The object of this research is agricultural land in highland areas that have the potential to be planted with garlic. The main problem solved is the difficulty of identifying and selecting optimal land for planting garlic efficiently and objectively, especially in large and geographically complex areas. Special focus is given to data and spatial parameters that affect land welfare. In addition, planting garlic can be a promising business opportunity, especially in areas that have environmental conditions that support its growth. However, garlic production in Indonesia is often unable to meet market demand, resulting in dependence on imports. This is a common problem that can increase the price of garlic on the market. This study aims to increase crop yields and resource utilization efficiency, but also provide adaptive solutions to climate challenges and support national food security by using principal component analysis (PCA) and K-means. Researchers use principal component analysis (PCA) to reduce data dimensions or simplify complex input variables such as altitude, rainfall, temperature, soil type, and others-without losing important information. After that, the K-means clustering algorithm was used to group the areas into several land suitability classes based on the results of the dimension reduction from PCA. The PCA and K-Means methods help in data-based decision making for more efficient agricultural land development. The clustering results can be used by farmers, governments, and agribusiness companies to determine the most suitable locations for planting garlic. The results of the spatial study of garlic cultivation land using PCA and K-means successfully determined spatial land by conducting a classification test with a test accuracy using Inerta of 0.49% and using the Silhouette Score test of 0.89%

Keywords: spatial land, garlic cultivation, machine learning, principal component analysis, k-means

UDC 631.67:528.9:004.021
DOI: 10.15587/1729-4061.2025.325340

OPTIMIZATION OF GARLIC CULTIVATION LAND SELECTION USING PCA AND K-MEANS APPROACH IN SPATIAL INTELLIGENT SYSTEM

Desilia Selvida
Corresponding author
Lecturer*
E-mail: desilia.selvida@usu.ac.id
Annisa Fadhillah Pulungan
Lecturer
Department of Information Technology**
Ade Sarah Huzaifah
Lecturer*
*Department of Computer Science**
**Universitas Sumatera Utara
Dr. T. Mansur str., 9, Medan, Indonesia, 20222

Received 03.04.2025
Received in revised form 23.05.2025
Accepted date 09.06.2025
Published date 30.06.2025

How to Cite: *Selvida, D., Pulungan, A. F., Huzaifah, A. S. (2025). Optimization of garlic cultivation land selection using PCA and K-Means approach in spatial intelligent system. Eastern-European Journal of Enterprise Technologies, 3 (2 (135)), 54–64. https://doi.org/10.15587/1729-4061.2025.325340*

1. Introduction

Garlic (Allium sativum) is one of the high-value horticultural commodities that plays an important role in the lives of Indonesians, both as a basic ingredient in cooking and in the health sector. Along with population growth and increasing national consumption needs [1]. In areas such as Berastagi, determining the right land is crucial to improving productivity and efficiency in cultivation. However, commonly used land selection methods still rely on statistical maps, manual assessments, or general advice that often does not take into account the spatial diversity of land characteristics. Therefore, a computation-based approach is needed that can manage and analyze large-scale spatial data more effectively [2].

Garlic (*Allium sativum L.*) is one of the horticultural commodities that has strategic value in Indonesia due to its role as a staple food and main spice in cooking. Unfortunately, Indonesia's dependence on garlic imports is still very high, reaching around 94% of national demand. This situation poses a major challenge to national food security and leads to price instability due to supply fluctuations and unpredictable weather conditions [3].

Optimizing the spatial selection of garlic cultivation areas is crucial, not only to enhance productivity but also as a foundation for formulating more targeted agricultural policies. This is the key to the success of a sustainable self-sufficiency program, by integrating technical, economic, and social dimensions into the national garlic commodity development strategy [4].

Within this framework, principal component analysis (PCA) and K-means Clustering offer promising approaches for analyzing and grouping regions based on the spatial suitability of land characteristics. PCA is useful for reducing the number of variables determining land suitability, while K-means enables the grouping of regions into specific categories based on their suitability levels. The use of spatial technology and statistical methods allows the identification of potential land to be conducted in a more objective, systematic, and efficient manner [5].

Therefore, efforts to increase garlic productivity are very important. Garlic plants require certain environmental conditions to grow optimally, such as an altitude of 600–1200 meters above sea level, cool temperatures (10–25°C), loose soil texture, neutral to slightly alkaline soil pH, and sufficient but not excessive water availability. With these special requirements, selecting the right land is the key to successful cultivation. However, manual and subjective land mapping is often unable to provide accurate and efficient recommendations [6].

Several studies have been conducted to solve the problem including in garlic cultivation, land characteristics are very important in cultivating land to achieve optimal garlic production. Based on previous research [7], it is possible to state that most garlic cultivation is done in the highlands with an altitude greater than 700 m above sea level. The government is currently implementing a policy to expand garlic cultivation. An important part of this policy is the development of technology to monitor garlic production. To support the government to increase garlic productivity, machine learning technology is needed to determine the spatial location of highland land in cultivation that has an impact on crop production [8].

Therefore, the research topic of developing a spatial intelligence-based land selection system with PCA and K-means approaches is very important to be studied further. This issue is not only practically relevant in the context of increasing national garlic production, but also in line with the development of science and technology in the field of precision agriculture. The application of PCA and K-means methods in spatial data-based systems has the potential to produce a more efficient, objective, and accurate land selection process. Amidst the challenges of climate change and limited productive land, this approach is expected to make a real contribution to efforts to increase land efficiency, agricultural productivity, and strengthen national food security [9–12].

2. Literature review and problem statement

The use of artificial intelligence (AI) technology in agriculture has opened up great opportunities to improve efficiency and accuracy in the management of garlic horticulture. However, various studies show that there are still significant challenges and shortcomings, both from a technical and implementation perspective.

Paper [13] discusses an IoT-based smart irrigation system and machine learning for garlic cultivation in Enrekang Regency, South Sulawesi. This system can measure soil moisture, temperature, and air humidity in real-time, and provide high-accuracy irrigation recommendations using a random forest algorithm (94% accuracy, AUC 0.90). However, the main limitation identified is the lack of integration with pest detection systems, which are important for supporting irrigation decisions that consider pesticide effectiveness.

Paper [14] presents a systematic review of 43 studies on IoT-based smart irrigation systems that incorporate advanced machine learning models. The study concludes that such systems can improve water use efficiency by up to 50%. However, several major challenges were identified, such as suboptimal integration of data from various sources, limited network connectivity in remote agricultural areas, and scalability constraints for widespread implementation.

Paper [15] develops an Arduino-based solar-powered automatic irrigation system for garlic cultivation in lowland areas, achieving water efficiency of 46%. While this system shows promise in terms of energy and water efficiency, significant weaknesses include the lack of real-time interoperability of microclimate data and minimal integration with other agronomic factors, such as pest management and fertilization.

