EURASIAN JOURNAL OF MATHEMATICAL
AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 13, Issue 3 (2025) 112 – 127

OPTIMIZING FRUIT CULTIVATION THROUGH AN UNSUPERVISED LEARNING APPROACH WITH DATA SCALING TECHNIQUES FOR STRATEGIC INSIGHTS

Thongnim P.[1], **Sreekajon J.**, **Pukseng T.**

Abstract This study employs K-Means clustering to analyze durian farm yield and area of production in Eastern Thailand, revealing patterns that could guide agricultural improvements. By employing advanced data preprocessing techniques including Min-Max scaling and Z-Score normalization, the research ensures robust, equitable consideration of all variables, enhancing the analytical process. Particularly, Min-Max scaling proved most effective for this dataset, optimizing the influence of each variable in the clustering process. The clustering categorizes farms by yield and land characteristics and uncovers optimal practices for durian cultivation. This approach offers actionable insights that promise to support sustainable farming practices and contribute to economic development in rural area. Determining that five was the optimal number of clusters was critical in identifying the most distinct and relevant patterns within the data. The findings highlight the potential of machine learning in agriculture, advocating for a data-driven strategy to optimize durian farm output and sustainability.

Key words: Clustering, Data Scaling, Data Preprocessing, Agriculture, Durian.

AMS Mathematics Subject Classification: 62H30, 62P99.

DOI: 10.32523/2306-6172-2025-13-3-112-127.

1 Introduction

Durian, widely known as the king of fruits plays a crucial role both culturally and economically in Southeast Asia, especially in Thailand [1]. The region of Eastern Thailand, with its favorable climate and soil conditions, stands out as a key area for durian cultivation, making a substantial contribution to the nation's agricultural exports. However, optimizing the yield and production of durian farms is a complex challenge that requires a detailed understanding of the environmental conditions and cultivation practices involved. Therefore, it demands an in-depth comprehension of various factors including local environmental conditions, advanced cultivation techniques, and precise farm management practices, all of which are essential for optimizing durian production.

Given the complexity involved, traditional methods of farm management and yield optimization are proving increasingly insufficient. This situation calls for a transition to more sophisticated, data-driven strategies. Such strategies are capable of analyzing and addressing the intricate aspects of agricultural production. These modern approaches employ sophisticated tools such as machine learning algorithms, drone imagery and internet of things (IoT) sensors to capture and analyze a vast array of data points across the farm [2]. By harnessing these detailed insights, strategies support the development of clustering models that group similar cultivation zones based on various parameters such as arable land, price and production. This allows for more detailed understanding and management of durian farms.

[1] Corresponding Author.

Additionally, these approaches enhance precision agriculture practices, enabling targeted interventions like specific agricultural application and optimal irrigation schedules tailored to each cluster's unique needs. This customized approach boosts the efficiency of durian production and also promotes sustainability by minimizing waste and reducing environmental impact [3],[4],[5]. Therefore, implementing these data-driven techniques can revolutionize durian farming making it more productive and sustainable and positioning it in line with global agricultural innovation trends.

Therefore, the primary objective of this study is to utilize K-Means clustering to conduct a comparative analysis of durian farm yield and area of production in Eastern Thailand. By categorizing farms based on various attributes related to yield and land characteristics, the study aims to identify optimal cultivation practices and effective land use patterns. To refine the methodology of this comparative analysis, incorporating data preprocessing techniques such as Min-Max scaling and Z-Score normalization is essential. Employing a robust dataset that captures a broad spectrum of variables, the research is designed to ensure a thorough analysis. This data-driven approach endeavors to provide actionable insights that support the sustainable intensification of durian cultivation which in turn promises to bolster economic development and enhance environmental sustainability in the region.

In addition, the application of K-Means clustering in agricultural contexts underscores the substantial potential of advanced analytics in boosting crop management and precision farming by integrating preprocessing steps before clustering, the robustness of the analysis is enhanced, ensuring that conclusions drawn from the study consider all relevant features equitably. This leads to more practical and impactful recommendations for durian farming in Eastern Thailand. The deployment of these machine learning techniques exemplifies their ability to augment agricultural productivity and sets a standard for leveraging data science in optimizing farm outputs. Additionally, the study meticulously outlines the methods used, shares insights from the comparative analysis of the clustering models and explores the wider implications for agricultural innovation and sustainability.

2 Literature Review

This section provides an overview of existing research related to optimizing fruit cultivation particularly durian through data-driven approaches and machine learning and clustering techniques.

2.1 Data-Driven Approaches in Agriculture

The agricultural sector has witnessed a significant transformation with the advent of data-driven approaches leveraging advanced technologies to optimize crop yield and resource management. Big data analytics has emerged as a powerful tool in agricultural decision-making as demonstrated with high accuracy in crop yield prediction using large datasets of weather patterns, soil conditions [6] and historical yield data [7]. The integration of Internet of Things (IoT) devices has further revolutionized precision agriculture showcasing how IoT sensors could monitor soil health [8], crop growth and environmental conditions in real-time resulting in a substantial increase in crop yield and a significant reduction in pesticide. Artificial Intelligence (AI) and Machine Learning (ML) applications have also found their place in agriculture developing an AI-powered system for early detection of plant diseases achieving high accuracy in disease identification [9].

Despite these advancements, challenges persist in implementing data-driven approaches in agriculture [10]. These include data quality issues, the need for specialized skills to interpret results and the initial cost of technology adoption particularly for small-scale farmers. It can predict that the integration of blockchain technology with existing data-driven systems will enhance traceability in the agricultural supply chain potentially revolutionizing food safety and quality control measures [11]. While these data-driven approaches have demonstrated significant potential in optimizing various aspects of agriculture, there remains a need to adapt and apply these technologies to the specific challenges and conditions of tropical fruit farming particularly in the context of durian cultivation in Eastern Thailand which is the focus of this study.

2.2 Machine Learning Applications in Fruit Cultivation

Supervised learning and unsupervised learning represent two fundamental approaches in machine learning, each with distinct characteristics and applications. Supervised learning operates on labeled datasets where the algorithm is trained to map input data to known output values and categories. This method is widely used for prediction and classification tasks such as forecasting crop yields and identifying plant diseases [12]. Common algorithms in supervised learning include linear regression [13], logistic regression, decision trees and neural networks [14]. The primary advantage of supervised learning is its ability to make precise predictions based on historical data but it requires a significant amount of labeled data for training which can be time consuming and expensive to obtain.

