Journal of Information System Exploration and Research

https://joiser.shmpublisher.com

p-ISSN 2964-1160 | e-ISSN 2963-6361

Clustering Analysis of Temperature Humidity Index (THI) and Supporting Weather for Food Crop Cultivation in Tegal

Willy Yudha Perdana[1*], Aji Supriyanto[2]

[1,2]*Department of Technology and Industry, Universitas Stikubank, Indonesia*

DOI: https://doi.org/10.52465/joiser.v4i2.11

Received 28 May 2026; Accepted 20 June 2026; Available online 20 June 2026

Article Information

Keywords:
THI;
K-Means;
DBSCAN;
Clustering;
Crops

Abstract

The food crop agriculture sector faces serious challenges due to global climate change that disrupts the stability of conventional production systems. The Temperature Humidity Index (THI) is a crucial indicator for measuring environmental thermal comfort, which, combined with supporting weather parameters, can map the risk of crop failure. This study aims to analyze the clustering of THI and weather variables in the Tegal area using a Machine Learning approach. The dataset used is daily historical weather data for 10 years (2016–2025) from BMKG, including temperature (T), relative humidity (RH), solar radiation (SR), wind speed (WS), and rainfall (RF). The method includes preprocessing, normalization, THI calculation, and clustering using K-Means and DBSCAN. K-Means identified agro-climate vulnerability zones: Optimal, Alert, and Critical for food crop growth. DBSCAN effectively detected dominant cluster patterns and outliers of extreme weather anomalies. Internal evaluation shows K-Means performs better, with a Silhouette score of 0.4057 and Davies-Bouldin Index of 0.8391, compared to DBSCAN with 0.3957 and 2.7432. The results are expected to support farmers and policymakers in determining adaptive cropping patterns and mitigating climate change impacts in Tegal.

1. Introduction

The agricultural sector is the backbone of the economy in many regions, including Tegal City and Regency. Increasing food crop productivity is a top priority to ensure regional food security. However, a major challenge currently faced by farmers is climate change and unpredictable fluctuations in environmental conditions, which directly affect crop growth and yields. One important indicator that reflects the level of comfort and potential environmental stress for plants is the Temperature Humidity Index (THI) [1] . The THI integrates air temperature and humidity into a single value to assess the thermal suitability of an area [2] . Extreme THI fluctuations, whether too high or too low, can cause thermal stress

* ***Corresponding Author:***

Willy Yudha Perdana,
Department of Technology and Industry,
Universitas Stikubank,
Jl. Kendeng V, Bendan Ngisor, Semarang 50233, Indonesia.
Email: willyyudha0003@mhs.unisbank.ac.id

in plants, resulting in decreased photosynthesis rates, stunted growth, and the risk of crop failure. However, THI analysis cannot stand alone in determining agricultural land suitability; it requires integration with other supporting weather parameters such as rainfall, solar radiation, and wind speed to provide a comprehensive agro-climate picture. Therefore [3] , a deep understanding of THI patterns and their interactions with supporting weather in the Tegal region is crucial. Unfortunately, this information need has not been met well in the field. Available weather and THI data are often only raw numerical records that are difficult for farmers to directly interpret. This gap in understanding creates a real urgency for a system capable of translating raw environmental data into practical, ready-to-use information.

To overcome these problems, a computational method is needed that is able to identify hidden patterns from complex weather data interactions. Clustering algorithms in data mining offer an effective solution for grouping regions based on similarities in their agro-climatic characteristics. Two popular clustering algorithms with different approaches are K-Means and DBSCAN [3] , [4] . According to [5] K-Means works by dividing data into a number of predetermined clusters, while DBSCAN is able to find clusters with arbitrary shapes and automatically identify outliers that represent extreme weather anomalies. Considering the differences in geographic and microclimatic characteristics between Tegal City, which tends to be urban-coastal, and Tegal Regency, which is dominated by agricultural to mountainous areas, a comparison of the performance of these two algorithms is very important [6] . This study aims to conduct a clustering analysis of THI and supporting weather to map food crop cultivation areas in Tegal into potential zones (Optimal, Alert, Critical). This comparison not only determines which algorithm is more accurate in modeling local climate, but also provides strategic recommendations for farmers in determining adaptive planting schedules and mitigating the risks of climate change impacts [7] .

2. Literature Review

Many studies have utilized clustering techniques in agriculture. For example, research by [8] applied the K-Means Clustering method to group rice crop productivity data in Central Java. The results of this study successfully divided regions based on yield metrics, demonstrating the ability of K-Means to effectively identify agricultural spatial patterns.

Another study that focuses on food commodities is a comparison of the performance of K-Means and K-Medoids in West Kalimantan Province [9] . which emphasizes the important role of clustering in the classification of food agriculture areas and the determination of the best algorithm based on internal evaluations such as the Davies-Bouldin Index (DBI). In addition to land productivity, clustering techniques are also starting to be widely applied in the analysis of meteorological and agroclimatological data. Previous research shows that grouping weather element parameters is very effective for mapping ideal micro-climate areas for certain commodities [10] . In the context of thermal indices, clustering of temperature and humidity variables has been shown to be able to identify the level of vulnerability to environmental stress in vegetation in a more structured manner [11] .

Specifically for the application of the DBSCAN algorithm to spatial and environmental data analysis, several previous studies have proven the effectiveness of this method in producing optimal cluster structures [12] . The main strength of DBSCAN lies in its ability to find clusters with arbitrary shapes while identifying noise or outliers (extreme data points that do not belong to any group) [13] , [14] . This adaptive ability becomes very crucial when dealing with agroclimatology datasets with non-uniform or fluctuating data point distributions, thus playing an essential role in environmental resilience analysis and food insecurity risk mitigation at the regional level [15] .

The relevance of using these two algorithms (K-Means and DBSCAN) lies in the characteristics of historical weather data which often have both general patterns and extreme anomalies. In contrast to the server log research by [12] which combined the two methods in a hybrid manner, this research in the Tegal region positions K-Means and DBSCAN as comparative methods. K-Means is used to map general agro-climatic zones (Optimal, Alert, Critical), while DBSCAN is focused on capturing specific cluster patterns while detecting outliers that represent extreme weather anomalies [16] .

3. Method

Optimizing food crop cultivation, such as efforts to increase crop yields in the Tegal region, is greatly influenced by the ability to mitigate the risks of local microclimate change. This positive impact can be achieved through the accurate determination of the thermal comfort index or Temperature Humidity Index (THI). To map these climate characteristics, this study utilized 10 years of historical daily average data (from 2016 to 2025) obtained from the BMKG Tegal Meteorological Station. In its implementation, this study adopted a comparative experimental model that compared two main clustering algorithms in data mining: K-Means Clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The comparison of these two unsupervised learning methods aims to analyze the most optimal, adaptive, and informative algorithm for grouping and determining the THI threshold for food crops in the Tegal region. Operationally, the entire series of experiments was carried out through three main phases: Input, Process, and Output. The workflow and systematic stages of this study are presented in detail in Figure 1.

