Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 113-122, February 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.55

RESEARCH PAPER

OPEN ACCES

Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports

Dwi Arman Prasetya[1], Anggraini Puspita Sari[2], Mohammad Idhom[1], and Angela Lisanthoni[1]

[1] Department of Data Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
[2] Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia

ABSTRACT

Agriculture is a key contributor to Indonesia's economic growth, with tubers representing the second most important food crop. Despite their significance, the export value of Indonesia's tuber crops has not yet reached its full potential given the decline in the value of tuber exports since 2021. One of the contributing factors is the restricted range of export market options. This study aims to analyze export trade patterns to identify the most high-potential markets for Indonesian tuber commodities. Clustering analysis is used as a key method to identify market locations by grouping countries based on similar trade characteristics. Clustering was conducted using the Gaussian Mixture Model (GMM), which enhanced by Particle Swarm Optimization (PSO) and evaluated by silhouette score and DBI. The dataset is collected from Indonesia's Central Bureau of Statistics from 2019 to 2023, focusing on 5 kinds of tuber exports with total of 455 entries and 8 columns. Using the AIC/BIC method, the optimal number of clusters obtained is 2 which are low market opportunities (cluster 0) and high market oppurtunities (cluster 1). Results showed that the GMM model without optimization has silhouette score of 0.7602 and DBI of 0.8398, while the GMM+PSO model achieved an improved silhouette score of 0.8884 and DBI of 0.5584. Both score are categorized as strong structure but, GMM+PSO has higher silhouette score and lower DBI score, demonstrating the effectiveness of PSO in enhancing the clustering model's performance. The key potential markets for Indonesian tuber exports are primarily concentrated in Asia, including countries such as China, Malaysia, Thailand, Vietnam, Hong Kong, and United States.

PAPER HISTORY

Received Dec. 10, 2024
Revised Jan. 20, 2025
Accepted Feb. 10, 2025
Published Feb. 25, 2025

KEYWORDS

Gaussian Mixture Model;
Silhouette Score;
Particle Swarm Optimization;
Export;
Clustering

CONTACT:

[2]anggraini.puspita.if@upnjatim.ac.id

1. INTRODUCTION

Agriculture plays a significant role in driving Indonesia's economic growth, as approximately 30% workforce from the agricultural sector and around 14.2% contributed to the nation's gross domestic product [1]. As an agrarian country, Indonesia's agricultural sector, particularly food crops, plays a crucial role in sustaining domestic consumption and contributing to international trade. It enhances foreign exchange earnings, which supports infrastructure improvements, and finance imports [2]. Tubers are the most significant food crops, holding the position of the second most important group. One of example of a tuber is cassava, which is the fourth most important crop in developing nations and the second-largest source of starch worldwide [3]. With the abundance of tubers, it can be leveraged to enhance export value. However, the condition that occurs is that the value of Indonesia's root crop exports has not yet reached an optimal level. Data from The Central Bureau of Statistics stated that by 2021, Indonesia's export of root crops experienced a sharp decrease with only reaching 28,797,892 kg and a trade value of 19,606,392 USD. By 2023, no significant recovery had been observed [4]. A key factor contributing to this issue is the restricted number of export destination markets. To address this, a potential strategy is to diversify the markets for exports [5], [6]. However, there is currently no efficient approach for identifying the most promising markets for root crops commodities. Therefore, this research will focus on analyzing trade patterns to locate the most suitable market locations.

Clustering serves as an approach for identifying market locations by categorizing data based on similar attributes [7]. Previous studies have explored various clustering approaches for export commodities. For instance, [8] clustered crumb rubber export product with Central Bureau of Statistics data from 2012 to 2022 using K-Means. The results obtained 2 clusters with silhouette score of 0.7325. [9] clustered frozen shrimp export products using data from the Central Bureau of Statistics for the years 2022 to 2023 with the K-Medoids algorithm. The results obtained 2 clusters with silhouette score of 0.725 and 0.755. There are other studies such as clustering of export products using the K-means and K-Medoids algorithms [10], clustering crude petroleum materials export product using the K-Medoids algorithm [11], clustered export commodities based on destination continents for 15 commodities [12], and more.

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

K-means is a widely used clustering algorithm, but it has limitations, particularly when dealing with large datasets or data with non-convex structures [13]. To address these challenges, this study will employ the Gaussian Mixture Model (GMM) which offers more flexibility by handling elliptical cluster shapes and capturing more complex data patterns through its probabilistic approach [14]. According to research by [15], GMM is more effective than K-means in identifying complex, non-linear patterns and managing overlapping clusters, making it a superior option for clustering. To improve the model's performance, the Particle Swarm Optimization (PSO) algorithm will be implemented in this study to determine the optimal model parameters. According to research by [16], PSO has been proven to enhance the accuracy and quality of clustering results, as being able to search for optimal features from the global feature set. Research by [17] [18], proves that GMM can be used to cluster markets in various fields other than exports which are suitable for market analysis.

This research focuses on the implementation of GMM and PSO to identify the most potential markets for root crops. The novelty of this study lies in the application of the GMM-PSO approach for clustering Indonesian export commodities, which currently has not been widely applied yet. Unlike previous research focused solely on sweet potato competitiveness, this study expands the analysis to include other tuber crops like cassava, yams, taro, and sago, offering broader insights into Indonesia's tuber exports. The findings provide valuable insights for the Indonesian government and businesses to refine export strategies. Exporters can shift focus to high-potential markets with strong demand for tuber exports, apply similar sales approaches to countries with shared trade patterns, and identifies the most in-demand tuber commodity. This allows exporters to focus strategically optimize production and resource allocation for higher trade value.