In contrast, unsupervised learning works with unlabeled data aiming to discover hidden patterns and structures within the dataset without predefined output variables. This approach is particularly useful for exploratory data analysis clustering similar data points and reducing the dimensionality of complex datasets. Popular unsupervised learning algorithms include K-Means clustering [15], hierarchical clustering [16] and principal component analysis (PCA) [17]. It excels at revealing underlying patterns and relationships in data that might not be immediately apparent. This makes it valuable for tasks such as customer segmentation in agricultural markets or identifying groups of plants with similar growth characteristics. The choice between supervised and unsupervised learning depends on the nature of the available data and the specific objectives of the analysis.

2.3 Clustering Techniques in Agricultural Optimization

Among these advanced techniques, clustering algorithms such as K-Means, Elbow method and Silhouette scores are particularly effective in revealing hidden patterns and similarities within agricultural data [18], [19]. These methods segment datasets into clusters based on common characteristics making it easier to identify specific areas of a durian farm that may benefit from tailored agricultural practices. For instance, by applying K-Means, farms can be divided into zones that exhibit similar product levels or yield requirements [20] allowing for more precise and efficient resource allocation. This enables more targeted intervention strategies that can significantly improve yield optimization and land use efficiency. By leveraging these clustering techniques, durian farmers can implement targeted actions that enhance production and also contribute to sustainable farming practices.

In addition to clustering techniques, the application of normalization methods such as Min-Max scaling and Z-Score normalization plays a crucial role in the preprocessing of agricultural data. Min-Max scaling adjusts the data values to a common scale of 0 to 1, which is essential for maintaining a consistent range across different features [21]. This method is particularly useful when the data encompasses attributes with varying units and scales ensuring that each feature contributes equally to the algorithm's performance. On the other hand, Z-Score normalization standardizes the data based on the mean and standard deviation [22]. This approach is beneficial for algorithms that assume data is normally distributed, enhancing the performance of models that are sensitive to outliers in the dataset. By integrating these scaling techniques, the data is better conditioned for clustering, facilitating more accurate and insightful segmentation of the durian farms based on their production characteristics and needs.

Compared to these, the K-Means clustering method provides a computationally efficient approach to grouping farms based on similar yield and soil characteristics. By clustering farms into distinct categories, K-Means can identify patterns in land use and yield distribution that may not be as easily detected with traditional methods. However, K-Means relies on proper data preprocessing and normalization techniques to ensure accurate clustering [21] and its effectiveness is dependent on selecting an appropriate number of clusters, which may require additional methods like the Elbow or Silhouette score for optimization.

Therefore, various machine learning techniques have been applied to agricultural optimization and the K-Means clustering method offers a unique balance of computational efficiency and interpretability. This approach is particularly suited for categorizing farms based on multiple variables potentially revealing insights that might be overlooked by other methods. This study builds upon this foundation applying K-Means clustering to the specific context of durian cultivation in Eastern Thailand. By incorporating advanced data preprocessing, this study aims to enhance the accuracy and reliability of our clustering results. Moreover, this study addresses the gaps in current research regarding durian farming optimization and provides a framework that can be adapted to other tropical fruit cultivation scenarios.

3 Methodology

3.1 Data Collection

This study utilized a comprehensive dataset collected from various durian farms located in Eastern Thailand. The dataset includes a wide range of variables crucial for understanding durian cultivation and optimizing yield. Data were gathered through a combination of Thai government agencies and agricultural collects from data center for smart farming of Burapha University at Chanthaburi [23]. Before proceeding with the analysis, the dataset underwent comprehensive preprocessing to enhance its quality and usability. This involved cleaning steps, such as removing outliers and normalization to scale numerical variables to a uniform scale ensuring that differences in value ranges did not distort the analysis. Additionally, feature selection was carried out to identify and select variables critical to durian yield and land utilization. This selection process was informed by both expert knowledge in the field and exploratory data analysis aiming to focus the study on the most relevant factors for robust and meaningful clustering outcomes.

3.2 Data Preprocessing

The data is cleaned with redundant columns removed and no missing values are present in the dataset. The process of data cleansing helps in making the dataset ready for analysis and modeling by ensuring accuracy, completeness, consistency, and relevance of the data. It is a critical step before data preprocessing because clean data leads to more reliable outputs and decisions.

In the context of a data processing pipeline, one-hot encoding is typically performed after data cleaning. In the dataset, the province information is categorical. Since K-Means requires numerical input, the categorical province data is transformed into a series of binary (0 or 1) columns, known as one-hot encoding [24]. Each column represents a province, and the presence of a province for a record is indicated by a 1, and absence by a 0. The methodology combines both geographical (location-based) and agricultural (production-based) features. This integrated approach is essential for clustering in this scenario because it ensures that both sets of variables that describe where the cultivation takes place and those that describe the cultivation's outcomes are considered in the analysis. It can be seen that Tab. 1 presents a comprehensive overview of the area production data set from 2012 to 2023. Its scaled using both Min-Max, Z-Score methods and the encoded province values offering a comprehensive view of the data preprocessing pipeline for the area production analysis. The yield data also transform the same scaling processes as the area production data

Moreover, data preprocessing is a critical step in the data analysis, especially in machine learning and statistics where the quality and structure of data can significantly impact the performance of models. Min-Max scaling and Z-Score normalization are two techniques used in data preprocessing to transform data in this study.

3.2.1 Min-Max Scaling

Min-Max scaling transforms the data to fit within 0 to 1 range. This method can exaggerate the distances between clusters if the original data contained outliers, as all data is strictly confined to this range [25]. The formula for Min-Max scaling is $x_{\text{scaled}} = (x - x_{\text{min}})/(x_{\text{max}} - x_{\text{min}})$, where x is the original value, x_{min} is the minimum value of the feature in the dataset, x_{max} is the maximum value of the feature in the dataset and x_{scaled} is the scaled value, which will fall within the range of 0 to 1.

