JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 8 Number 1 | June 2023: 1-10
DOI: 10.15575/join.v8i1.1031

Regression Analysis for Crop Production Using CLARANS Algorithm

Arie Vatresia[1], Ruvita Faurina[2], Yanti Simanjuntak[3]
[1]Research Center for Computing, The National Research and Innovation Agency, Bogor, Indonesia
[1,2,3]Informatics, Engineering Faculty, Universitas Bengkulu, Indonesia

Article Info

Article history:

Received Feb 17, 2023
Revised April 05, 2023
Accepted April 11, 2023

Keywords:

CLARANS
Clusters
Crop Production
Regression
Rainfall

ABSTRACT

Crop production rate relies on rainfall over Rejang Lebong district. Data showed a discrepancy between increased crop production and rainfall in Rejang Lebong District. However, the spatiotemporal distribution of the crop variable's dependencies remains unclear. This study analyses the relationship between rainfall and crop production rate in the Rejang Lebong district based on the performance of the machine learning method. In addition, this research also performed regression analysis to carry out rainfall clusters and crop production. This order provides information in the form of cluster results to determine how much the rainfall variable influences the crop production rate in each cluster. Harnessing the Elbow, CLARANS, Simple Linear Regression, and Silhouette Coefficient methods, this study used 231 rainfall data sourced from the Bengkulu BMKG and 110 data for plant production obtained from BPS Bengkulu Province from 2000-2022. This research found that the optimal clusters were 3 clusters. C_1 contains 106 data with the largest regression value for chili = 0.127, C_2 contains 15 data with the largest regression value for mustard greens = 0.135, and C_3 contains 110 data with the largest regression value for cabbage = 0.408, eggplant = 0.197, and carrots = 0.201. Furthermore, this research also found that the biggest correlation of crops with highly significant improvement would be cabbage commodity (Y=0.4114X+0.2013) and chili plantation with high RSME (0.9897).

Corresponding Author:

Arie Vatresia
Informatic, Engineering Faculty University Bengkulu
Jl. W. R. Supratman Kandang Limun, Bengkulu, Sumatera, Indonesia. 38371
Email: arie.vatresia@unib.ac.id

1. INTRODUCTION

Indonesia is part of the tropics with high rainfall intensity. This is due to differences in latitude, pseudo-motion of the sun, geographical location, topography, and the interaction of various kinds of air circulation, local, regional, and global [1], [2]. This will have a negative effect on various sectors such as agriculture, fisheries, tourism, and transportation. In agriculture, rainfall as one of the components of the climate plays an important role as a geographical and topographical source of water, greatly affecting the variation of rainfall spatially and temporally [3]–[6]. One of the potential agricultural areas in the Bengkulu region is Rejang Lebong Regency (BPDLD, 2007). The main commodities in the agricultural sector of vegetable crops in Rejang Lebong Regency are cabbage, eggplant, carrots, mustard greens, and large chilies. Based on data published by the Central Statistics Agency of Bengkulu city in 2019, Rejang Lebong was ranked first as a vegetable crop producing district in Bengkulu province[7]. The number of productions in these 5 plants is ranked in the top five in the annual production of vegetable crops in