2. MATERIALS AND METHOD

A. Gaussian Mixture Model (GMM)

The clustering method based on statistical models assumes that the data is a mixture of probability distributions, with each distribution representing a different cluster. When each combined statistical model follows a Gaussian distribution, it is referred to as GMM. If the actual data within clusters is not normally distributed, GMM may struggle to correctly capture the clusters. But, the parameter can be adjusted following the data. In the multivariate case, GMM using three parameters: mean (μ), covariance (Σ), and weights (w). The formula of GMM is shown in Eq. (1) [19].

$$p(x) = \sum_{j=1}^{k} w_j P(\boldsymbol{X}|\mu_j, \Sigma_j) \quad (1)$$

Where, k is the number of clusters, $\boldsymbol{X}$ is the data used for this research, $P(X|\mu_j, \Sigma_j)$ is the probability that data **x** belongs to a cluster, and $p(x)$ is the joint probability function of the data **x**.

GMM has parameters that use the Expextation-Maximation (EM) algorithm that works iteratively to optimize the estimation of the three GMM parameters [20]. GMM has the following process [15]:

1) Set the number k of clusters to be used. Next, initialize each GMM parameter including weight (w_j), mean (μ_j) with a vector dimension (d), and covariance (Σ_j) for j = 1, 2, ..., k.
2) E-step: This step involves computing the likelihood of how the data ($\boldsymbol{x}_i$) is fits into each Gaussian component (C_j) based on the current parameters. Then, the probability that the data point $\boldsymbol{x}_i$ belongs to a cluster C_j is calculated using Eq. (2) [15]:

$$Z_{ij} = \frac{\rho(\boldsymbol{x}_i|C_j).w_j}{\rho(\boldsymbol{x}_i)} \quad (2)$$

Where, $\rho(\boldsymbol{x}_i|C_j)$ is the probability $\boldsymbol{x}_i$ generated by the Gaussian distribution of the cluster C_j, w_j is the weight of the cluster C_j.

Then, calculate the value of the likelihood $\rho(\boldsymbol{x}_i|C_j)$ and total probability $\rho(\boldsymbol{x}_i)$ with Eq. (3) and Eq. (4) [15]:

$$\rho(x_i|C_j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \mu_j)^T \Sigma_j^{-1}(\boldsymbol{x}_i - \mu_j)\right) \quad (3)$$

$$\rho(x_i) = \sum_{j=1}^{k} \rho(x_i|C_j).w_j \quad (4)$$

3) M-step: concentrates on adjusting all three model parameters to optimize the likelihood. The formulas are shown in Eq. (5), Eq. (6), and Eq. (7) [15].

$$\mu_j^{new} = \frac{\sum_{i=1}^{n} z_{ij}.\boldsymbol{x}_i}{\sum_{i=1}^{n} z_{ij}} \quad (5)$$

$$\Sigma_j^{new} = \frac{\sum_{i=1}^{n} z_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} z_{ij}} \quad (6)$$

$$w_j^{new} = \frac{1}{n}\sum_{i=1}^{n} z_{ij} \quad (7)$$

4) Repeat steps 2 and 3 until convergence

B. Particle Swarm Optimization (PSO)

Particle Swarm Optimization is an optimization technique inspired by the collective behavior of birds in flocks or fish in schools. The main objective of PSO is to identify the optimal solution that best optimize the objective function by efficiently exploring the search space [21]. Here are the steps for using the PSO algorithm [22]:

1) Initialization of Particles
   Define the number of particles in the swarm and initialize each particle with a random position and velocity within the search space. Each particle has two important values: personal best (*pBest*), which is the best position found by that particle itself during the iterations, and global best (*gBest*), which is the best position found by the entire swarm.
2) Objective Function Evaluation
   Then evaluate the objective function or fitness function for each particle based on its position. If the current position is better than the particle's personal best, the personal best is updated. Additionally, if the

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

particle's position is better than the global best found by the swarm, the global best is updated as well.

3) Update Parameter Velocity
The next step is to update the velocity of each particle. The velocity is updated based on three main components: inertia, which helps maintain the particle's previous movement direction, *pBest* guiding the particle back to its own best-found position, and *gBest*, guiding the particle toward the best-found position in the swarm. The updated velocity is calculated using a mathematical formula shown in Eq. (8) [22].

$$v_i^{new} = w \,.\, v_i^{old} + c_1 \,.\, r_1 \,.\, (pBest_i - x_i) + c_2 \,.\, r_2 \,.\, (gBest - x_i) \quad (8)$$

4) Update Parameter Position
Once the velocity is updated, the position of each particle is updated accordingly using Eq. (9) [22].

$$x_i^{new} = x_i^{old} + v_i^{new} \quad (9)$$

5) Repeat the process until reach maximum iteration.

C. Silhoette Score

The silhouette score is a method used to assess the quality of a clustering model, with values ranging from -1 to 1. A higher score indicates better clustering [23]. The evaluation of the silhouette score can be categorized into four groups shown in Table 1 [24].

Table 1. Silhoeuette Score Categories

Category	Score	Criteria
1	0,71 – 1,00	Strong Structure
2	0,51 – 0,70	Good Structure
3	0,26 – 0,50	Weak Structure
4	≤ 0,25	Bad Structure

The silhouette score evaluates the average distance between data points within the same cluster and compares it to the average distance between points and those in neighboring clusters [25] with the formula provided in Eq. (10) [26].

$$silhoette(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))} \quad (10)$$

Where, *a(x)* is the mean distance between object *x* and all other objects within the same cluster and *b(x)* is the mean distance between object *x* and all other objects in neighboring clusters.