3.2.2 Z-Score Normalization

Z-Score normalization, also known as standard scaling, is a technique where the values for each numerical attribute in the data are centered around the mean and scaled by the standard deviation. The resulting transformed feature has a mean of zero and a standard deviation of one [26]. This technique is especially useful for data involving algorithms that assume a normal distribution of the input features, such as many machine learning algorithms, and it can improve the performance of algorithms sensitive to the variance in the data. The formula for Standard Scaling is $Z = (x - \mu)/\sigma$, where x is the original data point, μ represents the mean of the dataset, σ is the standard deviation of the dataset and Z is the Z-Score, resulting in a distribution with a mean of 0 and a standard deviation of 1. This normalization is particularly useful for algorithms that are sensitive to outliers or that assume normally distributed data.

Both techniques are used to ensure that the scale of the data does not distort the analysis and that the algorithms' performance is not adversely affected by the nature of the data. Standardizing data helps to remove biases caused by different scales and makes it easier to compare different variables on the same terms. Each method has its advantages and suitability depending on the nature of the data and the specific analysis to be performed. Therefore, the differences in clustering method between using Min-Max normalization and Z-Score normalization can be attributed to how these scaling methods affect the distribution and relative distances within the dataset.

3.3 Clustering Analysis

In preparing the dataset for machine learning analysis, one-hot encoding is applied to the province and district columns to transform these categorical variables into multiple binary columns, ensuring no ordinal relationships are implied, as year could be treated differently based on its role as either a categorical and a time series element. Additionally, Min-Max scaling and Z-Score normalization are used on the area of production and yield columns to scale these numerical values to a range between 0 and 1, facilitating equal contribution to model performance and improving algorithm convergence. These preprocessing steps are crucial for effectively training machine learning models, as they standardize feature scales and convert all input variables into a format suitable for analysis.

3.3.1 Steps for Clustering

1. Select numerical features that are relevant for clustering. This typically includes features such as area of production, yield, and possibly the one-hot encoded province columns in geographical segmentation.

2. Given the nature of K Means clustering, it is crucial to scale the data. Given the presence of binary columns that is one hot encoded provinces. Given that the area of production and yield values vary widely and could potentially overshadow the binary indicators using either Min-Max scaling and Z-Score normalization might be more effective for this variables.

3. Apply K-Means clustering on the scaled data to identify patterns and groups based on area, yield and geographical distribution.

3.3.2 K-Means Clustering

The K-Mean algorithm aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster. The formal objective function which the algorithm tries to minimize is given by the within cluster sum of squares (WCSS) [27]

Table 1: An Example of area production data set, scaling values
(Min-Max scaling, Z-Score), and province encoding from 2012 to 2023

Year	Province	Area of Production	Min-Max Scaled	Z-Score	One-Hot Encoding
2012	NAKHON NAYOK	97	0.0012	-0.62	1,0,0,0,0
2012	NAKHON NAYOK	30	0.0004	-0.75	1,0,0,0,0
2012	NAKHON NAYOK	8	0.0001	-0.80	1,0,0,0,0
2012	PRACHIN BURI	1638	0.0212	0.35	0,1,0,0,0
2012	PRACHIN BURI	342	0.0044	-0.45	0,1,0,0,0
2012	PRACHIN BURI	64	0.0008	-0.72	0,1,0,0,0
2012	PRACHIN BURI	15	0.0002	-0.78	0,1,0,0,0
2012	CHON BURI	20	0.0003	-0.77	0,0,1,0,0
2012	CHON BURI	67	0.0009	-0.71	0,0,1,0,0
2012	CHON BURI	41	0.0005	-0.74	0,0,1,0,0
2012	TRAT	4328	0.0561	0.85	0,0,0,1,0
2012	TRAT	12400	0.1605	2.10	0,0,0,1,0
2013	RAYONG	2001	0.0259	0.5	0,0,0,0,1
...	...	...	...	...	...

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|^2, \tag{1}$$

where k is the number of clusters, S_i is the set of observations in the i^{th} cluster, x is a single observation from the dataset and μ_i is the centroid of points in S_i.

Therefore, the K-Mean algorithm follows four steps

1. **Initialization** Select k initial centroids and usually choosing k observations at random from the dataset.

2. **Assignment step** Assign each observation to the cluster with the closest centroid. This is done by calculating the distance between each observation and centroid and then classifying the observation into the cluster associated with the nearest centroid.

3. **Update step** Recalculate the centroids as the mean of all observations in each cluster.

4. **Repeat** The assignment and update steps are repeated until the centroids no longer change significantly indicating convergence of the algorithm.

After that, evaluate the clusters to interpret the characteristics of different groups such as yield, production area and geographic variables.

3.4 The Optimal Number of Clusters

The Elbow method and the Silhouette score are two distinct approaches used in cluster analysis serving a different purpose and offering unique insights into the clustering process. The Elbow method is primarily employed to determine the optimal number of clusters within a dataset. It operates by plotting the WCSS against the number of clusters and identifying the Elbow point where further increases in the number of clusters lead to diminishing returns in terms of reducing the WCSS. This method is particularly useful for algorithms like K-Means which require the number of clusters.

In contrast, the Silhouette score measures the quality of clustering achieved by assessing how similar each data point is to its own cluster compared to other clusters. This metric calculates the average distance between each point and the points in the nearest cluster that it is not a part of, normalized by the maximum of this distance and the average distance to points in the same cluster. A high Silhouette score indicates well-defined clusters that are tightly packed together with a clear separation between different clusters, making it an effective tool for evaluating the cohesion and separation of the resulting clusters, irrespective of their number.

While the Elbow method provides a heuristic for determining a suitable number of clusters by balancing the model's complexity and its ability to explain variance within the data, the Silhouette score offers a more direct assessment of clustering quality, focusing on the separation and cohesion of clusters. Consequently, these two methods can be used complementary in clustering analysis. The Elbow method guides the selection of an optimal cluster count and the Silhouette score evaluates the effectiveness and quality of the clustering, ensuring that the chosen number of clusters results in meaningful and distinct groupings within the data.

3.4.1 The Elbow Method

The Elbow method is a heuristic used in determining the optimal number of clusters for K-Means clustering. It involves plotting the WCSS against the number of clusters and looking for the elbow point where the rate of decrease in WCSS significantly slows down indicating the optimal number of clusters [28]. The WCSS is calculated as shown in Eq. (1). Therefore, WCSS values are plotted against the number of clusters k to visually track the decrease in the sum of squares Then, identify the elbow point on the plot, which is the point where the reduction in WCSS becomes less pronounced, suggesting that increasing k further yields diminishing returns in clustering compactness.