Rejang Lebong Regency by BPS Bengkulu Province data in 2019. Rainfall affects the process of growth and crop production. Data showed that the production of cabbage plants reached 650,202 tons, eggplant production of 501,400 tons, carrot production of 352,082 tons, mustard plant production of 287,930 tons, and large chili plant production of 276,025 tons (BPS, 2019). Traditional schemes have shown the number before integrating technology that may improve the number of productions. Crop production relies on the value of rainfall. Accurate rainfall predictions [8]–[10] can help farmers make informed decisions about their crop management, leading to higher yields and better-quality produce. Rainfall produces water that is beneficial to plants, where water is where chemical processes and reactions take place. This impacts soil moisture conditions, so the soil can easily absorb water and reduce evaporation[11]–[13]. Predicting rainfall can be a crucial factor in improving crop production. With accurate predictions, farmers can plan their planting and harvesting schedules accordingly, saving resources and maximizing yields. Data Mining is a large-scale data processing method that aims to obtain information from data sets and transform that information into new information structures that can be easily understood[14]–[16]. Data mining can be a valuable tool for improving crop production by analyzing rainfall patterns and identifying trends. By analyzing historical rainfall data [17], farmers can identify patterns in rainfall, such as how much rain falls during a particular season or how often it rains in a given region. This information can help farmers plan their planting schedules, as well as their irrigation and fertilizer applications. Furthermore, farmers can accurately predict future rainfall patterns using advanced data mining techniques. This can help them plan their crops and irrigation systems accordingly, ensuring their crops get the water they need to grow.

Clustering is one of the data mining methods that is unsupervised. Clustering can group similar rainfall patterns, allowing farmers to identify regions or seasons that experience similar weather conditions [18], [19]. This information can be used to plan planting schedules, optimize irrigation systems, and develop pest control strategies. Clustering can also identify unusual weather patterns, such as extended periods of drought or unusually heavy rainfall. By detecting these patterns early, farmers can take appropriate action to protect their crops and minimize losses. This research was conducted in three stages, the cluster pre-analysis stage; the cluster analysis stage; and the cluster validation stage. At the pre-analysis stage of the cluster using the elbow method[20]–[22]. The next stage is cluster analysis using the CLARANS method[23]–[25]. Moreover, for the cluster validation stage using the Silhouette Coefficient method[26]–[29]. The results of the previous cluster process will be used as a regression [30]–[32] value search variable between plants and their rainfall clusters. From the problems found above, it is necessary to conduct a regression analysis as well as rainfall clusters and crop production to provide information in the form of cluster results and find out how much influence the rainfall variable has on the crop production rate in each cluster. Clustering crop production and rainfall is a common research topic in agricultural and environmental sciences. This research aims to understand the relationship between rainfall patterns and crop production rate and identify clusters of crop production based on rainfall patterns. This information can inform crop management strategies, improve agricultural productivity, and mitigate the effects of climate change. Several studies have been conducted to explore the relationship between crop production and rainfall patterns using clustering analysis [33]. The results showed that rainfall patterns and maize yields varied across different regions and that clustering analysis effectively identified patterns in the data. The study found that clustering analysis effectively identified distinct zones with similar rainfall patterns and crop production characteristics [5]. The information from this analysis can be used to develop crop management strategies tailored to specific regions, resulting in improved agricultural productivity and resilience. The results showed that clustering analysis effectively identified crop production clusters related to specific rainfall and soil characteristics. The information gained from this analysis can improve crop management strategies, enhance soil health, and increase regional agricultural productivity. This paper showed a new approach to use regression over clusters made by the model to see how the rainfall contributes to accelerating crop production over the Rejang Lebong regency.

2. METHOD

2.1. Data Collection

The data used in this study is rainfall data that occurred in Rejang Lebong Regency for 21 years (2000-2020) obtained from the Meteorological Climatology Agency, and Geophysics (BMKG) of the One-

Stop Integrated Service in Bengkulu. And for data on the production of 5 plants in Rejang Lebong Regency obtained from the Central Statistics Agency (BPS) of Bengkulu City. The data used in clustering are year, rainfall height, latitude, longitude, and number of cabbage, mustard, eggplant, chili, carrot productions. At the Data Preparation stage, the author will prepare the data for the model training process. This stage consists of normalizing the data and analyzing the data that has been obtained [34]–[36]. The process at the data preparation stage is by normalizing the data for both rainfall and crop production data. The process includes determining the minimum and maximum values of all data, entering values/data for that year, and then subtracting the minimum value. To subtract the maximum value from the minimum value, and to divide the value found in the second step by the third step, then the normalized value is found.