D. Davies-Bouldiun Index (DBI)

The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering by balancing two factors which are minimizing intra-cluster distances (similarity within clusters) and maximizing inter-cluster distances (differences between clusters). A lower DBI indicates better clustering performance, with more compact and well-separated clusters [27].The formula of DBI provided in Eq. (11) [28].

$$DBI = \frac{1}{n} \sum max \left(\frac{R_i + R_j}{d(C_i, C_j)} \right) \quad (11)$$

Where, n represents the number of clusters, R_i is the average distance between object in cluster i and the center of cluster i, $d(C_i, C_j)$ is the distance between the centers of cluster i and cluster j.

E. Research Methodology

Fig. 1 illustrates the research methodology process, which consists of several stages including data collection, preprocessing data, determining the best cluster using AIC/BIX, building GMM and PSO model, and the model is evaluated using silhouette score and DBI.

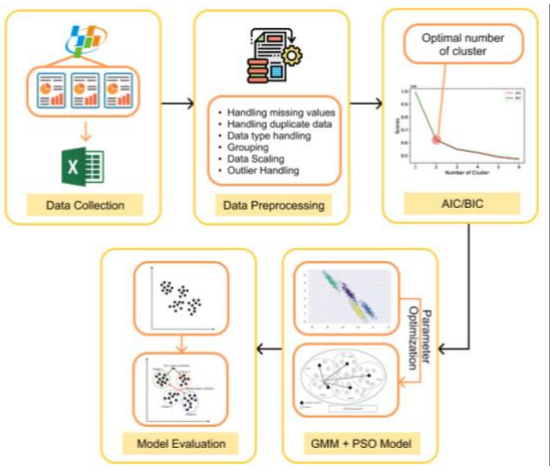


Fig. 1. Research Flow

1) Data Collection
The data used in this study are sourced from the Central Bureau of Statistics from 2019 to 2023 [4]. The data collection process was conducted manually and saved in Excel. Given that Central Bureau of Statistics is a trusted governmental organization, the data used in this study is assumed to be reliable without requiring additional validation. The commodity data is classified based on the Harmonized System (HS) code. This study specifically focuses on tuber crop export data at the HS6 level, which includes five HS6 codes: 071410 for cassava, 071420 for sweet potato, 071430 for yams, 071440 for taro, and 071490 for sago. The dataset consists of 455 entries, containing 8 columns which are "HS6 ID", "HS6", "HS8 ID", "HS8", "Country", "Year", "Quantity (in kg)", and "Trade Value (in US$)".

2) Preprocessing Data
Data preprocessing consists of employing various methods to enhance the quality of raw data to ensure the accuracy and result [29]. Since clustering is unsupervised learning, splitting data is not required. In this research, multiple preprocessing techniques are applied, including:

- Handling missing values aims to prevent analysis errors, which is addressed by imputing with the median value [30].

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

- Handling duplicate data ensures the integrity and efficiency of the data analysis, achieved by using the *drop_duplicates()* function.
- Data type handling ensures that each column in the dataset has the correct data type.
- Grouping the data by HS6 ID and country.
- Data scaling is applied to normalize the data, making the points more uniform. In this study, Min Max Scaler is the chosen method for scaling [31].
- Outlier handling aims to minimize errors in the analysis caused by unrepresentative data [32]. In this research, outlier detection is carried out using the z-score method with threshold used is 3, meaning that any data point with a z-score outside the range of -3 < x < 3 is considered an outlier and is removed from the dataset. This is because such data points are deemed to be too far from the overall data distribution, potentially skewing the analysis.

3) AIC and BIC

The optimal number of clusters will be determined using two methods, namely AIC and BIC, which have the formulas shown in Eq. (12) and Eq. (13) [33].

$$AIC = 2\,x\,k - 2\,x\,LL \quad (12)$$

$$BIC = \log_e(N) * k - 2 * LL \quad (13)$$

Where *k* is the number of clusters, *N* is the number of data points, and *LL* is the maximum likelihood of the objective function. The optimal number of clusters is identified when the AIC and BIC scores show a sharp decline, followed by a flattening, indicating that additional clusters no longer significantly improve the model.

4) Creating GMM Model and PSO Optimization

Clustering will be based on "quantity" and "trade value" as the primary factors, with the number of clusters determined through the AIC/BIC process. The clustering approach in this study is GMM, utilizing the PSO optimization algorithm to find the optimal GMM model parameters. In this study, all parameters of GMM and PSO including parameters of each cluster, the number of particles, the maximum number of iterations, inertia weight, acceleration coefficients, and velocity will used default from python library. This setting will serve as a starting point, with further fine-tuning attempted if necessary. It because the default parameter is more efficient, quick implementation, and have been shown to work well fo a wide range of problems.

5) Model Evaluation

The model is evaluated using silhouette score and DBI. The performance of the GMM is compared with the GMM + PSO model to assess whether the application of PSO improves the model's performance.

3. RESULTS

A. Result of Preprocessing Data

Table 2 presents the processed data, which consists of 88 entries and 5 columns: "HS6 ID", "HS6", "Country", "Trade Value", and "Quantity".