3.4.2 Silhouette Scores

The Silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The Silhouette score is particularly useful for determining the separation distance between the resulting clusters [29]. This score can be used with any clustering algorithm, including K-Means clustering. The Silhouette score for a single data point is given by $s = (b - a)/\max(a, b)$, where a is the mean distance between a sample and all other points in the same class, and b is the mean distance between a sample and all other points in the nearest cluster that the sample is not a part of.

The Silhouette score for the dataset is the average of the Silhouette score for each sample.

3.5 Data Visualization

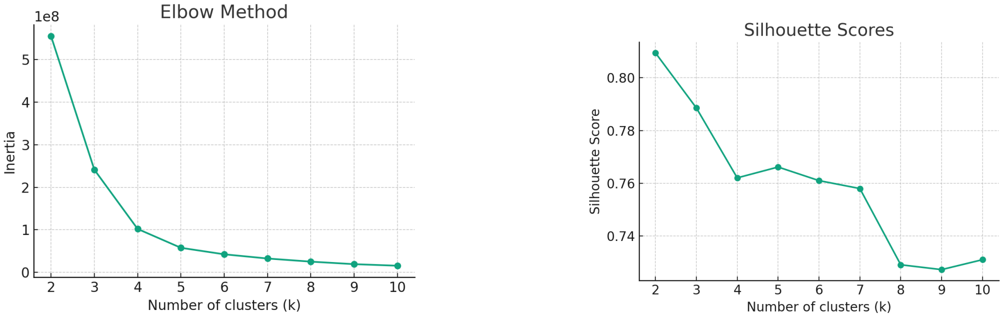
For visualization purposes using principal component analysis (PCA), the key formulas relate primarily to the projection of original data onto the principal components. In addition, PCA is a powerful tool in data science for reducing dimensionality, simplifying data structures, and enabling easier visualization and analysis. The principal components are the eigenvectors of the covariance matrix of the standardized data. These eigenvectors are derived from the eigen decomposition of the covariance matrix Σ. The covariance matrix of the data is computed as $\Sigma = \frac{1}{n-1} Z^T Z$, where Z is the standardized data and Z^T is its transpose. The covariance matrix is then decomposed into its eigenvalues and eigenvectors $\Sigma v = \lambda v$, where v are the eigenvectors and λ are the corresponding eigenvalues. The data is then projected onto the selected eigenvectors, reducing its dimensions but retaining essential features $T = ZV_k$, where V_k includes the top k eigenvectors. The projection of the data onto the first two principal components can be visualized on a plot, with principal component 1 on the x-axis and principal component 2 on the y-axis. This visualization helps in understanding the data's structure and clustering.

4 Results

The results demonstrate the steps involved in analyzing clustering processes within data science and machine learning applications in agriculture, focusing on the yield and area of production of durian in Eastern Thailand. The analysis facilitates clustering based on various locations, enabling a detailed examination of agricultural patterns and efficiencies specific to different regions.

4.1 Choose the Number of Clusters (k)

The Elbow point is where the curve starts to flatten, suggesting a good number of clusters (Fig. 1a). It appears that the curve begins to flatten around $k = 4$ or $k = 5$, indicating these could be suitable choices for Elbow method. However, higher silhouette scores indicate more clearly defined clusters (Fig. 1b). The plot suggests that $k = 2$ and $k = 3$ have relatively



(a) The plot illustrates the Elbow Method used in determining the optimal number of clusters.

(b) The plot illustrates the Silhouette method used in determining the optimal number of clusters.

Figure 1: Elbow and Silhouette to determine optimal clusters.

higher Silhouette scores, with a slight decrease for higher values of k. Therefore, these analyses can guide the decision on the optimal number of clusters. A higher silhouette score at $k = 2$ or $k = 3$ suggests good cluster separation and cohesion at these values, while the elbow method provides additional insight at $k = 4$ and $k = 5$ (Fig. 1).

4.2 Apply K-Means Clustering

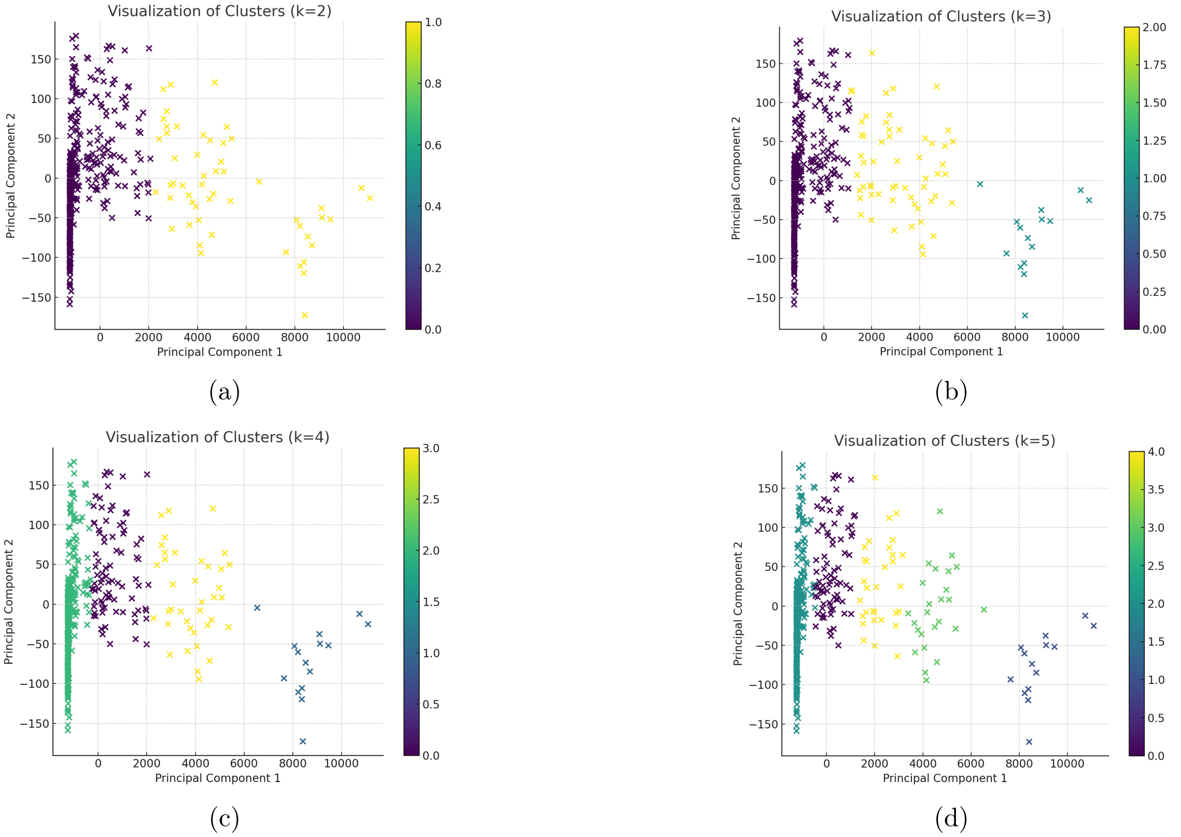
To apply K-Means Clustering using principal component analysis (PCA) for dimensionality reduction, the essential formula involves projecting the original data onto two principal components. These components are derived from the eigenvectors of the covariance matrix of the data, where each component is a linear combination of the original features. The plot has been reduced to two dimensions using PCA, which is common practice for visualizing high-dimensional data. For the clustering process, K-Means was implemented with varying values of k, specifically $k = 2, 3, 4$, and 5, to explore the most effective clustering resolution.