Paper [16] develops a method for recognizing the orientation of garlic cloves based on image contours and deep learning. Although recognition accuracy exceeds 98%, there are still issues related to the model's generalizability under varying field conditions, such as irregular clove shapes, remaining skin, and the complexity of integrating the system with diverse planting machines. Additionally, technical challenges in synchronizing image recognition with mechanical actuation in the field limit widespread application.

Paper [17] develops a land suitability classification system for garlic cultivation using the Spatial ID3 algorithm in the form of a web application. This system can visualize land suitability maps and recommend varieties based on region. However, this model cannot predict regions with conditions not represented in the training data, especially in microecosystems not covered by available spatial data. The complexity and cost of collecting large-scale data also pose challenges for national-level implementation.

Paper [18] proposes a more efficient approach through the application of principal component analysis (PCA) for environmental data dimension reduction, as well as K-means Clustering for grouping regions based on characteristics relevant to garlic cultivation. This approach is considered capable of addressing some of the issues in accurate and computationally efficient spatial classification.

Paper [19] shows that the integration of AI and IoT in modern hydroponic systems can improve the efficiency of nutrient management, irrigation, and disease detection. However, there are still challenges such as high initial costs, technical complexity, and low user skills in operating automated systems in real-time. Therefore, a modular system that is easy for farmers to use is needed.

Papers [20, 21] highlight the development of an IoT sensor-based crop recommendation system and soil nutrient analysis. The weakness lies in the limited technological infrastructure in rural areas and the low digital literacy of farmers. The proposed solution is to simplify the system into a modular form and an offline-first mobile application based on low-cost devices, although this approach does not yet include full integration of IoT and AI.

Papers [22, 23] investigate the relationship between habitat conditions and the content of bioactive compounds in Allium ursinum. The research shows that soil and climate factors significantly influence yield and compound content, but also reveals a mismatch between the optimal location for biomass yield and secondary metabolite content. Additionally, there is no predictive spatial model available that can fully integrate all environmental parameters.

Therefore, the main unresolved issue emerging from the reviewed literature is the absence of a land selection classification model that can be used in garlic cultivation, which is lightweight, accurate, and adaptive, and implemented in real-time.

There are several limitations in the land selection optimization model for garlic cultivation using PCA and K-means approaches in spatial intelligence systems, including: limitations in representing heterogeneous environmental data, the model's inability to capture local agronomic complexity, and the assumption of cluster homogeneity, which is inconsistent with the diversity of agroecosystems in the field. Additionally, this model remains static and does not consider temporal dynamics or integration with real-time data. These issues have prompted the current research to develop a more adaptive and contextual spatial intelligence system approach by integrating dimension reduction and spatial clustering methods into a framework capable of accommodating multi-source data, accounting for local ecological variability, and generating operationally

interpretable outputs to support decision-making in the development of sustainable garlic cultivation.

3. The aim and objectives of the study

The objective of this study is to develop a model for selecting garlic growing areas through a machine learning classification approach using PCA and K-Means. To achieve this aim, the following objectives are accomplished:

– to preprocess spatial data of highland areas as the main spatial variable using principal component analysis (PCA) and K-means;

– to evaluate the contribution of principal component analysis (PCA) and K-means clustering in identifying garlic cultivation areas by measuring the accuracy and stability of clustering results based on spatial data variation;

– to visualize and improve the effectiveness of the spatial soil clustering model by proposing an approach to refine the PCA and K-means parameters, as well as evaluating metrics such as precision, recall, F1-score, and representative spatial visualization.

4. Materials and methods

4. 1. Object and hypothesis of the study

This study focuses on the spatial characteristics of highland areas for garlic cultivation, which are used to develop and evaluate a model for selecting garlic farming areas. The main hypothesis is that the combination approach between principal component analysis (PCA) and K-means clustering in a spatial intelligence system is able to optimize the selection of garlic cultivation land by increasing the accuracy of land suitability based on environmental variables, and producing spatial information that can be used practically to support agronomic decisions in various regions. This study assumes that the characteristics of garlic cultivation land can be effectively represented through a combination of land variables reduced using principal component analysis (PCA), and that areas with similar characteristics can be accurately identified using K-means clustering. In addition, it is assumed that this approach, when implemented in a spatial intelligence system, is able to produce agronomically relevant land classifications and is operationally useful in sustainable garlic cultivation planning.

4. 2. Datasets

The dataset used in this research will be divided into two, namely training data and testing data. For the training process, the data used consists of valid data obtained from soil spatial data in Berastagi. The data used by the author uses garlic cultivation criteria. Table 1 of this study shows data on land characteristics.

The algorithms used in the research are principal component analysis (PCA) and K-means. In this context, the K-means clustering algorithm has been widely recognized as a basic method in spatial segmentation for land selection. However, limitations in handling high-dimensional data and the presence of redundant information often lead to unrepresentative clustering results. Therefore, the principal component analysis (PCA) approach is applied as a preprocessing technique to reduce the dimensionality of the data without significant information loss. Thus, the integration of PCA and K-means in an ensemble method enables the improvement of potential land zone classification accuracy as well as efficiency in processing complex spatial data. The combination of these two methods is realized within the framework of a spatial intelligent system, which serves as a geospatial data-based analytics platform to support decision-making in garlic cultivation. Thus, this research aims to provide a systematic solution to the challenge of cultivation land selection, while offering theoretical contributions in the development of multidimensional analytical methods that can be widely applied in the precision agriculture sector.

Table 1

Soil spatial characteristics data

Characteristics of garlic growing requirements		Cluster ID			
		1	2	3	Outline
1	Temperature	24	24–25	24	24
2	Rainfall	300–350	300–400	300–350	250–350
3	Elevation	Low	Rather low	Low	Low
4	Soil depth	In	In	In	Very deep
5	Soil texture	Fine	Medium	Fine	Fine
6	Relief	Slightly sloping	Steep	Slightly steep	Flat

4. 3. Principal component analysis (PCA)

Principal component analysis (PCA) is a dimension reduction technique used in data analysis and machine learning to transform high-dimensional datasets into lower dimensions while retaining as much variation in the data as possible. PCA works by finding principal components, which are linear combinations of the original variables that have the greatest variance [24, 25].

PCA is used to:

1. Reduce data dimension without losing too much information.

2. Eliminate redundancy in the dataset (multicollinearity).

3. Simplify data visualization in 2D or 3D.

4. Improve the efficiency of machine learning models by reducing the number of features.

As for the principal component analysis (PCA) formula, PCA consists of several main steps:

1. Data standardization. Data must be normalized to have a mean of 0 and a variance of 1 as shown in formula (1)

$$X' = \frac{x - \mu}{\sigma}, \tag{1}$$

with x – original value of the feature;

– μ – average features;

– σ – standard deviation of features.