2.2. Location

Rejang Lebong Regency is located at position 102° 19'39; East Longitude to 102° 57'39; Longitude East and 2° 22'39; 07''; South Latitude to 3° 31'39''; South Latitude is bordered by South Sumatra Province in the North and East, with Kepahiang Regency in the South, with Lebong Regency in the North, and with North Bengkulu Regency in the West. The area of Rejang Lebong Regency is around 151,576 ha which is located at an altitude of less than 100 meters to more than 1000 meters [37], [38]. Areas that have elevations of less than 100 meters above sea level are 2,250 hectares (1.48%), which have altitudes between 100 - 1000 meters above sea level, covering an area of 112,669 hectares (74.33%). The rest have elevations above 1000 meters.

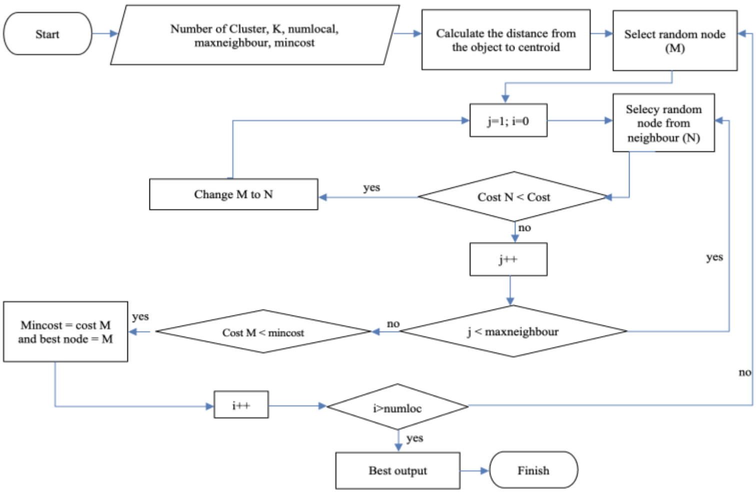
Start
Number of Cluster, K, numlocal, maxneighbour, mincost
Calculate the distance from the object to centroid
Select random node (M)
j=1; i=0
Selecy random node from neighbour (N)
Change M to N
yes
Cost N < Cost
no
j++
yes
Mincost = cost M and best node = M
yes
Cost M < mincost
no
j < maxneighbour
no
i++
i>numloc
yes
Best output
Finish

Figure 1. CLARANS method flowchart

CLARANS has two main parameters, max neighbor and num local. The flow of CLARANS can be seen in Figure 1. Maxneighbor is the maximum number of neighbor nodes examined from a selected node, while num local is the local minimum number which determines how many times a node is checked or how many iterations are performed. The max neighbor value used is $p\%\ x\ k\ (n-k)$ and the num local value used is 2. According to both, the value is between 1.25 and 1.5 where k represents the number of clusters formed, and n represents the number of objects in the data set. These values are chosen to balance runtime (operation time) and the quality of the clusters formed. Suppose there are n objects with p variables in a data set D. Then k clusters (k < n) will be formed, where the value of k is given to calculate the Euclidean distance. This research also calculates the regression that can improve crop production over Rejang Lebong area. Simple Linear Regression is a statistical method that tests the extent of a causal relationship between the Causal Factor Variable (X) and the Consequent Variable. This formula is used to estimate a study with one independent variable (Variable X) and one dependent variable (Variable Y), which is used to determine whether there is an effect of Variable X on Variable Y.

After analyzing using a simple linear regression formula, the next step is to test the hypothesis. This aims to see whether a proposed hypothesis is rejected or accepted. If the results of the null hypothesis test (Ho) are rejected, then the working hypothesis (Ha) is accepted, and vice versa. Hypothesis testing in this study was carried out by means of Test f. The t test serves to test how far the influence of variable X individually (partially) on variable Y is in a study. The way to test the hypothesis by means of the f test is to compare the calculated f value and the f table value, with the provision that if f count> f table, then there is an influence of variable X on variable Y, which means that the Working Hypothesis (Ha) is accepted, and the Null Hypothesis (Ho) rejected. The way to test the hypothesis by means of the f test is to compare the calculated f value to the f table value.