Fig. 2 displays the clustering result of a K-Means algorithm applied to Min-Max normalized data, with the number of clusters set to $k = 2, 3, 4$ and 5. Observations are plotted along the first and second principal components, demonstrating the grouping determined by the algorithm. Colors correspond to different clusters, showcasing the distinct segmentation achieved through the clustering process. This scaling can sometimes exaggerate the distance between clusters if the original data contained outliers.

Fig. 3 depicts a scatter plot of data points that have been clustered using the K-Means algorithm with $k = 2, 3, 4$ and 5, and the data has been normalized using Z-Score normalization. The results suggest that the dataset contains varied but discernible patterns that the K-Means algorithm has been able to group into 2, 3, 4, and 5 distinct clusters in this case. Each cluster may represent different characteristics or behaviors within the fruit cultivation data.

4.3 Evaluate the Results

The Silhouette scores improve as the number of clusters increases, with the highest score being for $k = 5$ (Fig. 4). This suggests that five clusters provide the best cohesion and separation for this dataset when the data is normalized using Min-Max scaling. However, the highest Silhouette score is for $k = 2$, suggesting that two clusters provide the best cohesion and separation for this dataset when the data is normalized using the Z-Score method.



(a)

(b)

(c)

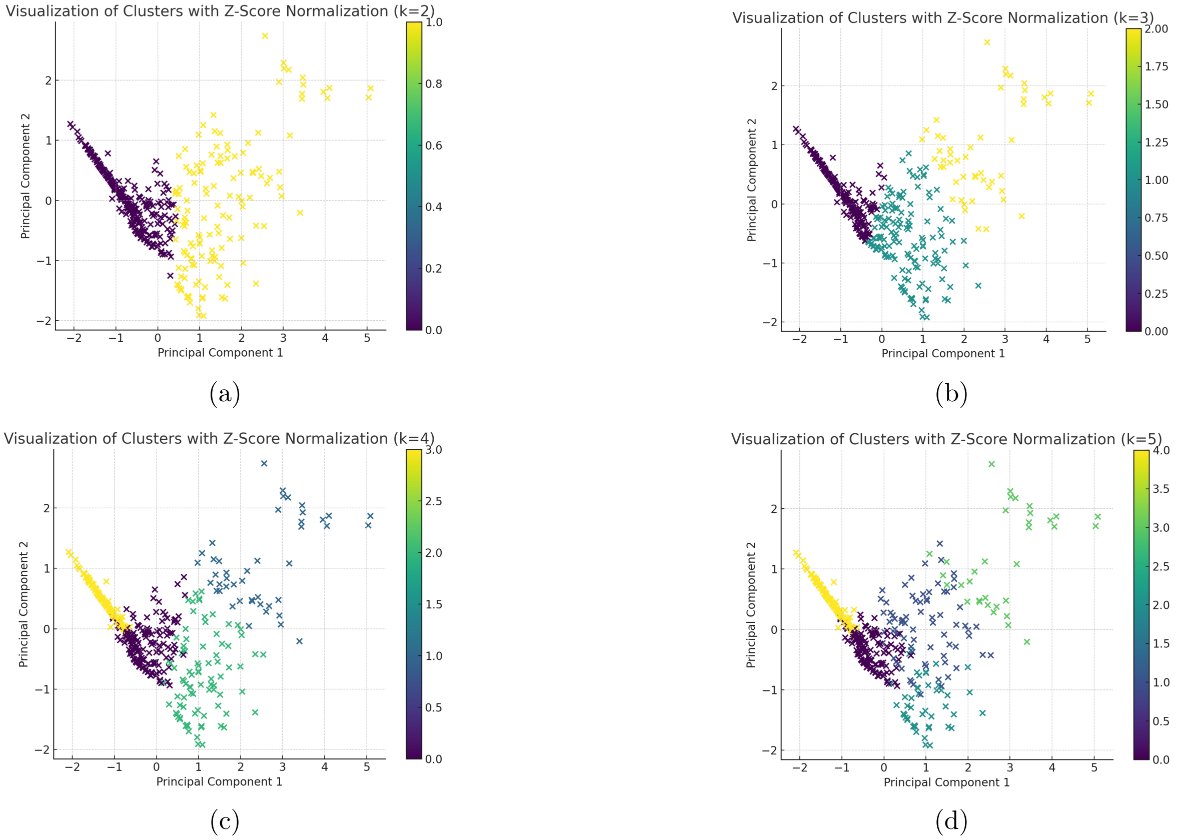
(d)

Figure 2: Scatter plot of K-Means clustering with $k = 2, 3, 4, 5$ from the min–max scaling data.