2. Forming the covariance matrix. The covariance matrix is calculated to understand the relationship between the features. The equation in the formula shown is

$$C = \frac{1}{m-1} X'^{t} X', \tag{2}$$

with X – the standardized data matrix, C – the covariance matrix.

3. Determination of eigenvalues and eigenvectors. Eigenvalues indicate the amount of variance explained by each eigenvector.

4. Selecting principal components. It is possible to sort the eigenvalues from largest to smallest and select the k eigenvectors with the largest eigenvalues to form the k transformation matrix.

5. Data transformation to new dimensions. It is possible to project the data to the new space with. The equation in the formula shown is

$$X_{pca} = X'V_K, \tag{3}$$

with V_k – data in a new dimension;

K – matrix of selected eigenvectors.

With these steps, PCA helps to reduce the dimensionality of the dataset without losing significant information.

4. 4. K-means

K-means is a clustering algorithm used to group data into K clusters based on similarity of features [26]. The algorithm works by dividing a set of data into K different clusters, where each data point will be assigned to the cluster with the closest centroid. K-Means is often used in data exploration, customer segmentation, image clustering, and various other machine learning applications. The algorithm is iterative based and seeks an optimal solution by minimizing the variance within each cluster. The main process of K-means involves several steps, which are described as follows [5, 6]:

1. Determining the number of clusters (K). Determination of the desired number of clusters K.

2. Centroid initialization. Randomly selection of K points as the initial centroid of the dataset.

3. Calculating the distance of each point to the centroid. The distance between each point and each centroid is calculated using the distance formula Euclidean. The equation in the formula shown is

$$d(Xi,Cj) = \sqrt{\sum_{m=1}^{n}\left(X_{im} - C_{jm}\right)^2}, \tag{4}$$

with $d(X_i, C_j)$ – the distance between data point X_i and centroid C_j;

X_{im} – to m feature value of the data point X_i;

C_{jm} – feature value of the centroid C_j; Check this please;

N – number of features in the dataset

4. Clustering data to the closest cluster.

Each data point is assigned to the cluster with the closest centroid based on the distance calculation.

5. Updating the centroid.

Once all points are clustered, a new centroid is calculated as the average of all points in the cluster. The equation in the formula shown is

$$C_j = \frac{1}{N_j}\sum_{i=1}^{N_j} X_i, \tag{5}$$

where C_j – new centroid for the cluster j;

N_j – number of data in the cluster j;

X_i – data points in the cluster.

6. Repeating the process until convergence. This process is repeated until the centroid no longer changes or the change is very small (convergent).

7. Evaluation of clusterization results. After clustering is complete, an evaluation is performed using Silhouette Score or WCSS (Within-Cluster Sum of Squares) to assess the quality of the clusters formed. The equation in the formula shown is:

$$WCSS = \sum_{j=1}^{K}\sum_{i=1}^{N_j}\left(X_i - C_j\right)^2, \tag{6}$$

where $WCSS$ – total squared distance in one cluster;

K – number of clusters;

N_j – number of points in the cluster j;

X_i – points in the cluster;

C_j – centroid Dari cluster j.

K-means is an effective clustering method in grouping data based on similarity. The selection of the optimal number of K clusters is usually done by the Inertia or Silhouette Score method. This algorithm is widely used in data analysis and machine learning to identify patterns in large data.

5. Optimization result of spatial problem of shallot cultivation with machine learning

5. 1. Implementing principal component analysis (PCA) and K-means optimization

This research focuses on the highland locations of Berastagi and Kabanjahe. The researchers collected environmental data based on variables, temperature, rainfall, elevation, soil depth, drainage, soil texture, soil acidity, CEC, KB, and relief. Furthermore, spatial data in the form of latitude and longitude. The collected data will be divided into two sets: one is used as training data, and the other is used as data for testing. The stages of the research methodology are shown in Fig. 1.

Fig. 1 illustrates the complete flow of the research methodological stages, starting from data input to producing land suitability analysis. This process includes several main stages as follows:

– input data: the data used consists of soil spatial data and climate/weather data. This data is obtained from various sources such as satellite imagery, soil mapping, and weather station data;

– training and testing data: input data is then divided into two subsets, namely training data for the model training process and testing data for testing model performance. This division is important to ensure that the model built can be evaluated objectively;

– cleaning and labeling (initial pre-processing): training data undergoes a cleaning process, namely cleaning from noise, missing values, or outliers. The labeling process is used if there is a need for initial annotation in internal validation or semi-supervised learning (although this is an unsupervised process in general);

– preprocessing – PCA (principal component analysis): covariance matrix is calculated to determine the relationship between variables.

Eigenvalue and eigenvector are calculated to find the principal components that represent the largest variation in the data.

PCA aims to reduce the dimensionality of the data, eliminate redundancy, and speed up the clustering process without losing important information;

– K-means clustering: K value is determined (e.g. through the elbow or silhouette method) to determine the optimal number of clusters.

The data is then classified into clusters based on the Euclidean distance between points and the cluster center. This produces groups of areas with similar suitability characteristics;

– Performance evaluation & analysis result: evaluation is carried out on the cluster results with accuracy, precision, and recall.