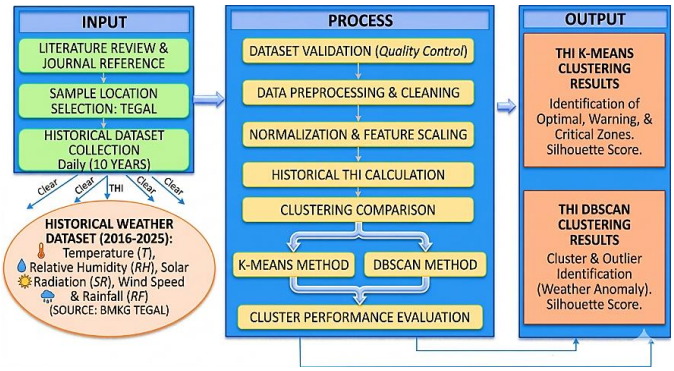


Figure 1. Flowchart of research stages

3.1. Data processing

In this stage, the research process is carried out systematically to transform raw data into meaningful information through several structured steps.

3.1.1. Input phase

This phase is the initial stage that underpins the entire agroclimatology research process. The process begins with a literature review and journal references to establish a strong theoretical foundation regarding microclimate parameters and their impact on food crops. The next step is the selection of sample locations: Tegal, which encompasses the Tegal Regency and City as the research locus due to its position as one of the most productive food crop areas in Central Java.

A crucial stage in this phase is the Collection of Daily Historical Datasets (10 Years) from 2016 to 2025, sourced directly from the Tegal Meteorology, Climatology, and Geophysics Agency (BMKG). The collected multivariate historical weather dataset includes five main parameters: Temperature (T), Relative Humidity (RH), Solar Radiation (SR), Wind Speed (WS), and Rainfall (RF). Through this scheme, temperature and humidity variables will be specifically extracted to form the Temperature Humidity Index (THI) value, while other supporting variables serve to enrich the context of the agroclimatology analysis of the agricultural environment in the Tegal region.

3.1.2. Process phase

The process phase describes the core of the computational methodology and Data Mining techniques applied in the research. The workflow begins with Dataset Validation (Quality Control) to ensure data integrity, followed by the Data Preprocessing & Cleaning stage to handle structural anomalies such as missing or corrupted data. Because the raw data has highly variable units and value ranges, Normalization and Feature Scaling are performed to equalize the data scale to avoid bias when calculating mathematical distances in the clustering algorithm.

After the data is in a clean and standard condition, a Historical THI Calculation is performed using a combined empirical formula of temperature and humidity. The THI value formed is then fed into the Clustering Comparison stage, where the data is processed in parallel using two different unsupervised learning algorithm approaches, namely the K-Means Method : Clustering data based on the closest distance (distance-based) to the center of mass (centroid), and the DBSCAN Method : Clustering data based on the

level of spatial density (density-based) under a certain radius. These two algorithms produce their respective clustering structures which are then evaluated objectively in the Cluster Performance Evaluation stage.

3.1.3. Output phase

The final phase is the output phase that presents the results of data processing into structured information that is ready to be interpreted for practical agricultural needs. The output from the process phase is specifically streamed into two main output components, namely: THI K-Means Clustering Results : This component produces a division of plant thermal comfort index zones that are clearly divided into 3 levels, namely Optimal, Alert, and Critical Zones. The density quality of this group formation is tested and validated using the Silhouette Score value. THI DBSCAN Clustering Results : This component produces natural density-based groupings to identify the main cluster patterns. The advantage of this DBSCAN output is its ability to identify Clusters and Outliers (Weather Anomalies). Days with extreme weather conditions that do not meet the minimum density limit are isolated separately as noise, and the quality of the separation is also validated using the Silhouette Score.

As a result of the research (Tegal Climate Agri-Tech) , feedback from the two clustering outputs was compared to draw conclusions about which method was most accurate. The final interpretation was used to map the risk of environmental stress on food crops in Tegal, which can be used by farmers and policymakers to determine climate adaptation strategies, such as adjusting planting calendars and irrigation management.

3.2. Temperature Humidity Index (THI)

The Temperature Humidity Index (THI) in food crops is an indicator used to measure the level of thermal comfort based on a combination of air temperature, relative humidity, solar radiation, wind speed, and rainfall. This index basically reflects how much influence temperature and humidity have on the ability of organisms, whether humans, animals, or plants, to adapt to the environment. According to [17] the THI concept was initially widely used in the field of animal husbandry to assess the level of heat stress in livestock, but has now been widely adapted in the field of agriculture to analyze the influence of microclimate conditions on plant productivity. The THI value is obtained through an empirical formula that combines the elements of temperature and humidity [18] . According to [19] one of the common formulas used as in Eq. (1).

$$\mathrm{THI} = \mathrm{T} - (0.55 - 0.0055 \times \mathrm{RH}) \times (\mathrm{T} - 14.5) \quad (1)$$

From the equation above, T is the air temperature (°C) and RH is the humidity (%) . This formula provides a numerical value that indicates the level of thermal comfort. The higher the THI value, the higher the level of heat stress experienced by the organism. In the context of plants, increased THI can cause physiological disorders, such as decreased photosynthesis rate, increased transpiration, and disruption of water balance in plant tissues.

3.3. K-Means machine learning

The K-Means algorithm is one of the most popular clustering methods in the partition-based clustering approach that aims to minimize the distance between data in one cluster to its centroid [20] . The main principle of K-Means according to [21] is to find the optimal position of the cluster center (centroid) so that intra-cluster variation is minimal, this algorithm works by iterating the centroid position update based on the average of each cluster member until the convergence condition is reached. The main advantages of K-Means are its simplicity, speed, and efficiency in processing large data [22] .

According to research [23] K-Means has been successfully used to cluster temperature and humidity data on agricultural land to accurately determine microclimate zones. According to [24] this algorithm is less effective when the resulting clusters have a non-spherical shape or high noise. Therefore, the integration of the K-Means method and the Temperature Humidity Index (THI) is an ideal solution for analyzing data with normal distribution characteristics and having clear separation boundaries between regions. Through this combination, THI plays an important role in simplifying microclimate indicators (temperature and humidity) into a single measurable thermal comfort index. Furthermore, the K-Means algorithm is tasked with objectively clustering based on the index values. This

hybrid approach not only improves the accuracy of zone mapping, but also simplifies the decision-making process in agricultural land management [25] .