The differences in Silhouette scores when using Z-Score normalization compared to Min-Max scaling for K-Means clustering, particularly when k=5 is considered the optimal number of clusters, highlight that Min-Max scaling tends to produce better results than Z-Score normalization. This is attributed to how these scaling methods affect the dataset's distribution and the relative distances among data points. Min-Max scaling maintains the original structure of the data better in this context, leading to more distinct and well-separated clusters, thereby enhancing the clustering performance.

Therefore, Min-Max scaling tends to perform better than Z-Score normalization in K-Means clustering primarily [30], [31] because it preserves the original relationships between data points by scaling the data to a fixed range, typically $[0, 1]$. This method maintains the geometric properties of the dataset, which is crucial when clustering, as it ensures that each feature contributes equally to distance calculations. In contrast, Z-Score normalization, which adjusts data based on the mean and standard deviation, can distort these relationships, particularly in datasets that are not symmetrically distributed or contain outliers [32].

Moreover, Min-Max scaling is less sensitive to outliers compared to Z-Score normalization. While both methods are influenced by extreme values, the mean and standard deviation used in Z-Score normalization are more susceptible to being skewed by outliers [33]. This can lead to less effective normalization for the majority of data points. Additionally, the consistent scaling across all features in Min-Max method promotes better-defined clusters and more stable convergence within the K-Means algorithm [34] enhancing the overall clustering performance such as Elbow and Silhouette methods.



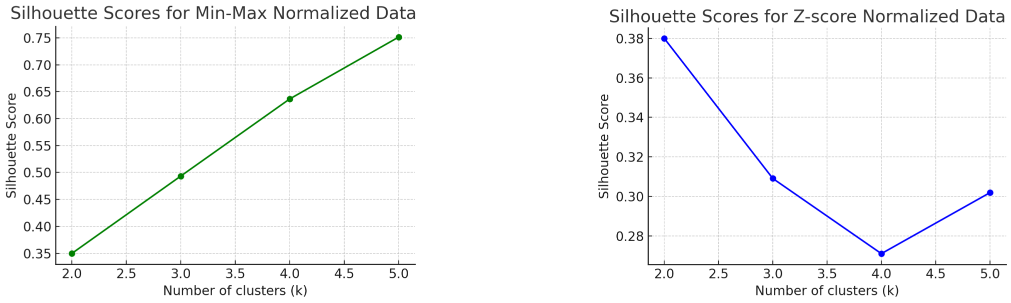
(a)

(b)

(c)

(d)

Figure 3: Scatter plot of K-Means clustering with $k = 2, 3, 4, 5$ from the Z-score normalized data.



(a) The plot illustrates the Silhouette scores used in determining the best solution of Min-Max scaling.

(b) The plot illustrates the Silhouette scores used in determining the best solution of Z-Score normalization.

Figure 4: Elbow and Silhouette to determine optimal clusters.

Clusters overlapping in two-dimensional visualizations after applying Z-Score normalization and dimensionality reduction suggests that the simplicity gained for visualization might come at the cost of losing important structural distinctions between clusters. Therefore, while dimensionality reduction aids in understanding and visualizing data, one must be cautious and consider complementing two-dimensional visual analyses with other methods and metrics to understand the data's true structure in its original dimensionality [35].

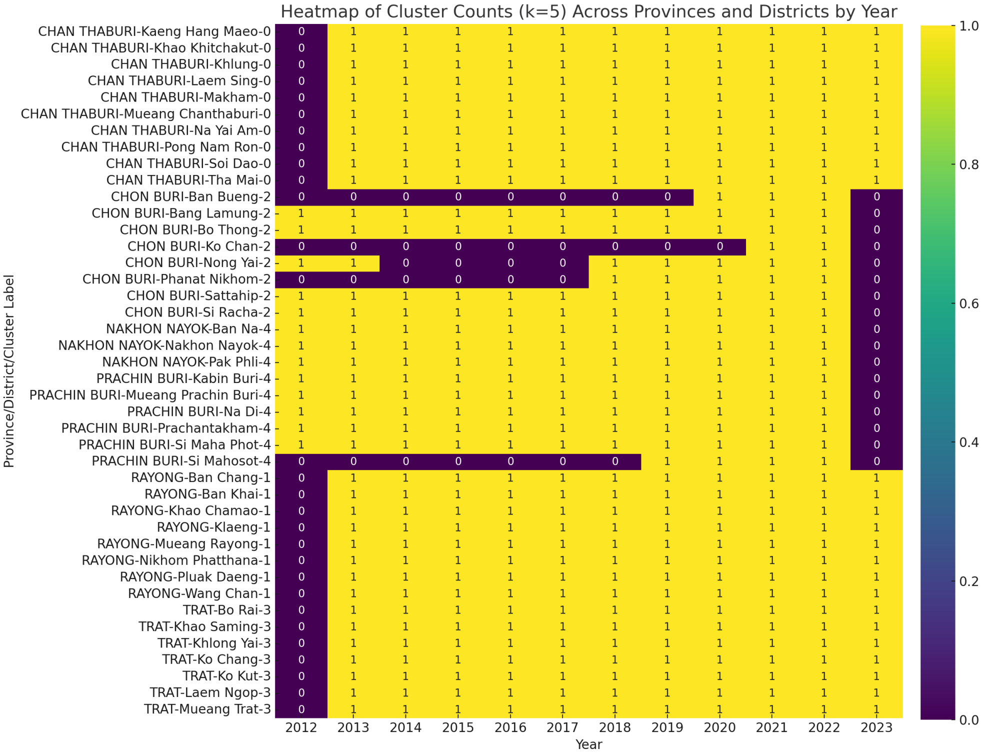


Figure 5: Yearly Heatmap Visualization of Cluster Allocation ($k = 5$) for Districts within Provinces from 2012 to 2023.

4.4 Visualize the Clusters

The x-axis represents the years, while the y-axis lists the province-district combinations (Fig. 5). The color intensity on the heatmap corresponds to the cluster counts, ranging from 0 to 1, as indicated by the color bar on the right. A score of 1 means the province-district combination belongs entirely to a particular cluster in that year, while a score of 0 means it does not belong to that cluster at all.

In addition, the heatmap illustrates a significant degree of temporal consistency in cluster assignments across various province-district combinations from 2012 to 2023. This persistence suggests stable underlying patterns or conditions within those regions that are captured by the clustering algorithm. The vertical streaks of consistent color across the years indicate that the characteristics defining these clusters.