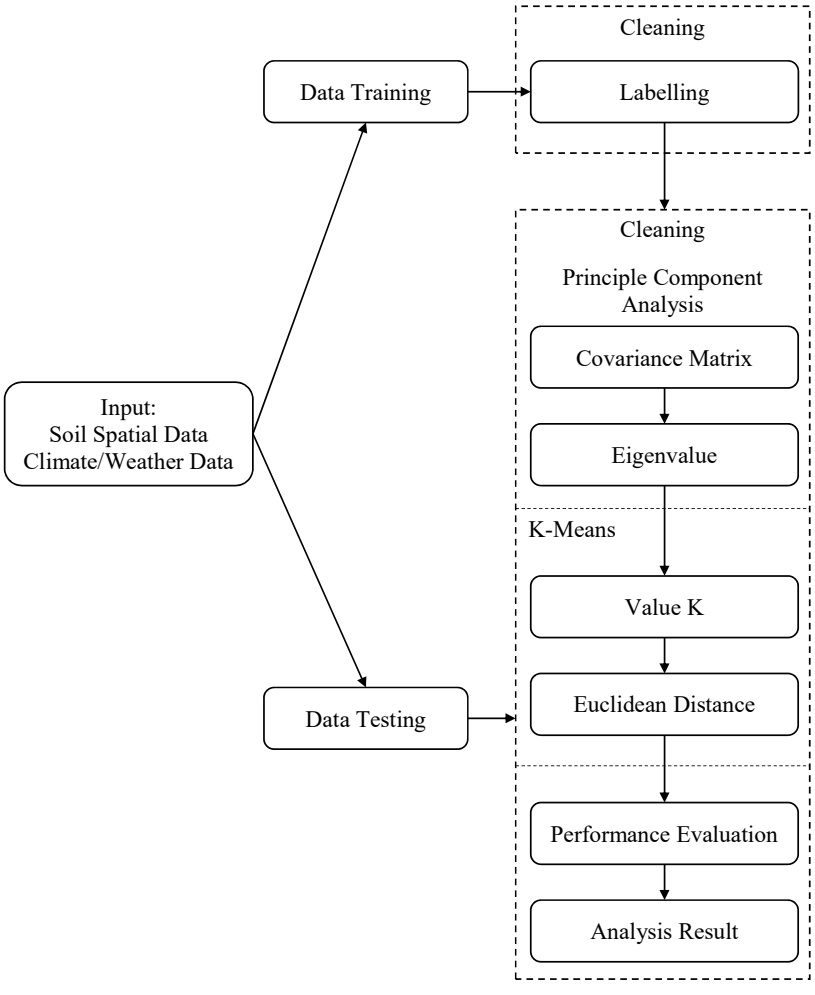

Fig. 1. Research methodology

5. 2. Evaluating the performance of principal component analysis

Spatial classification of garlic cultivation land using Google Colab. Google Colab is a cloud-based platform that allows users to perform data processing and machine learning without having to install additional software. In this research, Google Colab is used to train clustering models to determine groups of garlic cultivation areas based on land characteristics using Principal component analysis (PCA) and K-means clustering. At this stage, data training is the process of training the model to recognize patterns in the dataset, so that it can be used to make decisions or perform classification. In this study, the author conducted testing using Google Coolab as shown in Fig. 2.

The process in Fig. 2 shows the initial stage in building a K-means-based spatial classification system. The initial dataset serves as a base model for forming clusters, while the new dataset represents the target areas to be mapped based on similarities in environmental characteristics. This step is very important for grouping areas into garlic cultivation land suitability zones, and the results will be visualized in a digital map using the Python library visualization.

```
# Data acuan untuk menentukan cluster
data_awal = {
    'Temperatur': [24, 25, 24, 24],
    'Curah_Hujan': [350, 400, 350, 350],
    'Elevasi': [1, 2, 1, 1],
    'Kedalaman_Tanah': [1, 1, 1, 2],
    'Drainase': [1, 1, 1, 1],
    'Tekstur_Tanah': [1, 3, 1, 1],
    'Kemasaman_Tanah': [3, 2, 3, 3],
    'KTK': [3, 1, 3, 3],
    'KB': [5, 4, 5, 5],
    'Relief': [3, 5, 4, 1]
}

df_awal = pd.DataFrame(data_awal)
print(df_awal)
```

```
   Temperatur  Curah_Hujan  Elevasi  Kedalaman_Tanah  Drainase  Tekstur_Tanah  \
0          24          350        1                1         1              1
1          25          400        2                1         1              3
2          24          350        1                1         1              1
3          24          350        1                2         1              1

   Kemasaman_Tanah  KTK  KB  Relief
0                3    3   5       3
1                2    1   4       5
2                3    3   5       4
3                3    3   5       1
```

```
# Data baru yang ingin di-cluster
data_baru = {
    'Karo': [22, 271, 4, 1, 1, 3, 3, 1, 1, 6],
    'Simalungun': [25, 344, 2, 1, 1, 3, 3, 3, 1, 5],
    'Samosir': [29, 232, 2, 1, 1, 4, 2, 3, 5, 2],
    'Tapanuli Utara': [23, 387, 4, 1, 1, 2, 2, 1, 1, 6],
    'Sibolga': [25, 160, 1, 1, 1, 2, 3, 3, 1, 2],
    'Dairi': [21, 345, 1, 1, 1, 2, 2, 3, 3, 4],
    'Humbahas': [22, 228, 4, 1, 2, 3, 2, 5, 3, 2]
}
```

Fig. 2. Google Coolab training process

5. 3. Visualization in modeling the spatial problem of shallot cultivation with machine learning

This study has a design or flow of spatial classification analysis of garlic cultivation with algorithms. principal component analysis (PCA) and K-means. At this stage the researcher starts by collecting initial data then cleaning the data and combining it into one dataset, the data that has been collected will be examined and this stage provides an analytical basis for a study by identifying potential problems in the data. The data in this study were obtained from BMKG datasets and farmers' observation data. Here are the locations and garlic farmers. The author's data collection in Berastagi City to see directly and communicate with farmers is shown in Fig. 3, 4.

Fig. 3. Garlic farmers

Fig. 4. Location of garlic farms

Fig. 3, 4 show a large and well-organized garlic farming area in the Berastagi area. The garlic plants are planted in parallel rows, indicating a well-organized farming system. The land looks fertile, with soil that supports the growth of garlic plants. The land is likely intensively managed by farmers to ensure optimal yields. Land conditions greatly affect garlic cultivation. Then in Table 2 is the data used to determine the cluster.

Table 2 contains the main environmental variables that affect land suitability for garlic cultivation:

– temperature (°C) and rainfall (mm): main climate parameters;

– elevation: indicates the land height (numerical code, possibly in zoning classification);

– soil depth and drainage: soil characteristics that are important for root growth and resistance to waterlogging;

– soil texture: indicates the type of soil (e.g. clay, sand, loam) that greatly affects water and nutrient retention.

Data preprocessing.

In the initial stage, preprocessing is collecting relevant data for garlic cultivation in the highlands. The required data include: temperature, rainfall, elevation, soil depth, soil texture and relief. Once the data is collected, the next step of principal component analysis (PCA) is used to reduce the dimensions of complex data into fewer principal components. The steps are calculation of the covariance matrix of the normalized data to understand the relationship between variables. Calculation of the eigenvalue and eigenvector of the covariance matrix to determine the principal components. Selection of a number of principal components that explain a sufficient percentage of the variance (for example, 80–90% of the total variance). Once the data has been reduced in dimension, the next step is segmentation using K-means clustering. This helps in grouping the cultivation areas based on environmental and agronomic characteristics. Once the clustering is complete, it is possible to analyze the results to understand the patterns and characteristics of each cluster. After this, it is possible to visualize the analysis results using spatial maps to show the distribution of clusters and environmental characteristics. The next stage evaluates the performance using precision, recall, f1-score, inertia and silhouette score [27].

Table 3 is the data used to determine the cluster center point. This data consists of several features (variables) that represent characteristics This table is the basis for the cluster process.