3. RESULT AND DISCUSSION

The research used the Cross Industry Standard Process Model for Data Mining (CRISP-DM) method. Consist of several stages, one of them is the Business Understanding stage, explaining object knowledge, the process of building and obtaining data, and research objectives[39]. The purpose of this data processing is to determine the results of the regression analysis of rainfall and crop production data clusters with the attributes of year, rainfall height, latitude, longitude, cabbage production amount, mustard production amount, eggplant production amount, chili production amount, and carrot production amount. The results of the previous cluster process will be used as a regression value search variable between plants and their cluster's rainfall. This regression value will later be used to analyze the influence or relationship between rainfall and crop production (cabbage, mustard greens, chilies, eggplant, and carrots) to obtain information in the form of the best relationship between bulk variables and plant variables according to the clusters range [40]–[42]. The Data Understanding stage will gained information related to the data used in the research. The modeling stage describes the data process using data mining algorithms to be applied. At the modeling stage, the author uses two methods: (a) Elbow Method, a method to determine the optimal number of clusters, and (b) the CLARANS Method, a method for generating a cluster from rainfall data. The parameters in the CLARANS method are the number of local and maximum neighbor[43]–[45]. The evaluation stage is an interpretation of the output of data mining produced in the previous stage. The silhouette coefficient method serves to determine the quality of each cluster by calculating the average value of the silhouette index for each cluster[27], [29]. Deployment stage in this deployment stage, the information that has been obtained will be implemented into the system and reports in the information section. The results of clustering rainfall data will be displayed via WebGIS. Data processing showed the difference value calculated using the SSE cluster values 2 to 5. A comparison of the SSE values of each cluster is produced which is shown in table 2 below.

In this study, the modeling stage consisted of three stages (pre-cluster analysis, cluster analysis, and regression analysis). In the pre-cluster analysis stage, the authors used the elbow method; for the cluster analysis stage the authors used the CLARANS method; and for the regression analysis stage, the authors used a simple linear regression method. Optimal clusters number was defined in the Elbow method as seen in Figure 2.

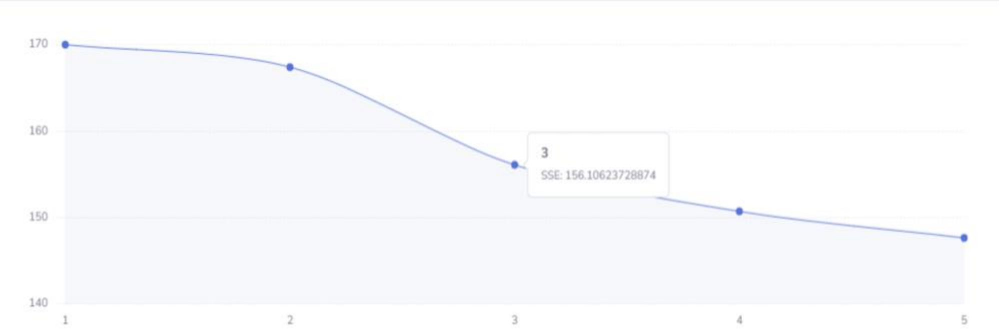

Figure 2. Elbow Result for Optimum Clusters

The determination of clusters number was provide based on SSE (Sum of Squares Error) values that showed the most significant error was 11.2866 lies in cluster 3. Furthermore, the result showed that the Cluster 1 silhouette coefficient value was 0.5899 with a Mean Square Error (MSE) value of 0.08576; Cluster 2 silhouette coefficient value was 0.6019 with a Mean Square Error (MSE) value of 0.0313; and Cluster 3 silhouette error coefficient value showed 0.5772 with MSE at 0.0906. The overall silhouette coefficient value was at 0.5897 with a Mean Square Error (MSE) of 0.06924. The three clusters range each variable, as shown in Table 1.