The real implementation of the theoretical integration is reflected in the formation of three agroclimatic zones of plants according to [26] identified objectively by K-Means. Cluster 1 (Optimal) is formed in the middle area with cool-warm conditions covering the ideal temperature range of 22°C–27°C and humidity of 70%–85%, where these real conditions support maximum photosynthesis rate and balanced transpiration for healthy plant growth. As microclimate fluctuations occur, the algorithm identifies Cluster 2 (Alert) when the temperature begins to rise to the range of 28°C–32°C with humidity decreasing to 60%–69%, or vice versa when the temperature drops below 21°C with humidity increasing above 85%; this pattern triggers real conditions of plants starting to become stressed and vulnerable to high evaporation or fungal attacks due to high humidity. The extreme critical limits were finally separated by K-Means into Cluster 3 (Critical) which represents hot and dry conditions characterized by temperatures above 32°C and humidity levels below 60%, which physiologically forces leaf stomata to close completely to conserve plant water [27] .

3.4. DBScan machine learning

According to a research study by [28] , DBSCAN is able to identify extreme climate anomalies such as high temperature and low humidity in multivariate weather data. This makes DBSCAN very relevant for Temperature Humidity Index (THI) analysis because climate data often contains high spatial and temporal variations. In addition, DBSCAN also excels in detecting natural distribution patterns in environmental data, which are often not spherical or normally distributed as assumed by K-Means. According to [29] , this advantage makes DBSCAN much more adaptive in handling fluctuating microclimate parameter variability. By determining the radius parameter (eps) and the minimum number of points (minPts), DBSCAN can map environmental zones based on the real density level of data in the field.

Research study by [30] used DBSCAN to map agroclimatic zones based on daily temperature and humidity data, and the results showed that DBSCAN was more accurate in identifying extreme areas than K-Means. Meanwhile, research by [31] applied DBSCAN to detect high heat stress zones in horticultural crops using THI data, and succeeded in producing microclimatic zoning maps with higher accuracy than traditional methods.

4. Results and Discussion

Quantitative comparative experimental design . The approach focuses on the application of applied data mining (specifically unsupervised learning) to analyze agroclimatology time-series data. The main objective of this study is to compare the efficiency, internal accuracy, and practical relevance of a distance-based algorithm (K-Means) with a density-based algorithm (DBSCAN) in identifying Temperature Humidity Index (THI) patterns in the Tegal region.

4.1. Data processing (data pre-processing and transformation)

In the preprocessing stage, the table 1. below shows all the data structured into a Pandas dataframe. It is important to define the attributes of each feature before further data processing. The data is stored in .csv format, and attributes that function as features or labels are added to each record, consisting of nine columns of information (such as station name, coordinates, date, temperature, humidity, rainfall, irradiance, and wind speed).

Table 1. Processing data frame

No	Column	Non-Null Count	Dtype
1	Station_Name	3653 non-null	str
2	Latitude	3653 non-null	float64
3	Longitude	3653 non-null	float64
4	Date	3653 non-null	str
5	temperature	3653 non-null	float64
6	humidity	3653 non-null	int64
7	rainfall	3653 non-null	str
8	radiation	3653 non-null	float64
9	Kec_angin	3653 non-null	str

Normalization and initial checks were performed to verify the completeness of the data (no missing data) and the appropriate data types. Based on Table 2. the data structure check, it was found that out of a total of 3,653 records (indices 0 to 3652), all features had a non-null count of 3,653. Therefore, there were no missing values, and all 3,653 valid records were used for further prediction analysis.

Table 2. Data transformation

No	Station_Name	Latitude	Longitude	Date	T	RH	RR	SS	FF
0	BMKG_tegal	-6.86843	109.12103	2016-01-01	28.7	82	0.0	8.0	2.0
1	BMKG_tegal	-6.86843	109.12103	2016-01-02	28.6	82	0.0	6.5	3.0
2	BMKG_tegal	-6.86843	109.12103	2016-01-03	28.8	84	21.0	6.5	2.0
3	BMKG_tegal	-6.86843	109.12103	2016-01-04	28.2	85	18.6	9.5	2.0
4	BMKG_tegal	-6.86843	109.12103	2016-01-05	28.9	81	0.0	4.5	2.0
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
3648	BMKG_tegal	-6.86843	109.12103	2025-12-27	27.1	87	6.5	4.0	2.0
3649	BMKG_tegal	-6.86843	109.12103	2025-12-28	28.8	82	38.7	2.0	2.0
3650	BMKG_tegal	-6.86843	109.12103	2025-12-29	27.6	85	35.6	8.0	2.0
3651	BMKG_tegal	-6.86843	109.12103	2025-12-30	28.2	83	8.6	6.6	2.0
3652	BMKG_tegal	-6.86843	109.12103	2025-12-31	26.0	91	0.6	6.3	1.0

4.2. Average temperature trend analysis 2016-2025

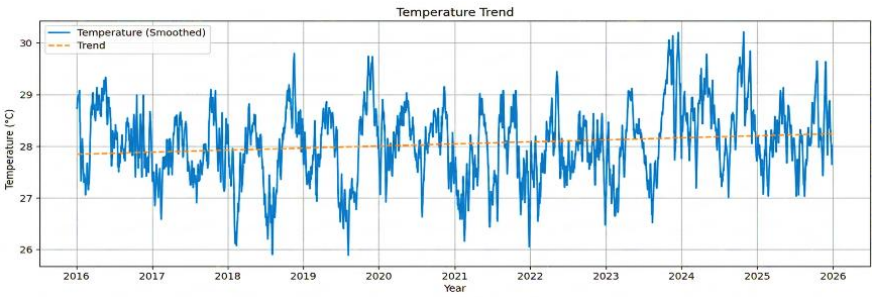


Figure 2. Average temperature trend

The Figure 2. "Temperature Trend" graph above shows the fluctuations in smoothed temperature data over a ten-year period from 2016 to the end of 2025. Although the daily or monthly data exhibits a highly dynamic and recurring pattern of fluctuations with the lowest temperatures reaching below 26°C (as in 2018 and 2019) and the highest temperatures reaching above 30°C (as in late 2023 and 2024) there is a consistent long-term upward trend in temperature overall. This increase is clearly depicted by the orange dotted line (Trend) which moves gradually upward from an average of around 27.8°C at the beginning of 2016 to nearly 28.3°C at the end of 2025, indicating a gradual warming in the region or location of observation over the past decade.