Table 3 is the data from the principal component analysis implementation process. In this process it is possible to reduce the number of features while maintaining the main information. The result is data with fewer features but still represents the characteristics of the garlic cultivation area. Furthermore, the data will be processed to cluster the region with K-Means. Furthermore, Tables 4, 5 are the results of data clustering and mapping based on location.

Table 2

Climate and soil condition dataset for land assessment

Temperature (C)	Rainfall (mm)	Elevation (m)	Soil depth (CM)	Drainage (SCORE)	Soil texture (SCORE)	Soil acidity (PH)	KTK (CMOL/KG)	KB (%)	Relief (SCORE)
24	350	1	1	1	1	0	3	3	5
25	400	2	1	1	1	1	2	1	4
24	350	1	1	1	1	2	3	3	5
24	350	1	1	2	1	3	3	3	5

Table 3

New data to be clustered

	0	1	2	3	4	5	6	7	8	9
Karo	22	271	4	1	1	3	3	1	1	6
Simalungun	25	344	2	1	1	3	3	3	1	5
Samosir	29	232	2	1	1	4	2	3	5	2
Tapanuli Utara	23	387	4	1	1	2	2	1	1	6
Sibolga	25	160	1	1	1	2	3	3	1	2
Dairi	21	345	1	1	1	2	2	3	3	4
Humbahas	22	228	4	1	2	3	2	5	3	2

Table 4 presents input data from 7 districts, followed by the cluster results generated by the K-Means algorithm. The numbered columns (0–9) are representations of environmental and soil variables that have been normalized and used as features, including: temperature, rainfall, elevation, soil depth, drainage, soil texture, soil acidity, CEC, base saturation, relief, etc.

Table 5 simplifies the clustering result information:

– district name: administrative area analyzed;

– cluster: clustering result from K-Means model;

– district number: represents the district index or ID in the database or mapping system. Fig. 5 is North Sumatra Maps based on clustering results. This picture shows the results of the classification of areas in North Sumatra into three clusters:

– cluster 0 (very suitable): marked in blue;

– cluster 1 (suitable): marked in green;

– cluster 2 (quite suitable): marked in orange.

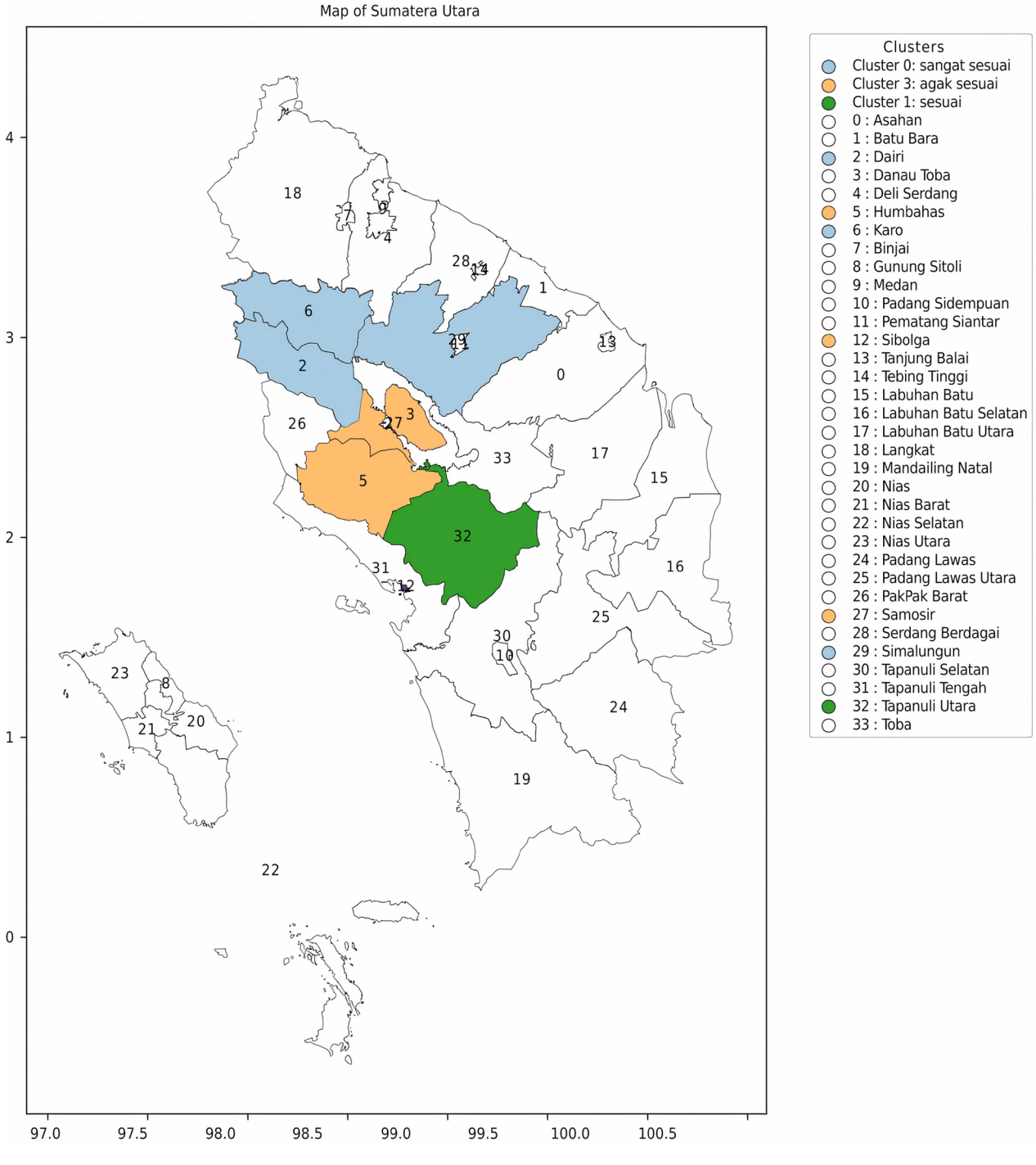

Fig. 5. Regional spatial cluster visualization results

Each region such as North Tapanuli, Central Tapanuli, and Humbang Hasundutan is classified according to its characteristics in the relevant cluster. The geographical location and spatial relationships between regions are also well depicted in this map.

Fig. 6 below is a map of North Sumatra province showing the results of regional clustering based on the level of suitability. In general, this map divides the regions in North Sumatra into three groups (clusters) based on color:

– light blue represents regions that are very suitable for certain criteria;

– purple indicates regions that are somewhat suitable;

– green indicates regions that are suitable.