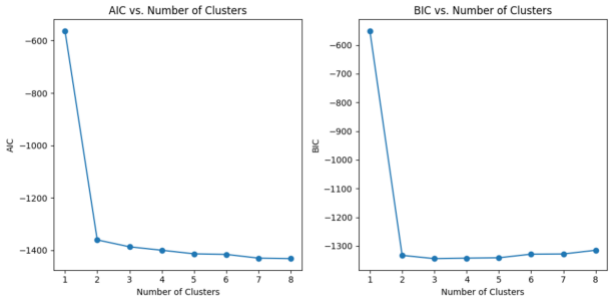
Table 2. Result of pre-processing data

HS6 ID	HS6	Country	Trade Value	Quantity
071410	Manioc (Cassava) (Fresh/Dried)	Australia	0.001108	0.000467
071420	Sweet Potatoes (Fresh/Dried)	China	0.008233	0.005745
071430	Yams (Dioscorea spp.)	Japan	0.001417	0.000722
071440	taro (Colocasia spp.)	China	0.194507	0.102738
071490	Arrowroot, Salep, etc. & Sago Pith (Fresh/Dried)	Hong Kong	0.000320	0.000197
⋮	⋮	⋮	⋮	⋮

The dataset consists of 39 different destination countries, including "Australia", "Brunei Darussalam", "Czech Republic", "East Timor", "France", "Germany", "Hong Kong", "Iran", "Israel", "Japan", "Korea", "Kuwait", "Lebanon", "Malaysia", "Maldives", "Myanmar", "Netherlands", "New Zealand", "Palau", "Papua New Guinea", "Puerto Rico", "Qatar", "Saudi Arabia", "Singapore", "Taiwan", "Thailand", "United Arab Emirates", "United Kingdom", "United States", "Viet Nam", "Bahrain", "China", "Christmas Islands", "India", "Serbia", "Canada", "Denmark", "Belgium", "Cambodia".

B. AIC and BIC Result

The AIC and BIC are utilized to identify the optimal number of clusters. These metrics are derived from formulas that based on the model's likelihood. Fig. 2 illustrate the AIC and BIC values for different numbers of clusters in GMM.



(a) (b)

Fig. 2. Plot of (a) AIC and (b) BIC Result

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 113-122, February 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.55

The plots show a sharp decrease in AIC and BIC at two clusters, followed by a plateau, indicating that two clusters are optimal. The negative AIC and BIC values result from the log-likelihood, which tends to be negative in probabilistic models like GMM. The focus, however, is on the relative differences between these values rather than the absolute values.

C. GMM Results

The most optimal parameters are presented in table 3, which summarizes the best-fitting values for GMM.

Table 3. Optimal parameters for clustering with GMM Model

k	0	1
μ	$[1.77\ x\ 10^{-3}, 1.08\ x\ 10^{-3}]$	$[1.01\ x\ 10^{-1}, 9.31\ x\ 10^{-2}]$
Σ	$\begin{bmatrix} 1.10\ x\ 10^{-5} & 5.63\ x\ 10^{-6} \\ 5.63\ x\ 10^{-6} & 4.93\ x\ 10^{-6} \end{bmatrix}$	$\begin{bmatrix} 1.01\ x\ 10^{-2} & 4.55\ x\ 10^{-3} \\ 4.55\ x\ 10^{-3} & 9.31\ x\ 10^{-3} \end{bmatrix}$
w	0.8152	0.1847

071440 Taro	Australia, Canada, Denmark, …	China, Malaysia, Netherland, Thailand
071490 Sago	Belgium, Cambodia, France	Taiwan, Thailand, Vietnam

Fig. 3 shows the visualization of the clustering results with GMM method with each icon shape and colour drawing a different cluster. Based on the clustering analysis, Thailand emerges as the country with the highest export potential for tuber products. This is evidence by Thailand has four out of five commodities fall into Cluster 1, which represents markets with high potential. Following Thailand, other high-potential markets include Taiwan, Malaysia, and Vietnam. However, it is also important to note that the overall export market for tubers is heavily concentrated within the Asian region.

D. GMM + PSO Results

Fig. 3. Clustering of Tuber Export with GMM Method

Table 4 shows the clustering results using the GMM method. The clustering process resulted in two distinct clusters: Cluster 1, which represents high market opportunities, and Cluster 0, which represents low market opportunities. Cluster 1 consists of 16 data points that are associated with higher trade values or larger quantities. Meanwhile, Cluster 0 includes 72 data points that reflect relatively weaker market opportunities, characterized by lower trade values and quantities.

Table 4. Clustering results with GMM Method

k	0	1
071410 Cassava	Australia, France, Germany, …	Malaysia, Netherland, Taiwan, Thailand, US, Vietnam
071420 Sweet Potato	Australia, Bahrain, China, India, ...	Hong Kong, Korea, Thailand
071430 yams	HongKong, Japan, Malaysia, ...	-

The most optimal parameters are presented in Table 5, which summarizes the best-fitting values for GMM+PSO. The weight (w) and mean (μ) parameters do not differ significantly from those of the GMM model. However, the covariance matrix (Σ) parameter is the key distinction. The covariance matrix represents the spread or variability within each cluster, and in the GMM+PSO model, this parameter is optimized, which can lead to more accurate clustering by accounting for the different shapes and orientations of the clusters.

Table 5. Optimal parameters for clustering with GMM + PSO

k	0	1
μ	$[1.77\ x\ 10^{-3}, 1.08\ x\ 10^{-3}]$	$[1.01\ x\ 10^{-1}, 9.31\ x\ 10^{-2}]$
Σ	$\begin{bmatrix} 1.19\ x\ 10^{-5} & 5.63\ x\ 10^{-6} \\ 5.63\ x\ 10^{-6} & 4.93\ x\ 10^{-6} \end{bmatrix}$	$\begin{bmatrix} 1.01\ x\ 10^{-1} & 4.55\ x\ 10^{-3} \\ 4.55\ x\ 10^{-3} & 9.31\ x\ 10^{-3} \end{bmatrix}$
w	0.8153	0.1847

Table 6 shows the clustering results using the GMM+PSO method. The clustering process resulted in two distinct clusters: Cluster 1, which represents high market

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

opportunities, and Cluster 0, which represents low market opportunities. Cluster 1 consists of 6 data points that are associated with higher trade values or larger quantities. Meanwhile, Cluster 0 includes 82 data points that reflect relatively weaker market opportunities, characterized by lower trade values and quantities.