4.3. Average humidity trends 2016-2025

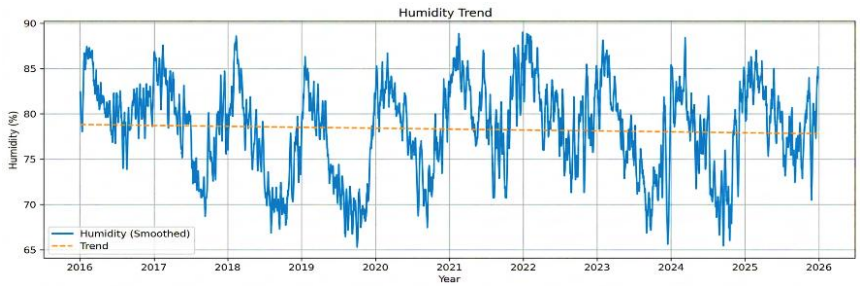


Figure 3. Average humidity trend

The "Humidity Trend" graph in the Figure 3. above shows the fluctuations in smoothed humidity values over a ten-year period from 2016 to the end of 2025. The data exhibits a highly fluctuating annual

cyclical pattern, with the highest humidity points periodically soaring to nearly 89% (as seen in mid-2018, 2021, and 2022) and the lowest points dropping to around 65% at the end of 2019 and 2024. Behind these sharp seasonal fluctuations, the orange dotted line (Trend) indicates a gradual, linear, long-term downward trend in humidity, moving down from an average of around 79% at the beginning of 2016 to nearly 78% at the end of 2025. This gradual decrease in humidity over the decade is directly proportional to the characteristics of the macroclimate where increases in air temperature are generally followed by a decrease in relative humidity in the observation area.

4.4. THI trend analysis with both temperature and humidity variables

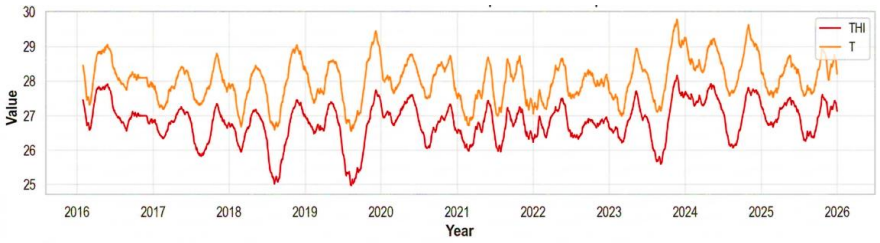


Figure 4. THI Trend with T variable

The Figure 4. above shows a comparison of the fluctuations in the THI (Temperature-Humidity Index, red line) and Air Temperature or T (Temperature, orange line) values from 2016 to the end of 2025. From this visualization, it is clear that there is a very harmonious (in-phase) movement pattern between the two variables, where every increase or decrease in the temperature (T) value is immediately followed by a change in the same direction in the THI value. Annual seasonal fluctuations show that the highest temperature (T) values approached or penetrated 30 at the end of 2023 and the end of 2024, which simultaneously pushed the THI value to its peak in the range of 28 to almost 29. Conversely, the lowest points of air temperature below 27 (such as in mid-2018 and mid-2019) were also linear with a sharp decline in the THI value to touch the lowest limit at 25. The consistency of the alignment of these curves confirms that air temperature variability plays a very dominant role and is directly proportional in controlling the rise and fall of the thermal comfort (THI) value at the observation location throughout the last decade.

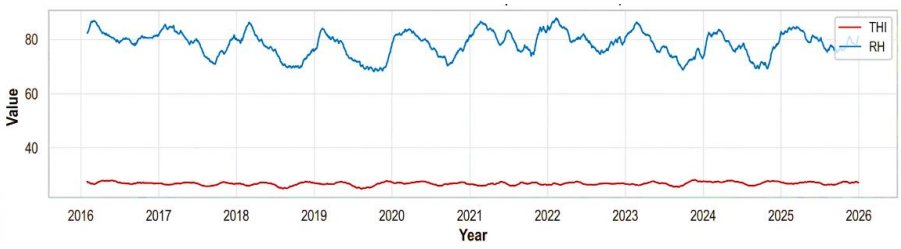


Figure 5. THI Trend with RH variable

The Figure 5. above shows a comparison of the fluctuations in the THI (Temperature-Humidity Index, red line) and relative humidity or RH (Relative Humidity, blue line) values over a ten-year period from 2016 to the end of 2025. From this visualization, a fairly clear negative or inverse correlation relationship can be seen between the two variables at their extreme moments, where when the humidity (RH) value drops to its lowest point in the range of 65% to 70% (such as in mid-2018, late 2019, and late 2024), the THI line actually moves up to reach its highest peak approaching 28. Conversely, when air humidity soars high to penetrate 85% (such as in early 2016, mid-2021, and early 2022), the THI index tends to press down to a lower level in the range of 25 to 26. This annual seasonal pattern confirms that the decrease in air humidity which is usually accompanied by an increase in extreme temperatures significantly contributes to the increase in the THI (air load index) value. heat) in the observation area over the past decade.

4.5. THI analysis with five variables from 2016-2025

The distribution of THI value variables with the main variables being temperature and humidity with the three supporting microclimate variables in the Tegal region consisting of Rainfall, Sunlight, and Wind Speed can be seen in the following table 3.

Table 3. Results of THI processing with five variables from 2016-2025

Statistik	Temperature	Humidity	Rainfall	Sunshine	Wind speed	THI
count	3653.000000	3653.000000	3653.000000	3653.000000	3653.000000	3653.000000
mean	28.041007	78.317000	5.214536	6.493676	3.661648	26.414428
std	0.911740	5.992029	14.667004	3.006059	1.933010	0.745302
min	24.500000	49.000000	0.000000	0.000000	0.000000	22.747750
25%	27.400000	74.000000	0.000000	4.500000	2.000000	25.988200
50%	28.100000	79.000000	0.000000	7.300000	3.000000	26.488750
75%	28.700000	83.000000	2.200000	8.800000	5.000000	26.919050
max	32.700000	95.000000	189.100000	42.800000	20.000000	29.697000

The descriptive statistics summary table above presents the characteristics of six meteorological variables with a total of 3,653 observations. From these data, the Temperature variable has an average value (mean) of 28.04°C with relatively stable fluctuations because its standard deviation (std) is less than one (0.91), while the average air humidity is at a fairly high level of 78.32% with a range between 49% and 95%. The THI (Temperature-Humidity Index) index recorded an average thermal comfort value of 26.41, with the most extreme conditions reaching a maximum index of 29.70. On the other hand, the Rainfall variable shows the most contrasting and highly unequal variability or data distribution, where the quartile value up to 50% is still at 0.00 mm (indicating that the majority of days do not experience rain), but the maximum value jumps drastically to 189.10 mm, indicating that very heavy rain events have occurred in the observation area.