Table 4

New clustered data

Karo	0	1	2	3	4	5	6	7	8	9	Cluster
	271	4	1	1	3	3	1	1	6	271	0
Simalungun	344	2	1	1	3	3	3	1	5	344	0
Samosir	232	2	1	1	4	2	3	5	2	232	3
Tapanuli Utara	387	4	1	1	2	2	1	1	6	387	1
Sibolga	160	1	1	1	2	3	3	1	2	160	3
Dairi	345	1	1	1	2	2	3	3	4	345	0
Humbahas	228	4	1	2	3	2	5	3	2	228	3

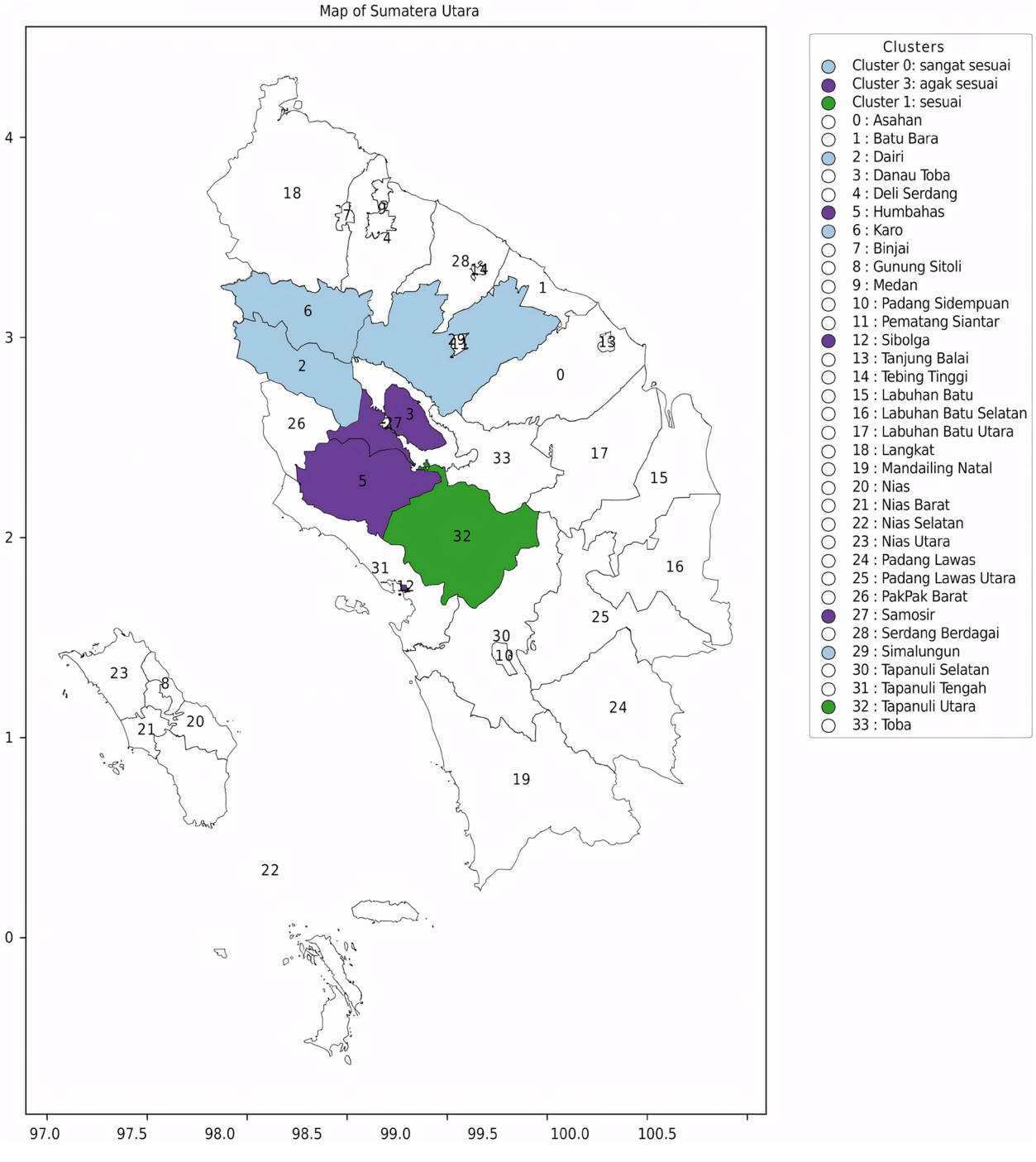

Fig. 6. Cluster visualization results based on color

Table 5

Cluster results for mapping

Name Kabupaten	Cluster	Number Kabupaten
Karo	0	6
Simalungun	0	29
Samosir	3	27
Tapanuli Utara	1	32
Sibolga	3	12
Dairi	0	2
Humbahas	3	5

Each number on the map represents one district or city, and a complete description can be seen in the legend on the right side. This type of clustering is usually used to support data-based decision making, such as in development planning, resource management, or determining regional priorities.

Based on the test results of Fig. 4, 5 PCA and K-means identified that the dominant factors affecting garlic yield are altitude, average temperature, and soil moisture. Regions with altitudes of 1200–1500 mdpl, temperatures of 15–20°C, and soil moisture are optimal. The resulting spatial map helps farmers in determining the optimal location for garlic cultivation. PCA and K-Means are effective in identifying the main factors and clustering garlic cultivation zones in the highlands. The results of spatial testing of regional location maps with cluster 0 stated that it is very suitable, spatial testing of regional location maps with cluster 3 is somewhat suitable and spatial testing of regional location maps with cluster 1 stated that it is suitable for the location of garlic cultivation land. The results of this study can be used to improve production efficiency and as a basis for spatial-based cultivation decision making. Evaluation testing is shown in Tables 6, 7.

Table 6

Test results

Datasets	Value K	Performance			
		Total dataset	Accuracy	Precision	Recall
Soil spatial	3	1500	89	92	84
	5	1500	89	90	90
	9	1500	92	97	96

The test results from the dataset used and based on the K value of clusters 3, 5, and 9 where the K = 9 value has accuracy = 92%, precision = 97% and recall = 96%. The test results are quite good, but need to be further evaluated depending on the priority of precision or recall.

Table 7

Test results

Datasets	Value K	Performance		
		Total customers	Inertia	Silhouette score
Soil Spatial	3	1500	0.49	0.89

Based on Table 7, spatial research of garlic cultivation land using PCA and K-Means successfully determined spatial land by conducting classification testing with testing accuracy using Inertia of 0.49% and using Silhouette score testing of 0.89%.