Table 1. Clustering Result From CLARANS

	C1	C2	C3
Latitude	(-3.63) – (-3.23)	(-3.63) – (-3.23)	(-3.63) – (-3.23)
Longitude	(102.44) – (102.8)	(102.44) – (102.8)	(102.44) – (102.8)

Rainfall	(0.0095)-(1)	(0.2583) – (0.7299)	(0) – (0.8279)
Cabbage	(0.1573) – (0.9825)	(0.9825) – (1)	(0) – (0.3818)
Mustard	(0) – (0.7691)	(0.3131) – (0.3374)	(0) – (1)
Eggplant	(0.1247) – (1)	(0.6687) – (1)	(0) – (0.40345)
Chili	(0.3165) – (0.7143)	(0.318063) – (1)	(0) – (0.617791)
Carrots	(0.3433) – (1)	(0.739134) – (1)	(0) – (0.418881)
Years	(2002) – (2020)	(2013) – (2014)	(2000) – (2019)
Data	106	15	110

Decision trees can help identify the most important features for clustering by identifying which variables are most important for separating the clusters. This can help select the relevant variables for clustering and exclude the less relevant ones. It also can help determine cluster boundaries by identifying the splits that provide the best separation between clusters. This can help to define the clusters better and improve the accuracy of the analysis. The cluster was then interpreted into the narration of the characters using a decision tree, and found that the root of the data was based on the time of crop produce in Rejang Lebong district (Figure 3).