Table 6. Clustering results with GMM + PSO Method

k	0	1
071410 Cassava	Australia, France, Germany, …	United states
071420 Sweet Potato	Australia, Bahrain, China, India, ...	Hong Kong
071430 yams	Hong Kong, Japan, Malaysia, ...	-
071440 Taro	Australia, Canada, Denmark, …	China, Malaysia
071490 Sago	Belgium, Cambodia, France	Thailand, Vietnam

Fig. 4 shows the visualization of the clustering results with GMM-PSO method with each icon shape and colour drawing a different cluster.

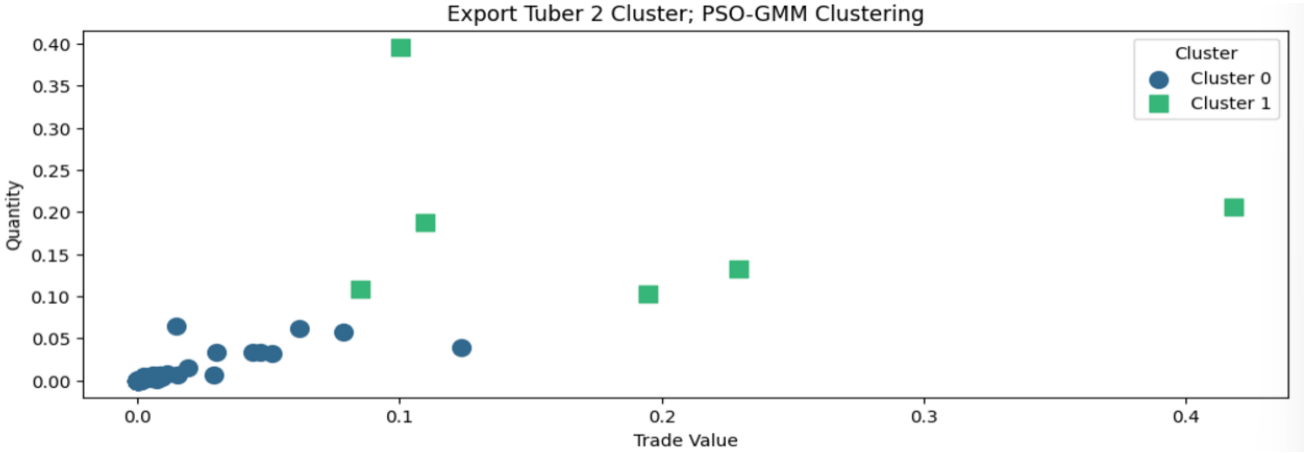


Fig. 4. Clustering of Tuber Export with GMM + PSO Method

The results of the PSO-GMM clustering analysis show that while the countries with potential remain similar, none stand out significantly, and the focus remains largely within the Asian region. Countries such as Hong Kong, China, Malaysia, Thailand, and Vietnam are identified as having market potential, each with their respective key commodities. The difference in the findings is that only the United States is classified as having high market potential for cassava, unlike the results from using only GMM.

4. DISCUSSION

A. Comparison of GMM and GMM + PSO

Fig. 5 shows the comparison of silhouette scores between the clustering models using GMM and GMM+PSO.

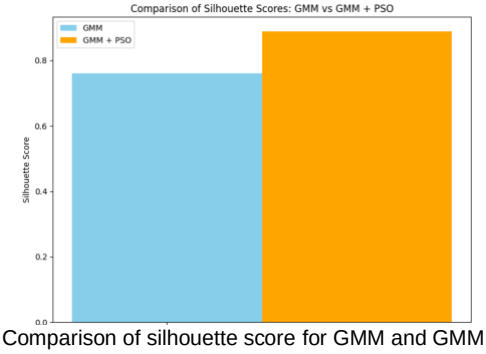


Fig. 5. Comparison of silhouette score for GMM and GMM + PSO

With the same number of clusters, specifically 2, the silhouette score for GMM was 0.7602, while the silhouette score for GMM+PSO was 0.8884. Both the GMM and GMM+PSO algorithms achieve strong structures, as indicated by their high silhouette scores, meaning that the clusters formed are both cohesive and well-separated. However, the higher silhouette score of 0.8884 for GMM+PSO suggests an even stronger structure compared to GMM, indicating that the data points in each cluster are more tightly grouped and better distinguished from other clusters. Fig. 6 shows the comparison of DBI between the clustering models using GMM and GMM+PSO.

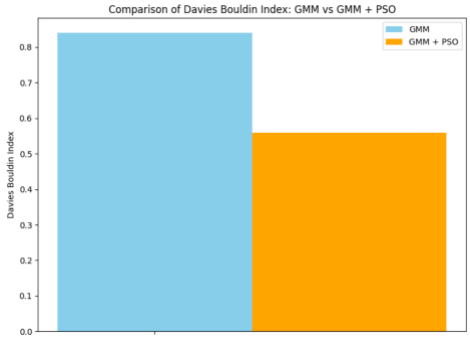


Fig. 6. Comparison of DBI score for GMM and GMM + PSO

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

With the same number of clusters 2, the DBI for GMM was 0.8398, while GMM+PSO achieved a lower DBI of 0.5584. Since a lower DBI indicates better clustering, these results suggest that both GMM and GMM+PSO produce well-structured clusters. However, the significantly lower DBI for GMM+PSO (0.5584) demonstrates even better clustering quality, with more compact and distinct clusters compared to GMM alone. Based on the evaluation scores, PSO is an effective optimization algorithm for enhancing the performance of GMM. The clustering results also show significant differences, particularly for the countries labeled as cluster 1, or high-value markets. Table 7 illustrates the clustering outcomes for both algorithms.