6. Discussion of the results of optimizing the spatial problem of garlic cultivation with machine learning

The results of this study indicate that the spatial research system for garlic cultivation land using PCA and K-means has significantly better classification performance compared to the study. This system combines the PCA and K-means algorithms directly into a digital spatial platform. so that the clustering results are directly visualized as an interactive map of potential garlic areas. This improvement is related to the selection of improved features by performing a cluster process unlike the features in previous studies [16]. The test results shown in Table 6 are based on the K values of clusters 3, 5, and 9, where the K value = 9 has accuracy = 92%, precision = 97% and recall = 96%. The results of determining spatial land by conducting a classification test with an accuracy test are shown in table 7 using Inerta of 0.49% and using the Silhouette Score test of 0.89%. This shows that the success of garlic cultivation spatially is greatly influenced by agroclimatic factors and the selection of features that produce garlic requires cold temperatures and good drainage, which are usually found in the highlands. Therefore, PCA not only organizes data, but also identifies the dominant factors in determining land availability. In contrast to previous studies [13], there are difficulties in adjusting irrigation systems based on atmospheric and soil variables that differ significantly between regions, as well as high costs associated with installing and maintaining WSN sensors in larger or remote agricultural areas.

The main issue raised is the spatial heterogeneity of land which causes low efficiency and yields of garlic cultivation in several areas.

Fig. 5 is a map of North Sumatra based on the results of clustering. This image shows the results of grouping areas in North Sumatra into three clusters:

– cluster 0 (very suitable): marked in blue;

– cluster 1 (suitable): marked in green;

– cluster 2 (quite suitable): marked in orange.

Each region such as North Tapanuli, Central Tapanuli, and Humbang Hasundutan is classified based on its characteristics in the relevant cluster. The geographical location and spatial relationships between regions are also well illustrated in this map.

Fig. 6 below is a map of North Sumatra Province showing the results of clustering areas based on the level of suitability. In general, this map divides areas in North Sumatra into three groups (clusters) based on color:

– light blue is an area that is very suitable for certain criteria;

– purple indicates an area that is somewhat suitable;

– green indicates an area that is suitable.

Each number on the map represents one district or city, and a complete description can be seen in the legend on the right side. This type of clustering is usually used to support data-based decision making, such as in development planning, resource management, or regional priority determination.

Using the PCA results, the K-means algorithm is applied to group the areas into 3 land suitability clusters:

– cluster 1: areas with cool temperatures, neutral pH, and high humidity – very suitable for garlic cultivation;

– cluster 2: areas with moderate temperatures, slightly acidic pH, and high rainfall – quite suitable;

– cluster 3: areas with high temperatures, too acidic or alkaline pH – not suitable. Mapping of the cluster results

shows that highland areas such as mountainous areas are included in Cluster 1 and are highly recommended as optimal locations.

Model evaluation based on K values of clusters 3, 5, and 9 where the K value = 9 has an accuracy of 92%, precision = 97% and recall = 96%. The test results are quite good, but need to be further evaluated depending on the priority of precision or recall, which indicates good cluster verification quality. Field validation of locations within clusters showed the best match between model results and actual high garlic productivity. These results support the overall objective of developing an accurate and practical garlic cultivation land selection system that is suitable for use in real-world agricultural environments, especially the Berastagi highlands area.

Although the proposed approach has several advantages, this study is not without limitations that need to be considered. One of the main shortcomings is the high dependence on the quality and coverage of the spatial data used, such as elevation maps, soil types, rainfall, temperature, and soil acidity levels. If the data has low resolution, is not up-to-date, or does not cover other important variables, the accuracy of the analysis results using PCA and K-means will decrease and may deviate from the actual land conditions. On the other hand, PCA as a linear-based dimension reduction technique has limitations in revealing non-linear patterns that may exist between environmental variables. Therefore, this method is less than ideal if the relationship between data is complex and non-linear. To overcome this, future research is recommended to combine more adaptive approaches such as Kernel PCA, Autoencoder, or other deep learning methods that are better able to capture non-linear relationships. The direction of future development also needs to expand the scope of the model from merely land suitability analysis to a predictive system for cultivation productivity. By combining agroclimatic data, agricultural management factors, and predictive machine learning techniques, the system built can provide more comprehensive recommendations, not only from an ecological perspective, but also based on estimates of possible production results.

7. Conclusions

1. The test results from the dataset used and based on the K value of clusters 3, 5, and 9 where the $K = 9$ value has accuracy = 92%, precision = 97% and recall = 96%. The test results are quite good, but need to be evaluated further depending on the priority of precision or recall.

2. The spatial research system of garlic cultivation land using PCA and K-means successfully determines the spatial land by conducting classification testing with testing accuracy using Inerta of 0.49% and using Silhouette Score testing of 0.89%.

3. PCA and K-means methods help in data-driven decision making for more efficient farmland development. The clustering results can be used by farmers, government, and agribusiness companies to determine the most suitable location for garlic cultivation. This technology is also in line with the government's goal of food self-sufficiency, thus supporting data and technology-based agriculture (smart farming).

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The author would like to thank the University of North Sumatra for the research grant and financial support for the USU Talent Applied Research Scheme Year 2023 with grant number 13388/UN5.1.R/PPM/2023.

Data availability

Manuscript has no associated data.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

Acknowledgment

The author would like to thank the University of North Sumatra for the research grant and financial support grant number Universitas Sumatera Utara.

References

1. Selvia, A., Dharanib, D., Gobikac, T., Harinid, S., Nithya, N. (2021). Onion Yield Prediction Based on Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12 (2). https://doi.org/10.17762/turcomat.v12i2.1972
2. Choi, J., Cho, S., Choi, S., Jung, M., Lim, Y., Lee, E. et al. (2024). Genotype-Driven Phenotype Prediction in Onion Breeding: Machine Learning Models for Enhanced Bulb Weight Selection. Agriculture, 14 (12), 2239. https://doi.org/10.3390/agriculture14122239
3. Tori, H., Dyanasari, Kholil, A. Y. (2023). Prospect Analysis of Onion (allium cepa L) Production in Indonesia. Indonesian Journal of Agriculture and Environmental Analytics, 2 (1), 1–14. https://doi.org/10.55927/ijaea.v2i1.2705
4. Sugiartini, E., Eris, F. R., Pancaningsih, E., Nurviani, O., Herawati, N. (2021). Studies on Cultivation of Several Varieties of Onion (Allium ascalonicum L.) under Plastic Shade during Rainy Season in Jakarta. IOP Conference Series: Earth and Environmental Science, 715 (1), 012044. https://doi.org/10.1088/1755-1315/715/1/012044
5. Utami, I. R. P., Roessali, W., Gayatri, S. (2023). Sustainability Analysis Of Onion Cultivation In Demak District, Central Java Province. Agrisocionomics: Jurnal Sosial Ekonomi Pertanian, 7 (3), 660–670. https://doi.org/10.14710/agrisocionomics.v7i3.17401