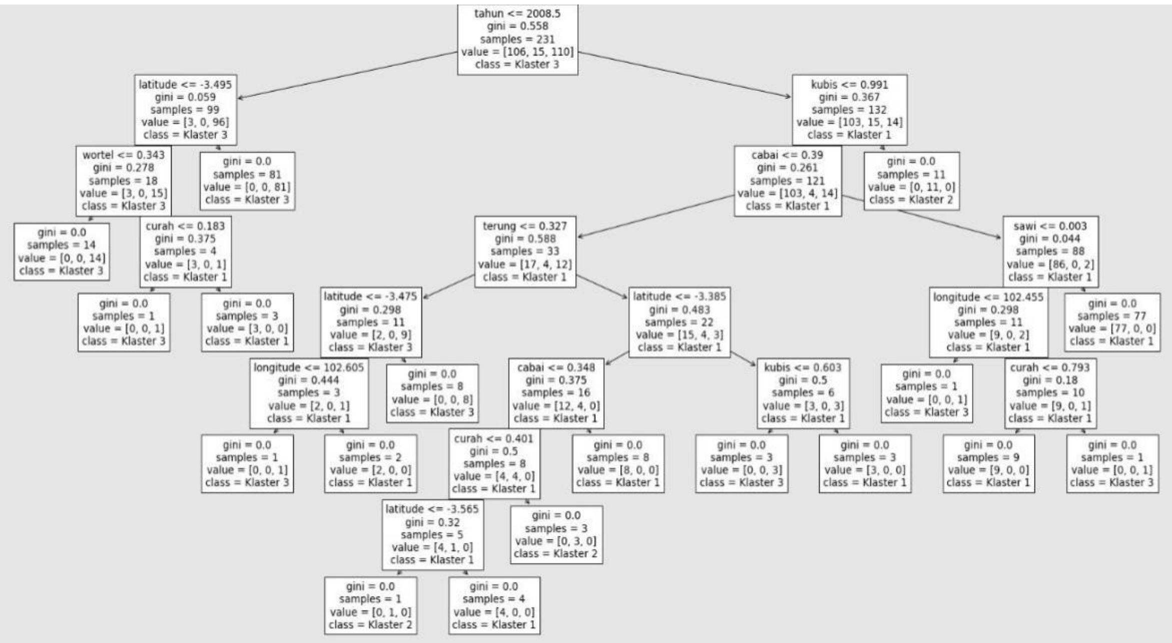
tahun <= 2008.5 gini = 0.558 samples = 231 value = [106, 15, 110] class = Klaster 3
latitude <= -3.495 gini = 0.059 samples = 99 value = [3, 0, 96] class = Klaster 3
kubis <= 0.991 gini = 0.367 samples = 132 value = [103, 15, 14] class = Klaster 1
wortel <= 0.343 gini = 0.278 samples = 18 value = [3, 0, 15] class = Klaster 3
gini = 0.0 samples = 81 value = [0, 0, 81] class = Klaster 3
cabai <= 0.39 gini = 0.261 samples = 121 value = [103, 4, 14] class = Klaster 1
gini = 0.0 samples = 11 value = [0, 11, 0] class = Klaster 2
gini = 0.0 samples = 14 value = [0, 0, 14] class = Klaster 3
curah <= 0.183 gini = 0.375 samples = 4 value = [3, 0, 1] class = Klaster 1
terung <= 0.327 gini = 0.588 samples = 33 value = [17, 4, 12] class = Klaster 1
sawi <= 0.003 gini = 0.044 samples = 88 value = [86, 0, 2] class = Klaster 1
gini = 0.0 samples = 1 value = [0, 0, 1] class = Klaster 3
gini = 0.0 samples = 3 value = [3, 0, 0] class = Klaster 1
latitude <= -3.475 gini = 0.298 samples = 11 value = [2, 0, 9] class = Klaster 3
latitude <= -3.385 gini = 0.483 samples = 22 value = [15, 4, 3] class = Klaster 1
longitude <= 102.455 gini = 0.298 samples = 11 value = [9, 0, 2] class = Klaster 1
gini = 0.0 samples = 77 value = [77, 0, 0] class = Klaster 1
longitude <= 102.605 gini = 0.444 samples = 3 value = [2, 0, 1] class = Klaster 1
gini = 0.0 samples = 8 value = [0, 0, 8] class = Klaster 3
cabai <= 0.348 gini = 0.375 samples = 16 value = [12, 4, 0] class = Klaster 1
kubis <= 0.603 gini = 0.5 samples = 6 value = [3, 0, 3] class = Klaster 1
gini = 0.0 samples = 1 value = [0, 0, 1] class = Klaster 3
curah <= 0.793 gini = 0.18 samples = 10 value = [9, 0, 1] class = Klaster 1
gini = 0.0 samples = 1 value = [0, 0, 1] class = Klaster 3
gini = 0.0 samples = 2 value = [2, 0, 0] class = Klaster 1
curah <= 0.401 gini = 0.5 samples = 8 value = [4, 4, 0] class = Klaster 1
gini = 0.0 samples = 8 value = [8, 0, 0] class = Klaster 1
gini = 0.0 samples = 3 value = [0, 0, 3] class = Klaster 3
gini = 0.0 samples = 3 value = [3, 0, 0] class = Klaster 1
gini = 0.0 samples = 9 value = [9, 0, 0] class = Klaster 1
gini = 0.0 samples = 1 value = [0, 0, 1] class = Klaster 3
latitude <= -3.565 gini = 0.32 samples = 5 value = [4, 1, 0] class = Klaster 1
gini = 0.0 samples = 3 value = [0, 3, 0] class = Klaster 2
gini = 0.0 samples = 1 value = [0, 1, 0] class = Klaster 2
gini = 0.0 samples = 4 value = [4, 0, 0] class = Klaster 1

Figure 3. Root of the data was based on the year of crop produce in Rejang Lebong district