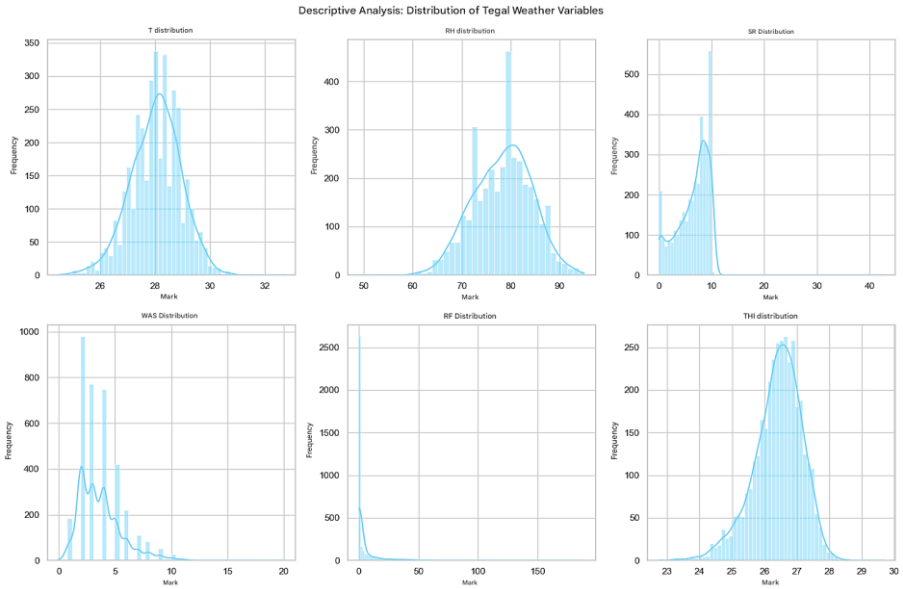


Figure 6. THI Trend with five variables

Based on the Figure 6. above, it displays a histogram graph along with a density estimation curve (Kernel Density Estimation/KDE) to see the frequency distribution of six meteorological parameters (T, RH, SR, WS, RF, and THI). Based on this visualization, the temperature variable (T Distribution) and the thermal comfort index (THI Distribution) show a normal (symmetrical) distribution pattern or approach a bell shape. In the T parameter, the data is centered around the value of 28°C, while in the THI it is centered around the value of 26.5. This symmetrical pattern indicates that daily fluctuations in temperature and thermal index in the observation area tend to be stable and consistent at their average values, with extreme events (too cold or too hot) being very rare.

Meanwhile, the RH (Humidity) distribution graph shows a negative slope pattern or is skewed to the left (left-skewed). The peak of the data population dominates the high value area, namely in the range of 75% to 85%, with the distribution tail extending towards lower values (around 50%). This emphasizes the characteristics of tropical regions that are dominated by humid air conditions throughout the year. In contrast to humidity, the SR (Solar Radiation) and WS (Wind Speed) distribution graphs show a right-skewed tendency. In the SR distribution, the data peaks below a value of 10 and is immediately interrupted, indicating the limits of daily radiation intensity, while in wind speed (WS), the highest frequency is gathered at low speeds (2–5) and slopes down to values 10 and above, indicating that strong winds are a rare phenomenon.

4.6. Correlation matrix between microclimate parameters

Based on the correlation matrix in the heatmap Figure 7. below, it can be concluded that the strongest inverse (negative) relationship occurs between humidity and solar radiation (-0.53) and humidity and temperature (-0.39), which indicates that the higher the intensity of sunlight and air temperature, the lower the humidity level. Conversely, the most significant unidirectional (positive) relationship was found between humidity and rainfall (0.31), indicating that humid air contributes positively to the potential for rain. Meanwhile, most other parameters such as rainfall and wind speed (-0.08) or temperature (0.11) show very weak correlations, indicating that these weather factors do not dominate each other linearly and are influenced by more complex atmospheric dynamics.

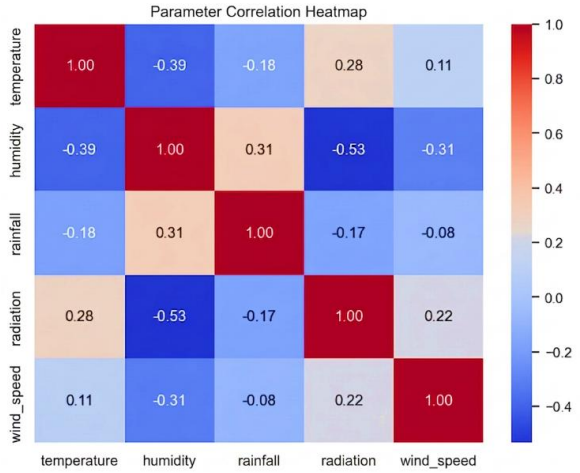
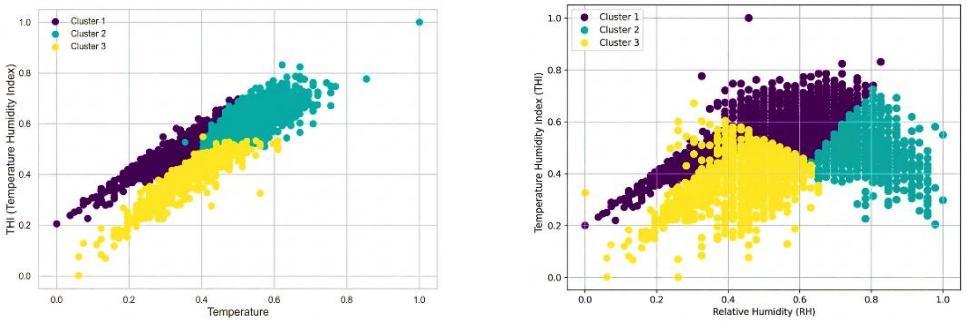


Figure 7. Heatmap of correlation between parameters

4.7. Results of K-Means clustering implementation



(A) (B)

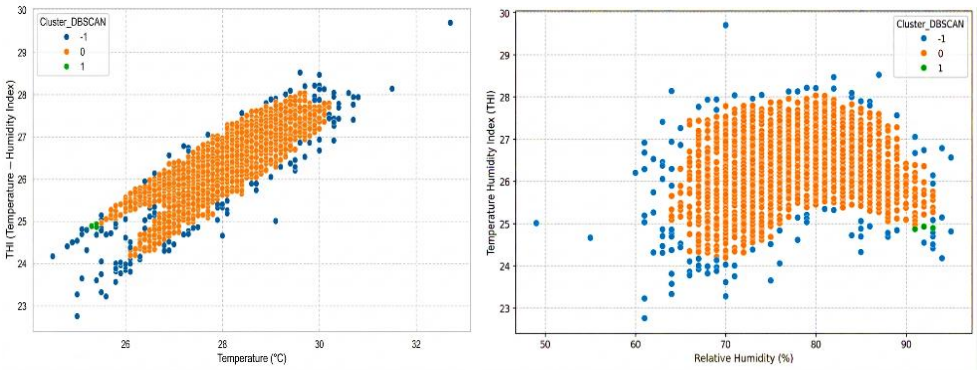
Figure 8. K-Means clustering results

Based on Figure (A), based on microclimate characteristics and agricultural implications, the three clusters divide the research area into three different comfort threshold levels. Cluster 1 (Purple) locks the data points in the lower left area of the graph, representing the Optimal Microclimate Zone (Ideal Temperature). In this stress-free zone, low surface air temperatures (24 ° C – 27.5 ° C) maintain the stability of plant physiological processes without the risk of dehydration due to excessive evaporation.