Table 7. Comparison Result for cluster 1

Commodity	GMM	GMM+PSO
071410 Cassava	Malaysia, Netherland, Taiwan, Thailand, US, Vietnam	United States
071420 Sweet Potato	Hong Kong, South Korea, Thailand	Hong Kong
071430 yams	-	-
071440 Taro	China, Malaysia, Netherland, Thailand	China, Malaysia
071490 Sago	Taiwan, Thailand, Vietnam	Thailand, Vietnam

In clustering using the GMM algorithm, many markets are categorized into cluster 1, or high-value markets, even though their trade value and quantity are not particularly high. This suggests that GMM tends to be more lenient in classifying countries into the high-value market cluster. In contrast, when using the GMM+PSO algorithm, only six countries are classified into cluster 1, or high-value markets. From the visualization of the clustering results, GMM+PSO provides better grouping, as it only includes countries with genuinely high trade value and quantity in the high-value market cluster. This indicates that the combination of GMM with PSO improves the quality of the clustering by better identifying key factors that determine market value.

Countries identified as high-potential markets are often influenced by various factors, with one of the primary drivers being consumption trends or demand within each country. According to a report by the FAO [34], tuber crops rank second in terms of the largest cultivated land area across the Asia-Pacific region, following cereals. The demand for tuber crops has grown since 2021, surpassing the growth rate for cereals, which indicates a significant demand for tuber crops in the Asia-Pacific region. Additionally, another contributing factor is the need to meet food security requirements for a growing population, further increasing the demand for these crops.

B. Model Evaluation and Validation

Table 8 presents a comparative analysis of Silhouette Score, between the proposed model and previous research. Both studies aim to improve export clustering methodologies; however, previous research focuses on different commodities, and clustering using similar commodities has yet to be explored. The proposed model, which uses a combination of GMM and PSO is applied to tubers crops. In contrast, previous research explored different clustering algorithms, such as K-Medians, K-Medoids, and K-Means, applied to various commodities. The table compares the clustering quality for each method and commodity across different K values.

Table 8. Comparison of silhouette scores with previous research

Model	Commodity	Method	k	Avg./Scr.
Proposed	Root Crops/Tubers	GMM-PSO	2	0.8884
		GMM	2	0.7602
[13]	HS2: 15,26,27,38,40,44,48, 61,62,64,71,72,84,85,87	K-Medians	3/4/5	0.6430
[10]	frozen shrimp	K-Medoids	2	0.7400
[9]	crumb rubber	K-Means	2	0.7300

The proposed model, utilizing GMM and GMM-PSO, shows superior performance with Silhouette Scores of 0.7602 and 0.8805, respectively. This outperforms previous studies, which had lower silhouette score values. The combination of GMM with the PSO optimization algorithm significantly improves the model's performance across both metrics, demonstrating that PSO is a suitable optimization method when combined with GMM. Table 9 presents a comparative analysis of model evaluation using Silhouette Score and DBI across different clustering algorithms on the same dataset.

Table 9. Comparison of evaluation scores with other algorithms

Algorithm	k	Silhouette Score	DBI
GMM	2	0.7602	0.8398
GMM + PSO	2	0.8884	0.5584
K-means	2	0.8588	0.6280
Hierarichal	2	0.8588	0.6280
MeanShift	10	0.7809	0.2212

The comparative analysis of the clustering algorithms, as shown in the table, indicates that the combination of GMM and PSO achieved the best

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

performance. GMM+PSO achieved the highest Silhouette Score of 0.8884 and the DBI of 0.5584, indicating well-separated and cohesive clusters. In comparison, standard GMM had a Silhouette Score of 0.7602 and a DBI of 0.8398, showing significant improvement when enhanced with PSO. K-means and Hierarchical clustering both performed similarly, with a strong Silhouette Score of 0.8588 and a lower DBI of 0.6080. Meanwhile, MeanShift produced a moderate Silhouette Score of 0.7809 but achieved a highly favorable DBI of 0.2212, However, MeanShift created 10 clusters, which makes the results harder to interpret and less practical. This results show GMM+PSO is the optimal choice for cluster performance based on these metrics. Table 10 presents a comparative analysis of model evaluation using Silhouette Score for different number of clusters which are the most influential parameters.

Table 10. Comparison of evaluation scores with other number of clusters

k	GMM		GMM + PSO	
	Silhouette	DBI	Silhouette	DBI
2	0.7602	0.8398	0.8884	0.5584
3	0.7369	1.0722	-	-
4	0.7795	0.5708	-	-
5	0.7816	0.4045	-	-

The table shows a comparison of silhouette scores between GMM and GMM+PSO for different numbers of clusters. For 2 clusters, GMM+PSO achieves the highest silhouette score of 0.8884, while GMM alone scores 0.7602. However, for 3, 4, and 5 clusters, no PSO results are defined. This is because PSO stops optimization at the optimal number of clusters, which is 2 in this case. The change in the number of clusters can significantly impact both the clustering results and the evaluation scores. Adding more clusters does not always improve the model, and in this research, the optimal solution was found with 2 clusters based on AIC/BIC.