The result represents the $k = 5$ clustering distribution across different geographical and temporal dimensions, structure data to include cluster labels along with the province and district in the row indices. Each row in the data matrix will then correspond to a unique combination of province, district, and cluster Label. The columns should represent different years.

However, the data also reveals instances of sporadic changes in certain districts, like Chonburi-Bang Lamung and Prachinburi-Maha Phot, which experienced a shift in cluster assignment in specific years. These outliers could indicate areas undergoing changes due to a variety of factors. Meanwhile, the uneven distribution of clusters suggests some conditions or characteristics are more prevalent in certain districts. The dominance of particular clusters in specific regions might point to commonalities in the province-district combinations that are grouped together. Over time, the evolution of these clusters could serve as a valuable indicator of how external factors are impacting the provinces and districts, reflecting the dynamic interplay between the local conditions and broader regional or national trends.

5 Discussion

For practical implications and further action, it would need to delve deeper into the specifics of each cluster, correlating them with known factors from the domain of fruit cultivation. Most districts consistently belong to the same cluster over the years, indicating that the characteristics defining each cluster have remained stable over time for these districts. There are instances where districts change clusters from one year to the next, as shown by Chonburi (Bang Lamung), which transitions from cluster 0 in the years 2012-2016 to cluster 1 in 2017, remaining in that cluster through 2023. The majority of the cells are colored to indicate cluster 1, with a very consistent presence across all years. This indicates that many districts share the attributes of this particular cluster.

It is also noticeable that some clusters have a very sparse representation. For example, cluster 4 seems to be less common than others, suggesting that fewer districts share its defining characteristics. In terms of temporal trends, there do not appear to be any significant shifts in clustering over time, as most districts remain in the same cluster. This could mean that whatever features are being used to define these clusters are not experiencing substantial changes year over year. To draw more concrete conclusions, one would need to understand the features used for clustering and the domain context. For instance, if this data pertains to socioeconomic indicators, a consistent cluster membership over time could imply stable economic conditions in those districts.

Economic variables could include a range of data points such as GDP growth, unemployment rates, income levels, industrial output, and investment patterns [36]. During a boom, regions may show improvements in these economic indicators, potentially moving into clusters characterized by higher income levels or industrial growth. Conversely, during a recession, regions might shift into clusters with higher unemployment rates or lower GDP growth. For example, a district that falls into a cluster with high economic activity might suddenly shift to a cluster with lower activity during an economic downturn. This could manifest in the heatmap as a change in color intensity or a shift from one vertical line of consistent color to another, reflecting the changing economic status of that district.

Clusters may also be influenced by environmental changes [37]. If the clustering incorporates environmental variables like average temperatures, rainfall, agricultural output, or pollution levels, shifts in these factors due to climate change could lead to changes in cluster assignments. For instance, an increase in temperature might affect agricultural districts, leading to shifts in clusters due to reduced crop yields. Similarly, districts might shift clusters if they are affected by natural disasters, which could impact a range of indicators from infrastructure damage to population displacement. Changes in land use, such as deforestation or urbanization, could also be reflected in cluster movements [38]. A district that undergoes rapid urbanization might move from a cluster characterized by agricultural land use to one characterized by urban land use, which would be identified by shifts in the associated variables within the clustering algorithm.

In summary, the visualization indicates that the K-Means algorithm has identified five distinct groups within the dataset after min-max scaling. These clusters are based on inherent patterns in the data, possibly reflecting different types of fruit cultivation practices, economics and environmental factors.

6 Conclusion

The K-Means clustering applied to durian farm yield and area of production in Eastern Thailand yields clusters with distinct characteristics. The application of data preprocessing techniques like Min-Max scaling and Z-Score normalization has been pivotal to refining the methodological approach. The analysis showcases the substantial potential of advanced analytics in agricultural settings, particularly highlighting the role of machine learning techniques like K-Means clustering in augmenting agricultural productivity. By ensuring an equitable consideration of all relevant features through robust preprocessing steps, the study leads to practical recommendations for optimizing durian farming outputs in Eastern Thailand. This effort exemplifies the broader implications for agricultural innovation and underlines the importance of leveraging data science for strategic insights in crop management and precision farming. In future research, time-series analysis could be incorporated to forecast trends in durian production and yield. Such analysis would build upon historical clustering patterns and consider external variables, including market demand and the impacts of climate change.

References

[1] Van Hau T., Doan H.T., Nguyen M.T., Huynh K., Tran T.O.Y., Mai V.T., Nguyen V.H., and Tran S.H. Durian. In *Tropical and Subtropical Fruit Crops*, pages 161–200. Apple Academic Press, 2023.

[2] Daud M.M., Abualqumssan A., Rashid F.N., Hanif M., and Saad M. Durian disease classification using transfer learning for disease management system. *Management*, 8(33):67–77, 2023.

[3] Tang R., Aridas N.K., and Talip M.S.A. A durian yield prediction method based on an improved multiple regression model. In *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 140–145. IEEE, 2023.

[4] Gopalakrishnan Y., Al-Gheethi A., Abdul Malek M., Marisa Azlan M., Al-Sahari M., Radin Mohamed R.M.S., Alkhadher S., and Noman E. Removal of basic brown 16 from aqueous solution using durian shell adsorbent, optimisation and techno-economic analysis. *Sustainability*, 12(21):8928, 2020.

[5] Suanpang P., Pothipassa P., Jermsittiparsert K., and Netwong T. Integration of kouprey-inspired optimization algorithms with smart energy nodes for sustainable energy management of agricultural orchards. *Energies*, 15(8):2890, 2022.

[6] Zhang X., Zhang J., Li L., Zhang Y., and Yang G. Monitoring citrus soil moisture and nutrients using an iot based system. *Sensors*, 17(3):447, 2017.

[7] Thongnim P., Sreekajon J., and Pukseng T. Enhancing durian cultivation efficiency through data-driven smart farming using cluster analysis and machine learning. In *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, pages 67–72. IEEE, 2024.

[8] Tudi M., Daniel Ruan H., Wang L., Lyu J., Sadler R., Connell D., Chu C., and Phung D.T. Agriculture development, pesticide application and its impact on the environment. *International journal of environmental research and public health*, 18(3):1112, 2021.