1. Cluster 1 consists of 106 data that have characteristics:
    a. 3 data were clustered based on latitude values <= -3,495 with a gini of 0.59, carrot production >= 0.343 with a gini of 0.278, and rainfall magnitude of >= 0.183 with a gini of 0.375.
    b. 2 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.39 with gini 0.261, eggplant production <= 0.327 with gini 0.558, latitude <= -3.475 with gini 0.298, and longitude >= 102.605 with gini 0.444.
    c. 8 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production >= 0.348 with gini 0.375, eggplant production >= 0.327 with gini 0.588, and latitude <= -3.385 with gini 0.483.
    d. 4 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.348 with gini 0.375, eggplant production >= 0.327 with gini 0.558, rainfall amount <= 0.401 with gini 0.5, and latitude >= -3.565 with gini 0.32.
    e. 3 data were classified based on cabbage production >= 0.603 with gini 0.5, chili production <= 0.39 with gini 0.261, eggplant production >= 0.327 with gini 0.588, and latitude >= -3.385 with gini 0.483.
    f. 77 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production >= 0.39 with gini 0.261, and mustard production >= 0.3 with gini 0.4.
    g. 9 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production >= 0.39 with gini 0.261, mustard production <= 0.3 with gini 0.4, longitude >= 102.455 with gini 0.298, rainfall magnitude <= 0.793 with gini 0.18.
2. Cluster 2 consist of 15 data that have characteristics:

   a. 3 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.39 with gini 0.261, eggplant production >= 0.327 with gini 0.558, latitude <= -3.385 with gini 0.483, chili production <= 0.348 with gini 0.375, rainfall amount >= 0.401 with gini 0.5.
   b. 1 data data is classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.39 with gini 0.261, eggplant production >= 0.327 with gini 0.558, latitude <= -3.565 with gini 0.32, chili production <= 0.348 with gini 0.375, rainfall amount <= 0.401 with gini 0.5.
   c. 11 data clustered based on cabbage production >= 0.0991 with gini 0.367.
3. Cluster 2 consist of 110 data that have characteristics:
   a. 14 data were clustered based on latitude <= -3,495 with a gini of 0.59. and carrot production <= 0.343 with a gini of 0.278.
   b. 1 data was clustered based on latitude <= -3,495 with a gini of 0.59, carrot production >= 0.343 with a gini of 0.278, and rainfall magnitude of <= 0.183 with a gini of 0.375.
   c. 81 data were clustered based on latitude >= -3,495 with a gini of 0.59.
   d. 8 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.39 with gini 0.261, eggplant production <= 0.327 with gini 0.588, and latitude >= -3.475 with gini 0.298.
   e. 1 data is classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.39 with gini 0.261, eggplant production <= 0.327 with gini 0.588, and latitude <= -3.475 with gini 0.298, and longitude <= 102.605 with gini 0.444.
   f. 3 data were classified based on cabbage production <= 0.991 with gini 0.367, chili production <= 0.39 with gini 0.261, eggplant production >= 0.327 with gini 0.588, latitude >= -3.385 with gini 0.483, and cabbage production <= 0.603 with gini 0.5.
   g. 1 data is classified based on cabbage production <= 0.991 with gini 0.367, chili production >= 0.39 with gini 0.261, mustard production <= 0.3 with gini 0.4, and longitude <= 102.455 with gini 0.298.
   h. 1 data is classified based on cabbage production <= 0.991 with gini 0.367, chili production >= 0.39 with gini 0.261, mustard production <= 0.3 with gini 0.4, longitude >= 102.455 with gini 0.298, and rainfall amount >= 0.793 with gini 0.18.