5. CONCLUSION

Clustering analysis plays a vital role in identifying potential markets for commodities by grouping countries based on similar trade patterns. After preprocessing, the dataset was reduced to 88 rows and 5 columns: "HS6 ID", "HS6", "Country", "Trade Value", and "Quantity". The results of the GMM and GMM-PSO models demonstrated significant improvements in silhouette scores, with the GMM achieving 0.7602 and DBI of 0.8398. Meanwhile, GMM-PSO achieving 0.8884 and DBI of 0.5584. These results highlight that PSO, as an optimization technique, successfully enhances the GMM model's performance by optimizing the parameter estimation process. Furthermore, the analysis revealed that most high-potential markets for tuber commodities are in the Asia region, such as Hong Kong, China, Malaysia, Thailand, and Vietnam. However, the United States is also identified as a high-potential market. In contrast, among the five types of tubers, yams (code 071430) were found to be less in demand, as no country was assigned to cluster 1, indicating low market potential for this particular commodity. The clustering analysis in this study is subject to several limitations that could impact the results. One key limitation is the influence of external factors such as changes in demand over time and varying regulations from different governments, which are not fully discussed in the research. These external factors may cause shifts in export patterns, potentially leading to biased interpretations of market behavior. Future research could address these limitations by incorporating time-series data and external variables, such as economic policies or shifts in global demand, to assess their impact on export markets more comprehensively.

REFERENCES

[1] D. H. S. Keefe, H. Jang, and J. M. Sur, “Digitalization for agricultural supply chains resilience: Perspectives from Indonesia as an ASEAN member,” *Asian Journal of Shipping and Logistics*, Dec. 2024, doi: 10.1016/j.ajsl.2024.09.001.

[2] H. Habib Witjaksono, A. Agustina Annisaa Sholihah, F. Bagus Kurniawan, L. Selviani Setyo Ningrum, N. Haliza Asta Palupi, and B. N. Asiyah, “The Role of Exchange Rates, Foreign Exchange Policies, and Foreign Exchange Reserves on the Stability of the Islamic Economy in a Country,” in *Proceedings of Islamic Economics, Business, and Philanthropy*, 2024. doi: 10.1088/1755-1315/1323/1/012013.

[3] Suhartini, B. Waluyo, D. W. Irawanto, B. Nofal, D. S. Lasitya, and B. N. Jihad, “The Role of Root and Tuber Crops on Food Diversification Facing the Climate Change in East Java Indonesia,” in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics, 2024. doi: 10.1088/1755-1315/1323/1/012013.

[4] Directorate of Distribution Statistics, “INDONESIA FOREIGN TRADE STATISTICS,” Jakarta, 2023.

[5] G. E. Charles and M. H. Rangen Jaya, “Credit For Export Bow Penetration: Catalyzing A Breakthrough In The International Trade Sector,” *Journal of Sustainable Economics*, vol. 1, no. 1, pp. 24–28, May 2023, doi: 10.32734/jse.v1i1.12066.

[6] A. Sulaiman, M. S. S. Ali, and A. Ahmad, “Encouraging comparative advantages of export-oriented Indonesian agriculture products,” in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Oct. 2020. doi: 10.1088/1755-1315/575/1/012073.

[7] S. Solikhun, V. Yasin, and Donni Nasution, “Optimization of the Number of Clusters of the K-Means Method in Grouping Egg Production Data in Indonesia,” *International Journal of Artificial Intelligence & Robotics (IJAIR)*, vol. 4, no. 1, pp. 39–47, Jun. 2022, doi: 10.25139/ijair.v4i1.4328.

[8] E. Annisa Octaria, D. Maulani, and M. Daris Syafiq, “Clustering Crumb Rubber Exports Based on Destination Countries Using the K-Means Method,” *Journal of International Trade*, vol. 2, no. 2, pp. 46–51, 2023, doi: 10.32832/jit.

[9] E. P. A. Akhmad and B. Priyono, “Classification of Indonesian Frozen Shrimp Export Data Using K-Medoids Clustering,” *Technology, and Business (JETBIS)*, vol. 3, no. 5, 2024, doi: https://doi.org/10.57185/jetbis.v3i5.106.

[10] H. A. Ulvi and M. Ikhsan, “Comparison of K-Means and K-Medoids Clustering Algorithms for Export and Import Grouping of Goods in Indonesia,” *Sinkron*, vol. 8, no. 3, pp. 1671–1685, Jul. 2024, doi: 10.33395/sinkron.v8i3.13815.

[11] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, “Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country,” in *IOP Conference Series: Materials*

