Low Production—which offer meaningful insights for regional agricultural planning. The High Production cluster includes 251 villages, demonstrating a significant concentration of agricultural output. The clustering model achieved acceptable evaluation scores, indicating its suitability for further interpretation and application. A web-based information system was developed to display agricultural data, clustering outcomes, and evaluation metrics interactively. This system provides practical value in assisting agricultural authorities and local governments in identifying priority areas and implementing data-driven policy interventions. For future work, integrating time-series data, incorporating geospatial mapping, and comparing GMM with alternative clustering methods

such as DBSCAN or hierarchical clustering are recommended to enhance model robustness and analytical depth.

REFERENCES

[1] J. Porsiana, J. Riry, and M. K. Lesilolo, "Pengujian Kadar Air Biji Kakao dengan Suhu Tinggi dan Rendah terhadap Kualitas Biji Kakao (Testing the moisture content of cocoa beans at high and low temperatures on the quality of cocoa beans)," *J. Pertan. Kepul.*, vol. 8, no. 1, pp. 7–12, 2024.

[2] Muhammad Fajar Zikri, Cut Mulyani, and Yenni Marnita, "Tingkat Serangan Penyakit Busuk Buah Kakao (Phytophthora Palmivora L.) Dan Kehilangan Hasil Tanaman Kakao Di Kecamatan Darul Ihsan Kabupaten Aceh Timur," *J. Penelit. Agrosamudra*, vol. 9, no. 2, pp. 11–20, 2022, https://doi.org/10.33059/jupas.v9i2.6565.

[3] R. Ariani, "Jurnal Sains Pertanian Analisis produktivitas kakao di Kabupaten Aceh Tenggara Analysis of cocoa productivity in Southeast Aceh District," vol. 6, pp. 96–98, 2022.

[4] J. Riyono, S. D. Puspa, and C. E. Pujiastuti, "Simulasi Clustering Provinsi di Indonesia dalam Penyebaran Covid-19 Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Algoritma Gaussian Mixture Model," *Majamath J. Mat. dan Pendidik. Mat.*, vol. 5, no. 1, pp. 43–60, 2022.

[5] N. N. Alyarahma, G. Kholijah, and C. Sormin, "Pengelompokan Provinsi di Indonesia Menggunakan Gaussian Mixture Model Berdasarkan Indikator Kemiskinan," vol. 6, no. 2, pp. 158–167, 2024, https://doi.org/10.31605/jomta.v6i2.4032.

[6] D. Siregar, W. Rahayu, B. M. Wardana, Ketrin Natasya Stefany, and Bayu Wibisono, "Characteristics of Provinces in Indonesia Based on JKN Indicator Outcomes by Gaussian Mixture Model with Expectation-Maximization Algorithm and Biplot," *J. Stat. dan Apl.*, vol. 8, no. 1, pp. 17–30, 2024, https://doi.org/10.21009/jsa.08102.

[7] M. Harahap, A. W. D. R. Zamili, M. A. Arvansyah, E. F. Saragih, S. Rajen, and A. M. Husein, "K-Means Clustering Algorithm Approach in Clustering Data on Cocoa Production Results in the Sumatra Region," *J. Resti*, vol. 6, no. 6, pp. 905–910, 2022, https://doi.org/10.29207/resti.v6i6.4199.

[8] A. Kartakoullis, N. Caporaso, M. B. Whitworth, and I. D. Fisk, "Gaussian mixture model clustering allows accurate semantic image segmentation of wheat kernels from near-infrared hyperspectral images," *Chemom. Intell. Lab. Syst.*, vol. 259, no. February, p. 105341, 2025, https://doi.org/10.1016/j.chemolab.2025.105341.

[9] F. Riza, S. Safwandi, and S. Fadlan, "Klasifikasi Varietas Kopi Berdasarkan Kondisi Tanah dan Suhu Menggunakan Algoritma Gaussian Naïve Bayes," *J. Algoritm.*, pp. 312–323, 2025, https://doi.org/10.33364/algoritma/v.22-1.2311.

[10] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," *J. Ilm. Inform. dan Ilmu Komput.*, vol. 3, no. 1, pp. 46–56, 2024, https://doi.org/10.58602/jima-ilkom.v3i1.26.

[11] N. Nur Afidah, "Penerapan Metode Clustering dengan Algoritma K-means untuk Pengelompokkan Data Migrasi Penduduk Tiap Kecamatan di Kabupaten Rembang," *Prism. Pros. Semin. Nas. Mat.*, vol. 6, pp. 729–738, 2023.

[12] Taufik Hidayat, Mohamad Jajuli, and Susilawati, "Clustering daerah rawan stunting di Jawa Barat menggunakan algoritma K-Means," *Infotech J. Inform. Teknol.*, vol. 4, no. 2, pp. 137–146, 2023, https://doi.org/10.37373/infotech.v4i2.642.

[13] R. K. Dinata, N. Hasdyna, S. Retno, and M. Nurfahmi, "K-means algorithm for clustering system of plant seeds specialization areas in east Aceh," *Ilk. J. Ilm.*, vol. 13, no. 3, pp. 235–243, 2021, https://doi.org/10.33096/ilkom.v13i3.863.235-243.

[14] S. R. Kasa and V. Rajan, "Avoiding inferior clusterings with misspecified Gaussian mixture models," *Sci. Rep.*, vol. 13, no. 1, pp. 1–13, 2023, doi: 10.1038/s41598-023-44608-3.

[15] Y. L. Nainel, E. Buulolo, and I. Lubis, "Penerapan Data Mining Untuk Estimasi Penjualan Obat Berdasarkan Pengaruh Brand Image Dengan Algoritma Expectation Maximization (Studi Kasus: PT. Pyridam Farma Tbk)," *Jurikom (Jurnal Ris. Komputer)*, vol. 7, no. 2, p. 214, 2020, https://doi.org/10.30865/jurikom.v7i2.2097.

[16] Suhada, G. L. Ginting, and R. K. Hondro, "Penerapan Data Mining Untuk Memprediksi Besarnya Pembayaran Pajak Kendaraan Pada: (Badan Pengelolaan Pajak Dan Retribusi Daerah Upt Samsat Medan Selatan) Menggunakan Algoritma Expectation Maximization," *Bull. Inf. Technol.*, vol. 2, no. 2, pp. 69–75, 2021.

[17] J. You, Z. Li, and J. Du, "A new iterative initialization of EM algorithm for Gaussian mixture models," *PLoS One*, vol. 18, no. 4 April, pp. 1–17, 2023, https://doi.org/10.1371/journal.pone.0284114.

[18] N. W. Yanto, H. Sukoco, and S. N. Neyman, "Pemodelan Proxy Anonim Menggunakan Algoritme Expectation Maximization Dengan Data Balancing," *J. Ilmu Komput. dan Bisnis*, vol. 12, no. 1, pp. 60–69, 2021, https://doi.org/10.47927/jikb.v12i1.91.

[19] P. P. Allorerung, A. Erna, and M. Bagussahrir, "Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," vol. 9, no. 3, pp. 178–191, 2024.

[20] I. Permana and F. N. S. Salisah, "Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," *Indones. J. Inform. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022, https://doi.org/10.57152/ijirse.v2i1.311.