Furthermore, Cluster 2 (Tosca Green) occupies a transition area in the middle of the graph, depicting the Alert Microclimate Zone (Light Stress). This area reflects the average daily weather baseline of the tropical region, where plants begin to adapt to light stomata but have not yet disrupted the plant's core metabolism. Meanwhile, Cluster 3 (Yellow) groups the data that accumulates in the upper right side, indicating the Heat Microclimate Zone (Severe Stress). This period reflects very hot and dry days (29.5 ° C > 31 ° C), where high THI values risk triggering heat stress which can reduce plant productivity and disrupt the comfort of farmers working in the field.

The comparison graph of clustering results in figure (B) above shows the contrasting characteristics of the K-Means and DBSCAN methods in grouping data distribution based on the variables of Air Humidity (RH) and THI Calculation Results. In the K-Means graph, the algorithm divides the data rigidly into three clean cluster regions without gaps based on geometric distance, where the boundaries between groups are separated linearly following the gradient of decreasing humidity values and increasing THI index. In contrast, the DBSCAN graph presents a more flexible mass density-based approach by successfully detecting one very dense core group in the center, while data points that fall outside the minimum density limit (especially in conditions of very low humidity below 65% or very high humidity above 90%) are objectively identified as outliers (outliers or noise with a red label -1). Through this visual comparison, it is clear that DBSCAN is superior in maintaining the integrity of the natural cluster structure of the data while filtering out extreme weather anomalies, while K-Means is very effective if the objective of the analysis is to map all data points into absolute category zones.

4.8. DBScan clustering implementation results



(A) (B)

Figure 9. DBSCAN clustering results

Based on the scatter plot graph (A) of the DBSCAN clustering results above, it can be seen that the algorithm successfully identified two main data groups (cluster 0 in orange and cluster 1 represented by the green dot in the lower left corner) and separated a large number of outliers or anomalous data (label -1 in blue) scattered around the main pattern. The data is massively dominated by cluster 0 which forms a dense positive linear trend, indicating a very strong direct correlation where the increase in temperature values (ranging from 25°C to more than 30°C) is consistently followed by an increase in THI values (between 24 and 28). Meanwhile, anomalous points (label -1) were successfully filtered by DBSCAN in the outer edge area of the density, including one extreme outlier in the upper right corner with a temperature approaching 33°C and a THI almost reaching 30, which indicates very unusual weather or climate conditions compared to the majority of other historical data.

The graph (B) shows the distribution of the new DBSCAN clustering results, the data clustering now focuses on the relationship between Air Humidity (%) and THI (Temperature Humidity Index), where the algorithm successfully separates most of the data into dense main groups (cluster 0 in orange), detects very small minor clusters (cluster 1 in green at the high humidity limit of around 91-93%), and filters out outlier points (label -1 in blue). The main data pattern of cluster 0 forms a curved distribution (parabolic/arc-like pattern) that gathers in the air humidity range of around 63% to 93% with a fluctuating THI index value between 24 to 28. The separation of noise data (label -1) looks very effective in areas outside the density limit, especially in low air humidity conditions below 60% as well as one extreme outlier at the top of the graph with a humidity level of around 70% but has a very high THI value approaching 30, indicating the recording of anomalous or unusual weather conditions in the Tegal area.

4.9. Comparative Analysis of KMeans and DBScan Performance

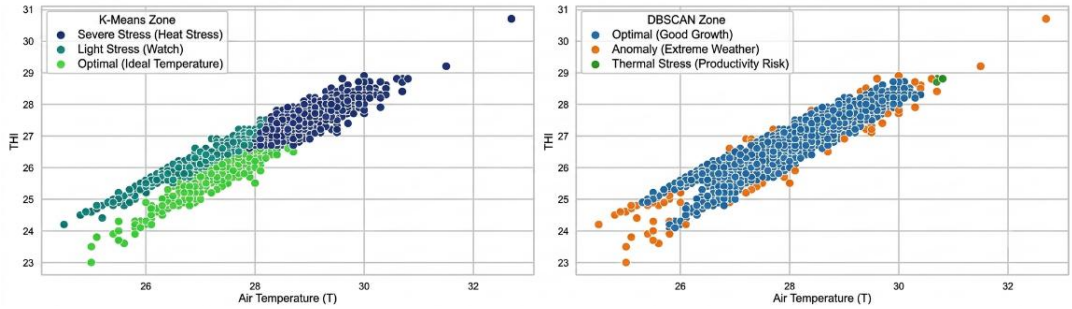


Figure 10. Comparison of KMeans and DBScan clustering (temperature and THI features)

Visualization in the Figure 10. above, shows a comparison of the clustering results of the relationship between Air Temperature (T) and THI using two different algorithms, namely K-Means and DBSCAN. In the Zona_KMeans graph (left), the algorithm divides the data in a structured manner into three stress level regions based on certain value limits: "Optimal (Ideal Temperature)" in the low temperature area, "Mild Stress (Alert)" in the middle area, and "Severe Stress (Heat Stress)" which dominates when the temperature and THI soar high. In contrast, the Zona_DBSCAN graph (right) groups the majority of the data into one large cluster assessed as "Optimal (Good Growth)" conditions, while data points that are outside the main density, either too low, too high, or scattered on the periphery, are detected as "Anomalies (Extreme Weather)", leaving a few points at the upper end as "Thermal Stress (Productivity Risk)". Overall, this analysis shows that K-Means is more suitable for creating consistent categorization of nested regions, while DBSCAN is very effective in separating normal data patterns from extreme anomalous points or outliers.

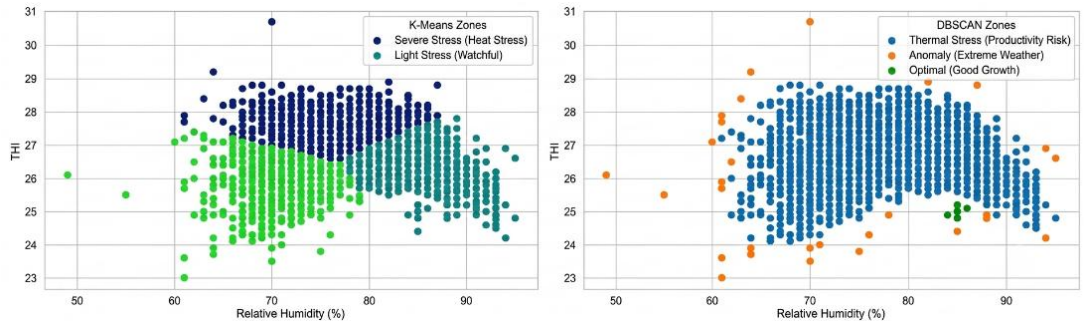


Figure 11. Comparison of KMeans and DBScan clustering (humidity and THI features)