[9] Madeira M., Porfírio R.P., Santos P.A., and Madeira R.N. Ai-powered solution for plant disease detection in viticulture. *Procedia Computer Science*, 238:468–475, 2024.

[10] Shaikh T.A., Rasool T., and Lone F.R. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, 198:107119, 2022.

[11] Rehman K.U., Andleeb S., Ashfaq M., Akram N., and Akram M.W. Blockchain-enabled smart agriculture: Enhancing data-driven decision making and ensuring food security. *Journal of Cleaner Production*, 427:138900, 2023.

[12] Domingues T., Brandão T., and Ferreira J.C. Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. *Agriculture*, 12(9):1350, 2022.

[13] Fashoto S.G., Mbunge E., Ogunleye G., and Van den Burg J. Implementation of machine learning for predicting maize crop yields using multiple linear regression and backward elimination. *Malaysian Journal of Computing (MJoC)*, 6(1):679–697, 2021.

[14] Hara P., Piekutowska M., and Niedbała G. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land*, 10(6):609, 2021.

[15] Golubovic N., Krintz C., Wolski R., Sethuramasamyraja B., and Liu B. A scalable system for executing and scoring k-means clustering techniques and its impact on applications in agriculture. *International Journal of Big Data Intelligence*, 6(3-4):163–175, 2019.

[16] Essary C., Fischer L.M., Irlbeck E., et al. A statistical approach to classification: A guide to hierarchical cluster analysis in agricultural communications research. *Journal of Applied Communications*, 106(3):3, 2022.

[17] Fang W. Assessment of agricultural development level based on hierarchical cluster analysis and principal component analysis: Evidence from china. In *Proceedings of the 2024 9th International Conference on Mathematics and Artificial Intelligence*, pages 44–52, 2024.

[18] Mouret F., Albughdadi M., Duthoit S., Kouamé D., Rieu G., and Tourneret J.-Y. Reconstruction of sentinel-2 derived time series using robust gaussian mixture modelsвЂ"application to the detection of anomalous crop development. *Computers and Electronics in Agriculture*, 198:106983, 2022.

[19] Liu B., Liu C., Zhou Y., Wang D., and Dun Y. An unsupervised chatter detection method based on ae and merging gmm and k-means. *Mechanical Systems and Signal Processing*, 186:109861, 2023.

[20] Salehnia N., Salehnia N., Ansari H., Kolsoumi S., and Bannayan M. Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and k-means approaches. *International journal of biometeorology*, 63:861–872, 2019.

[21] Obaid H.S., Dheyab S.A., and Sabry S.S. The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In *2019 9th annual information technology, electromechanical engineering and microelectronics conference (iemecon)*, pages 279–283. IEEE, 2019.

[22] Setiyawati N., Bangkalang D.H., and Purnomo H.D. Comparison of k-means & k-means++ clustering models using singular value decomposition (svd) in menu engineering. *JOIV: International Journal on Informatics Visualization*, 7(3):871–877, 2023.

[23] Thongnim P., Yuvanatemiya V., and Srinil P. Smart agriculture: Transforming agriculture with technology. In *Asia Simulation Conference*, pages 362–376. Springer, 2023.

[24] Al-Shehari T. and Alsowail R.A. An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, 23(10):1258, 2021.

[25] Sinsomboonthong S. et al. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *International Journal of Mathematics and Mathematical Sciences*, 2022, 2022.

[26] Nkikabahizi C., Cheruiyot W., and Kibe A. Chaining zscore and feature scaling methods to improve neural networks for classification. *Applied Soft Computing*, 123:108908, 2022.

[27] Sinaga K.P. and Yang M.-S. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020.

[28] Shi C., Wei B., Wei S., Wang W., Liu H., and Liu J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, 2021:1–16, 2021.

[29] Thongnim P., Charoenwanit E., and Phukseng T. Cluster quality in agriculture: Assessing gdp and harvest patterns in asia and europe with k-means and silhouette scores. In *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pages 1–5. IEEE, 2023.

[30] Hadouga H. Leadership in agriculture: Artificial intelligence for modelling and forecasting growth in the industry. 2023.

[31] Rani S., Mishra A.K., Kataria A., Mallik S., and Qin H. Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, 13(1):15997, 2023.

[32] Cao X.H., Stojkovic I., and Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC bioinformatics*, 17:1–10, 2016.

[33] Rahmad Ramadhan L. and Anne Mudya Y. A comparative study of z-score and min-max normalization for rainfall classification in pekanbaru. *Journal of Data Science*, 2024(04):1–8, 2024.

[34] Tzortzis G. and Likas A. The minmax k-means clustering algorithm. *Pattern recognition*, 47(7):2505–2516, 2014.

[35] Acevedo A., Duran C., Kuo M.-J., Ciucci S., Schroeder M., and Cannistraci C.V. Measuring group separability in geometrical space for evaluation of pattern recognition and dimension reduction algorithms. *IEEE Access*, 10:22441–22471, 2022.

[36] Zhang Q., Qu Y., and Zhan L. Great transition and new pattern: Agriculture and rural area green development and its coordinated relationship with economic growth in china. *Journal of Environmental Management*, 344:118563, 2023.

[37] Ma L., Zhang Y., Chen S., Yu L., and Zhu Y. Environmental effects and their causes of agricultural production: Evidence from the farming regions of china. *Ecological Indicators*, 144:109549, 2022.

[38] Baig M.F., Mustafa M.R.U., Baig I., Takaijudin H.B., and Zeshan M.T. Assessment of land use land cover changes and future predictions using ca-ann simulation for selangor, malaysia. *Water*, 14(3):402, 2022.

Pattharaporn Thongnim
Statistics, Department of Mathematics,
Faculty of Science,
Burapha University, Chonburi, Thailand
Email: pattharaporn@buu.ac.th

Jakkrapan Sreekajon
Information Technology and Data Science,
Faculty of Science and Arts
Burapha University, Chanthaburi, Thailand
Email: jsreekajon@gmail.com

Thanaphon Phukseng
Data Center, Faculty of Science and Arts,
Burapha University, Chanthaburi, Thailand
Email: thanaph@buu.ac.th.

Received 03.05.2024 , Accepted 30.10.2024, Available online 30.09.2025.