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

Science and Engineering, Institute of Physics Publishing, May 2020. doi: 10.1088/1757-899X/835/1/012058.
[12] R. L. Islami and P. R. Sihombing, "Application Biplot and K-Medians Clustering to Group Export Destination Countries of Indonesia's Product," *Advance Sustainable Science, Engineering and Technology*, vol. 3, no. 1, Apr. 2021, doi: 10.26877/asset.v3i1.8451.
[13] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," Aug. 01, 2020, *MDPI AG*. doi: 10.3390/electronics9081295.
[14] H. A. Santoso and S. C. Haw, "Improvement of k-Means Clustering Performance on Disease Clustering using Gaussian Mixture Model," *Journal of System and Management Sciences*, vol. 13, no. 5, pp. 169–179, 2023, doi: 10.33168/JSMS.2023.0511.
[15] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 158–167. doi: 10.1016/j.procs.2020.04.017.
[16] S. Eh Chonga, M. Md. Siraj, N. A. Rahmat, and M. M. Din, "Integration of PSO and Clustering algorithms for privacy preserving data mining," *International Journal Artificial Intelligent and Informatics*, vol. 2, no. 2, pp. 108–116, Apr. 2022, doi: 10.33292/ijarlit.v2i2.42.
[17] M. H. Raditya, Indwiarti, and Aniq Atiqi Rohmawati, "House Prices Segmentation Using Gaussian Mixture Model-Based Clustering," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 866–871, Nov. 2022, doi: 10.29207/resti.v6i5.4459.
[18] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi: 10.3390/analytics2040042.
[19] Z. Wahidah and D. T. Utari, "Comparison of K-Means and Gaussian Mixture Model in Profiling Areas by Poverty Indicators," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 2, pp. 0717–0726, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0717-0726.
[20] D. Y. Faidah, A. M. Hudzaifa, and R. S. Pontoh, "Clustering of Childhood Diarrhea Diseases using Gaussian Mixture Model," *Communications in Mathematical Biology and Neuroscience*, 2024, doi: 10.28919/cmbn/8365.
[21] A. G. Gad, "Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review," *Archives of Computational Methods in Engineering*, vol. 29, no. 5, pp. 2531–2561, Aug. 2022, doi: 10.1007/s11831-021-09694-4.
[22] M. Jain, V. Saihjpal, N. Singh, and S. B. Singh, "An Overview of Variants and Advancements of PSO Algorithm," *Applied Sciences (Switzerland)*, vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178392.
[23] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International journal of online and biomedical engineering*, vol. 19, no. 4, pp. 174–182, 2023, doi: 10.3991/ijoe.v19i04.37059.
[24] M. Yohansa, K. A. Notodiputro, and E. Erfiani, "Dynamic Time Warping Techniques for Time Series Clustering of Covid-19 Cases in DKI Jakarta," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 13, no. 2, pp. 63–73, Nov. 2022, doi: 10.21512/comtech.v13i2.7413.
[25] S. Renaldi. S, D. A. Prasetya, and A. Muhaimin, "Analisis Klaster Partitioning Around Medoids dengan Gower Distance untuk Rekomendasi Indekos (Studi Kasus: Indekos di Sekitar Kampus UPNVJT)," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 3, pp. 2060–2069, Jul. 2024, doi: 10.33379/gtech.v8i3.4898.
[26] S. Myagmarsuren, "Exploring the Use of Silhouette Score in K-Means Clustering for Image Segmentation (Exploring the Use of Silhouette Score in K-Means Clustering for Image Segmentation)," *International Journal of Engineering Research & Technology (IJERT)*, vol. 13, no. 4, 2024, [Online]. Available: http://www.ijert.org
[27] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," in *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 306–310. doi: 10.1109/ICCMC48092.2020.ICCMC-00057.
[28] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi, and I. W. A. S. Darma, "The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 13–24, 2024, doi: 10.20473/jisebi.10.1.13-24.
[29] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front Energy Res*, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.
[30] W. M. Hameed and N. A. Ali, "Comparison of Seventeen Missing Value Imputation Techniques," *Journal of Hunan University Natural Sciences*, vol. 49, no. 7, pp. 26–36, Jul. 2022, doi: 10.55463/issn.1674-2974.49.7.4.
[31] V. Sharma, "A Study on Data Scaling Methods for Machine Learning," *International Journal for Global Academic & Scientific Research*, vol. 1, no. 1, Feb. 2022, doi: 10.55938/ijgasr.v1i1.4.
[32] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.
[33] M. N. Shakib, M. Shamim, M. N. H. Shawon, M. K. F. Isha, M. M. A. Hashem, and M. A. S. Kamal, "An Adaptive System for Detecting Driving Abnormality of Individual Drivers Using Gaussian Mixture Model," in *5th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/ICEEICT53905.2021.9667850.
[34] OECD/FAO, "OECD-FAO Agricultural Outlook 2021-2030," OECD, Jul. 2021. doi: 10.1787/19428846-en.

AUTHOR BIOGRAPHY

Dwi Arman Prasetya received his B.E. degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 2004. He received M.E. degree from Universitas Brawijaya, Malang, Indonesia in 2010. He received Dr. Eng. degree from Tokushima University, Tokushima, Japan in 2013. He is an Associate Professor in Data Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia. His current research interest includes robotics, swarm robots, artificial intelligent, virtual reality, and internet of things. He is a member of the Electrical Engineering Education Forum Indonesia (Fortei Indonesia), the Institute of Electrical and Electronics Engineers (IEEE), and Deputy head of the certification department of The Institution of Engineers Indonesia.

Anggraini Puspita Sari received her B.E.and M.E. degrees from Universitas Brawijaya, Malang, Indonesia in 2009 and 2012 respectively. She received Dr. Eng. degree from Tokushima University, Tokushima, Japan in 2021. She is a Assistant Professor in Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia. Her current research interest includes forecasting, wind power, artificial intelligent, microelectronic, and Electrical Engineering. She is a member of the Electrical Engineering Education Forum Indonesia (Fortei Indonesia), a member of the Institute of Electrical and Electronics Engineers (IEEE), and a member of the Institution of Engineers Indonesia.

Mohammad Idhom received his B.S.A. degree from Universitas Brawijaya, Malang, Indonesia in 2007. He received his B.C.S. degree from Informatics Engineering, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia in 2012. He

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.

received M.E. degree from Universitas Atma Jaya Yogyakarta, Yogyakarta, Indonesia in 2015. He received Dr. degree from Universitas Negeri Surabaya, Surabaya, Indonesia in 2023. He is a Assistant Professor in Data Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia. His current research interest includes IT service management, software engineering, network security and audit IT. He is a member of the Institution of Engineers Indonesia.

Angela Lisanthoni is data science undergraduate student at the Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, currently in 4th year. Her research interest focused on data analysis and machine learning.

Corresponding author: Anggraini Puspita Sari, anggraini.puspita.if@upnjatim.ac.id, Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya Gunung Anyar, Surabaya, 60294, Indonesia.