Based on the visualization of the comparison of the humidity Figure 11. above, a very contrasting characteristic difference is seen between the K-Means and DBSCAN algorithms in grouping data on the relationship between Air Humidity (%) and THI for food crops. On the one hand, K-Means (left graph) divides the entire data variability space rigidly and based on distance into three clean linear agroclimatic zones, namely Optimal/Ideal Temperature at the bottom (THI < 27), Mild Stress in the middle (THI 27–28), and Severe Stress/Heat Stress at the top (THI > 28). On the other hand, DBSCAN (right graph) works flexibly based on the density of data points, thus successfully filtering out random points on the outer edges as Anomaly/Extreme Weather data (orange color) including one extreme outlier with THI close to 31 at 70% humidity while the majority of the main data distribution is locked into one large, dense cluster of Thermal Stress/Productivity Risk (blue color), leaving a small minor cluster of Optimal/Good Growth (green color) clustered exclusively in the very high humidity area above 85%.

Table 4. Evaluation Metrics of K-Means and DBSCAN performance comparison

No	Model Testing Metrics	K-MEANS	DBSCAN
1.	Limit Validation Accuracy	99.86%	99.59%
2.	Relative Accuracy of Structure	70.28%	69.78%
3.	Silhouette Score	0.4057	0.3957

4.	Davies-Bouldin Index	0.8391	2.7432
5.	Number of Main Clusters	3	3

The performance comparison evaluation metrics table 4. in the image above shows that the K-Means algorithm demonstrates superior performance compared to DBSCAN in grouping data into 3 main clusters. The superiority of K-Means is clearly visible from the higher Silhouette Score value (0.4057 vs 0.3957) and the much smaller Davies-Bouldin Index value (0.8391 vs 2.7432), which indicates that the clusters generated by K-Means have a clearer separation and a denser internal structure. In addition, K-Means also leads slightly in terms of accuracy, both in Boundary Validation Accuracy (99.86% vs 99.59%) and Relative Structure Accuracy (70.28% vs 69.78%). Overall, this metric analysis confirms that for these data characteristics, the K-Means centroid-based approach is able to form a much more optimal and efficient region partition than the density-based approach of DBSCAN.

.

5. Conclusions

Based on the results of the analysis and discussion, it can be concluded that the application of the K-Means and DBSCAN methods has proven to be very effective and reliable in mapping the Temperature Humidity Index (THI) to detect the risk of thermal stress in food crop cultivation in the Tegal City/Regency area. The K-Means method has the advantage of providing a very proportional and spatially balanced zoning division, where Tegal's climate variability is successfully grouped regularly into three distinct agro-climate physical levels , namely the Optimal Zone (Ideal Temperature), Alert Zone (Light Stress), and Heat Stress Zone (Severe Stress) with Internal evaluation shows that the K-Means method produces a more distinct zoning grouping with a Silhouette score of 0.4057 and a Davies-Bouldin Index of 0.8391, outperforming the DBSCAN method which has a Silhouette score of 0.3957 and a Davies-Bouldin Index of 2.7432.

Acknowledgements

The authors would like to express their sincere gratitude to the Meteorology, Climatology, and Geophysics Agency (BMKG) Tegal Meteorological Station for providing the historical weather dataset used in this study. The authors also thank the Department of Technology and Industry, Universitas Stikubank, for the academic support and research facilities that contributed to the completion of this research.

Credit Authorship Contribution Statement

Willy Yudha Perdana : Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Visualization, Validation, Writing – original draft, Project administration. **Aji Supriyanto**: Conceptualization, Methodology, Supervision, Validation, Resources, Writing – review & editing.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The dataset used in this study consists of daily historical weather data from 2016 to 2025 obtained from the BMKG Tegal Meteorological Station, including temperature, relative humidity, solar radiation, wind speed, and rainfall data. The processed data and source code supporting the findings of this study are available from the corresponding author upon reasonable request.

Use of Artificial Intelligence

The authors used artificial intelligence tools solely to assist with language refinement, grammar checking, and improving the readability of the manuscript. Artificial intelligence was not used to generate research ideas, collect data, perform data analysis, interpret the results, or formulate the conclusions.

All scientific content, analyses, interpretations, and the final version of the manuscript were reviewed, verified, and approved by the authors, who take full responsibility for the content of this article.

Reference

[1] A. Harmin, M. Isra, M. Bravikawati, and A. Naziah, "Analysis of Temperature Humidity Index (THI) on Thermal Comfort with the Addition of Green Plants to Multi-Storey Buildings," vol. 9, no. 1, pp. 294–298, 2025.

[2] Afandi and A. Supriyanto, "Thermal Comfort Projection on the Northern Coast of Central Java Using Machine Learning," *J. RESTI* , vol. 9, no. 4, pp. 865–878, 2025, doi: 10.29207/resti.v9i4.6537.

[3] S. Iskandar, L. Sebastian, and E. Rusdiyanto, "Analysis of Thermal Comfort Levels Using the Discomfort Index Method in South Putussibau District, Kapuas Hulu Regency," vol. 13, no. 1, pp. 224–230, 2025.

[4] Gladys Pra Adissha Nadira and Rizaldy Khair, "Comparative Analysis of K-MEANS and DBSCAN Clustering Algorithms on Community Economic Conditions in Pulo Brayan Darat 1 Subdistrict," *J. Sist. Inf. and Comput. Science.* , vol. 3, no. 2, pp. 86–95, 2025, doi: 10.59581/jusiik-widyakarya.v3i2.5230.

[5] S. Mutiah, Y. Hasnataeni, A. Fitrianto, and LMRD Jumansyah, "Comparison of K-Means and DBSCAN Clustering Methods in Identifying Household Groups Based on Socio-Economic Facilities in West Java In today's digital era, the amount of data available from various fields, including social and economic, continues," vol. 09, no. September, pp. 247–260, 2024.

[6] BL Syaefullah *et al.* , "The Effect of Land Cover Distribution on Temperature and Humidity Index in Batu City," *J. Animal Science. and Vet. Trop. (Journal Trop. Anim. Vet. Sci.* , vol. 11, no. 3, pp. 113–122, 2022, doi: 10.46549/jipvet.v11i3.167.

[7] MA Septianto, A. Faqih, and AR Rinaldi, "Clusterization of Agricultural Production Data in Cirebon Regency Using the K-Means Algorithm," *J. Inform. and Tech. Electro Terap.* , vol. 13, no. 2, 2025, doi: 10.23960/jitet.v13i2.6174.

[8] S. Wijayanto and M. Yoka Fathoni, "Grouping Rice Plant Productivity in Central Java Using the K-Means Clustering Method," *Jupiter* , vol. 13, no. 2, pp. 212–219, 2021.