6. Razali, Nasution, Z., Rahmawaty, Hanum, C. (2023). Effect of Soil Texture on the Productivity of Two Shallot Varieties. Indonesian Journal of Agricultural Research, 6 (01), 43–50. https://doi.org/10.32734/injar.v6i01.8217
7. Gomes, J. D. J., Widaryanto, E., Ariffin, Wicaksono, K. P. (2019). The Test of Genotype Adaptation of Several Garlic Varieties on the Highland. OnLine Journal of Biological Sciences, 19 (4), 203–212. https://doi.org/10.3844/ojbsci.2019.203.212
8. Mukhibah, D., Sitanggang, I. S., -, A. (2023). Classification of Garlic Land Based on Growth Phase using Convolutional Neural Network. International Journal of Advanced Computer Science and Applications, 14 (6). https://doi.org/10.14569/ijacsa.2023.01406100
9. Ugarkovic, A., Oreski, D. (2022). Supervised and Unsupervised Machine Learning Approaches on Class Imbalanced Data. 2022 International Conference on Smart Systems and Technologies (SST), 159–162. https://doi.org/10.1109/sst55530.2022.9954646
10. Nurjulaiha, S., Kurniawan, R., Dikananda, A. R., Suprapti, T. (2025). Optimizing Naïve Bayes Algorithm Through Principal Component Analysis To Improve Dengue Fever Patient Classification Model. Journal of Artificial Intelligence and Engineering Applications (JAIEA), 4 (2), 1005–1015. https://doi.org/10.59934/jaiea.v4i2.798
11. Abed Mohammed, A., Sumari, P., Attabi, kassem. (2024). Hybrid K-means and Principal Component Analysis (PCA) for Diabetes Prediction. International Journal of Computing and Digital Systems, 15 (1), 1719–1728. https://doi.org/10.12785/ijcds/1501121
12. Jansson, N. F., Allen, R. L., Skogsmo, G., Tavakoli, S. (2022). Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, Sala area, Sweden. Journal of Geochemical Exploration, 233, 106909. https://doi.org/10.1016/j.gexplo.2021.106909
13. Rafrin, M., Muh. Agus, Putri Ayu Maharani (2024). IoT-Based Irrigation System Using Machine Learning for Precision Shallot Farming. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 8 (2), 216–222. https://doi.org/10.29207/resti.v8i2.5579
14. Kalpana, P., Smitha, L., Madhavi, D., Nabi, S. A., Kalpana, G., Kodati, S. (2024). A Smart Irrigation System Using the IoT and Advanced Machine Learning Model. International Journal of Computational and Experimental Science and Engineering, 10 (4). https://doi.org/10.22399/ijcesen.526
15. Wulandari, E., Al Hakim, R. R., Saputri, L. D., Syahdiar, I. A., Pangestu, A., Jaenul, A. (2021). Mr. Rytem, An IoT-Based Smart Irrigation System Application Design for Cultivation Engineering of Allium sativum Garlic in Lowland Conditions. Prosiding Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika. Available at: https://www.researchgate.net/publication/351270518_Mr_Rytem_An_IoT-Based_Smart_Irrigation_System_Application_Design_for_Cultivation_Engineering_of_Allium_sativum_Garlic_in_Lowland_Conditions
16. Liu, J., Yuan, J., Cui, J., Liu, Y., Liu, X. (2022). Contour Resampling-Based Garlic Clove Bud Orientation Recognition for High-Speed Precision Seeding. Agriculture, 12 (9), 1334. https://doi.org/10.3390/agriculture12091334
17. Sitanggang, I. S., Nurkholis, A., Annisa, Agmalaro, M. A. (2019). Garlic Land Suitability System based on Spatial Decision Tree. Proceedings of the International Conferences on Information System and Technology, 206–210. https://doi.org/10.5220/0009908002060210
18. Jiang, Y., Wang, T., Zhao, H., Shao, X., Cui, W., Huang, K., Li, L. (2019). Big Data Analysis Applied in Agricultural Planting Layout Optimization. Applied Engineering in Agriculture, 35 (2), 147–162. https://doi.org/10.13031/aea.12790
19. Rathor, A. S., Choudhury, S., Sharma, A., Nautiyal, P., Shah, G. (2024). Empowering vertical farming through IoT and AI-Driven technologies: A comprehensive review. Heliyon, 10 (15), e34998. https://doi.org/10.1016/j.heliyon.2024.e34998
20. Mohammed, M. A., Akawee, M. M., Saleh, Z. H., Hasan, R. A., Ali, A. H., Sutikno, T. (2023). The effectiveness of big data classification control based on principal component analysis. Bulletin of Electrical Engineering and Informatics, 12 (1), 427–434. https://doi.org/10.11591/eei.v12i1.4405
21. Iliyas, I. I., Boukari, S., Gital, A. Y. (2024). A Proposed Multilayer Perceptron Model and Kernel Principal Analysis Component for the Prediction of Chronic Kidney Disease. International Journal of Artificial Intelligence, 11 (2), 99–113. https://doi.org/10.36079/lamintang.ijai-01102.783
22. Ahmad, I. A., Jawad Al-Nayar, M. M., Mahmood, A. M. (2023). A comparative study of Gaussian mixture algorithm and K-means algorithm for efficient energy clustering in MWSN. Bulletin of Electrical Engineering and Informatics, 12 (6), 3727–3735. https://doi.org/10.11591/eei.v12i6.5707
23. Oti, E. U., Olusola, M. O., Eze, F. C., Enogwe, S. U. (2021). Comprehensive Review of K-Means Clustering Algorithms. International Journal of Advances in Scientific Research and Engineering, 07 (08), 64–69. https://doi.org/10.31695/ijasre.2021.34050
24. Guslendra, G., Defit, S., Bastola, R. (2021). K-Means and K-NN Methods For Determining Student Interest. International Journal of Artificial Intelligence Research, 6 (1). https://doi.org/10.29099/ijair.v6i1.222
25. Astorino, A., Gorgone, E., Gaudioso, M., Pallaschke, D. (2011). Data preprocessing in semi-supervised SVM classification. Optimization, 60 (1-2), 143–151. https://doi.org/10.1080/02331931003692557
26. Sami, O., Elsheikh, Y., Almasalha, F. (2021). The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease. International Journal of Advanced Computer Science and Applications, 12 (6). https://doi.org/10.14569/ijacsa.2021.0120695
27. Park, H.-J., Koo, Y.-S., Yang, H.-Y., Han, Y.-S., Nam, C.-S. (2024). Study on Data Preprocessing for Machine Learning Based on Semiconductor Manufacturing Processes. Sensors, 24 (17), 5461. https://doi.org/10.3390/s24175461