[9] T. Margareta, N. Satyahadewi, and R. Pertiwi, "Comparison of the Performance of the K-Means and K-Medoids Algorithms in Clustering Individual Food Crop Farming Businesses in West Kalimantan Province," *J. Forum Anal. Stat.* , vol. 5, no. 1, pp. 35–46, 2025, doi: 10.57059/formasi.v5i1.91.

[10] MI Apriyatama, A. Tusi, and WR Diding, "Jurnal of Agricultural Biosystem Engineering Monitoring of Vapor Pressure Deficit (VPD) in a Naturally Ventilated Greenhouse," pp. 1–7, 2025.

[11] S. Susiyanti, N. Nasrul, M. Maddatuang, R. Ruliana, and R. Maru, "Spatio Temporal Modeling in Analyzing Temperature Humidity Index Using Google Earth Engine in South Sulawesi: Impact Analysis and Sustainable Mitigation Efforts," *Jambura Geosci. Rev.* , vol. 7, no. 2, pp. 105–117, 2025, doi: 10.37905/jgeosrev.v7i2.30644.

[12] BN Sari and A. Primajaya, "Application of Dbscan Clustering for Rice Farming in Karawang Regency," *J. Inform. and Comput.* , vol. 4, no. 1, pp. 28–34, 2019.

[13] Y. Lukman, IS Sitanggang, and MK Dewi, "Spatiotemporal Analysis of Fire Hotspots and PM2.5 in Riau, Jambi, and South Sumatra," vol. 12, no. 2021, pp. 212–227, 2023.

[14] NK Lay, D. Arisandi, and HJ Christanto, "Comparative Analysis of K-Means and DBSCAN Algorithms for Clustering Disaster-Prone Areas," vol. 7, no. 3, pp. 1875–1886, 2025, doi: 10.47065/bits.v7i3.8727.

[15] A. Ramadhan, K. Prawita, MA Izzudin, and G. Amandha, "Analysis of national food security strategies and clustering in facing the Covid-19 pandemic," *Technol. Pangan Media Inf. and Komun. Ilm. Technol. Pertan.* , vol. 12, no. 1, pp. 110–122, 2021, doi: 10.35891/tp.v12i1.2179.

[16] SDK Wardani, AS Ariyanto, M. Umroh, and D. Rolliawati, "Comparison of K-Means, Db Scanner & Hierarchical Clustering Method Results for Market Segmentation Analysis," *JIKO (Journal of Inform. and Computer)* , vol. 7, no. 2, p. 191, 2023.

[17] F. Fauzi, I. Kharisudin, R. Wasono, TW Utami, and IW Harmoko, "Thermal Stress Projection Based on Temperature-Humidity Index (Thi) Under Climate Change Scenario," *J. Meteorol. and Geofis.* , vol. 24, no. 1, pp. 65–73, 2023, doi: 10.31172/jmg.v24i1.867.

[18] M. Asghari, GF Ghalhari, EA Pirposhteh, and SF Dehghan, "Spatio-Temporal Evolution of the

Thermo-Hygrometric Index (THI) during Cold Seasons: A Trend Analysis Study in Iran," *Sustain.* , vol. 14, no. 24, 2022, doi: 10.3390/su142416774.
[19] R. Wu *et al.* , "Effects of the Bamboo Communities on Microclimate and Thermal Comfort in Subtropical Climates," *Forests* , vol. 14, no. 6, 2023, doi: 10.3390/f14061231.
[20] CJ Silalahi, A. Situmorang, and JF Naibaho, "Implementation of the K-Means Clustering Method to Map Potential Rice Producing Areas in North Sumatra Province," *Methotika J. Ilm. Tek. Inform.* , vol. 2, no. 2, pp. 49–57, 2022.
[21] H. Februariyanti, T. Khristianto, A. Jananto, and E. Nurraharjo, "Analysis of ANFIS-IoT-Based Humidity Control System for Plant Cultivation Room Analysis," vol. 8275, no. July, pp. 119–133, 2025.
[22] D. Ratna, J. Sari, and NY Permana, "Use of Differential Evolution Algorithm for Parameter Optimization in Weather Prediction Models," *J. ICT Inf. Commun. Technol.* , vol. 14, no. 2, pp. 2086–7867, 2023.
[23] UT Suryadi, P. Studi, T. Informatics, K. Clustering, and N. Red, "*1 , #2," vol. 13, no. 2, pp. 130–140, 2020.
[24] G. Andrian, D. Arisandi, and T. Handhayani, "Clustering Meteorological Data in Eastern Indonesia Using the K-Means and Fuzzy C-Means Methods," *INTI Nusa Mandiri* , vol. 18, no. 2, pp. 100–106, 2024, doi: 10.33480/inti.v18i2.5039.
[25] QA Kartika, R. Hidayat, and RH Virgianto, "Changes in Temperature Humidity Index (THI) on Java Island from 1981 to 2019," 2021. doi: 10.22146/mgi.63432.
[26] F. Fauzi, I. Kharisudin, R. Wasono, TW Utami, and IW Harmoko, "THERMAL STRESS PROJECTION BASED ON TEMPERATURE- HUMIDITY INDEX (THI) UNDER CLIMATE CHANGE SCENARIO," no. 18, pp. 65–73, 2023.
[27] JI Lingkungan, T. Wati, IP Iklim, and B. Meteorologi, "Analysis of Comfort Levels in DKI Jakarta Based on the THI Index (Temperature Humidity Index)," vol. 15, no. 1, pp. 57–63, 2017, doi: 10.14710/jil.15.1.57-63.
[28] S. Wibisono, MT Anwar, A. Supriyanto, and IHA Amin, "Multivariate weather anomaly detection using DBSCAN clustering algorithm," *J. Phys. Conf. Ser.* , vol. 1869, no. 1, 2021, doi: 10.1088/1742-6596/1869/1/012077.
[29] MR Zuhdi, HS Aljauhar, AAR Fernandes, and NWS Wardhani, "Comparison of Dbscan and K-Means Cluster Analysis With Path-Anova in Clustering Waste Management Behavior Patterns," *J. Tek. Inform.* , vol. 6, no. 1, pp. 105–112, 2025, doi: 10.52436/1.jutif.2025.6.1.4183.
[30] A. Kristianto, E. Sediyono, and KD Hartomo, "Implementation of dbscan algorithm to clustering satellite surface temperature data in Indonesia," *Register. J. Ilm. Teknol. Sist. Inf.* , vol. 6, no. 2, pp. 109–118, 2020, doi: 10.26594/register.v6i2.1913.
[31] X. Tu, C. Fu, A. Huang, H. Chen, *and J Environ. Res. Public Health* , vol. 19, no. 9, 2022, doi: 10.3390/ijerph19095153.