4. DISCUSSION

Three clusters were formed after grouping all rainfall production data with plants over Rejang Lebong Regency from 2000 to 2020. Based on the clustering results that have been found, the data will be processed using simple linear regression, wherein the simple linear regression method with one independent variable (X) and one dependent variable (Y). Furthermore, the cluster validation stage, in this study used the silhouette coefficient method[26], [27]. Cluster one coefficient silhouette value = 0.5899 (feasible or compliant cluster), cluster two coefficient silhouette value = 0.6019 (feasible or corresponding cluster), and the silhouette coefficient value of cluster three = 0.5772 (feasible or compliant cluster), and for the value silhouette coefficient as a whole = 0.5897 (feasible or compliant cluster). After the cluster process was carried out for all rainfall production data with crops in Rejang Lebong Regency from 2000 to 2020, 3 clustering were formed with the amount of data on cluster one, consist of 106 data; cluster two consist of 15 data; and cluster 3 consist of 110 data (Table 2). Based on the clustering results that have been found, the data will be processed using simple linear regression, where in this simple linear regression method using one free variable (X) and one bound variable (Y)[30], [32], [46]. Then the following results are obtained in Table 3.

Table 2. Value of Coefficient of Determination (R^2)

	Cluster 1	Cluster 2	Cluster 3
Cabbage	0.0372	0.0158	0.8581
Mustard	0.0009	0.1359	0.0944
Eggplant	0.0031	0.0441	0.1975
Chili	0.1276	0.0151	0.0886
Carrot	0.0817	0.0434	0.2012

The greatest regression relationship in cabbage plants is in cluster 3 with a coefficient of determination value (R^2) of 0.4080 in the rainfall range of 0-0.8279, cabbage production amount of 0-0.381781, and is between 2000-2019. Mustard plants are in cluster 2 with a coefficient of determination value (R^2) of 0.135921554 in the rainfall range of 0.258331-0.7298923, mustard production is 0.31305-0.337411, and is between 2013-2014. In eggplant plants, it is in cluster 3 with a coefficient of determination value (R Square) of 0.1975 in the rainfall range of 0-0.8279, the amount of eggplant production is 0- 0.4035, and is between 2000-2019. Furthermore, chili plants lied in cluster 1 with a coefficient of determination value (R^2) of 0.1276 in the rainfall range of 0.0095-1, the total chili production is 0.3165- 0.7143, and is between 2002-2020. And in carrot crops are in cluster 3 with a coefficient of determination value (R^2) 0.2012 in the rainfall range of 0-0.8279, the amount of carrot production is 0- 0.4189, and is between 2000-2019. The most large and influential linear regression relationship between rainfall and five plants is in cluster 3, with the total value of the overall coefficient of determination of 0.9897 (chili plantation).

Table 3. Rainfall and Crop Regression Equation

Cluster 1	
Cabbage	*Y=0.1712+0.08532X*
Mustard	*Y=0.0042+0.0113X*
Eggplant	*Y=0.0932+0.01099X*
Chili	**Y=0.5963+0.2682X**
Carrot	*Y=0.2121+0.0833X*
Cluster 2	
Cabbage	*Y=0.1013+0.03114X*
Mustard	***Y=0.0374+0.3319X***
Eggplant	*Y=0.1319+0.01279X*
Chili	*Y=0.1213+0.0982X*
Carrot	*Y=0.0942+0.0533X*
Cluster 3	
Cabbage	**Y=0.2013+0.4114X**
Mustard	Y=0.0087+0.03184X
Eggplant	Y=0.4369+0.0428X
Chili	Y=0.2962+0.1382X
Carrot	Y=0.5001+0.3328X

The simple regression model formed in Table 4 contains a positive regression constant value (X) for every 1% increase in rainfall, it can increase the amount of crop production by Y times. This research is research that produces clusters and regressions for rainfall with plants in Rejang Lebong district. Furthermore, from the clustering process using the CLARANS method, tree rainfall groupings with plants were obtained, and the simple linear regression method is the method used to measure the relationship between rainfall and plants, whereas in previous studies used a prediction with genetic algorithms and the result is the level of rainfall as one of the parameters determining plant species with certainty factor[47], [48]. This study has successfully occupied 21 year long data series to see perform the regression. The previous research on clustering was only one year and only based on the nature of rain [49], [50]. Clustering results are validated and MSE calculations are carried out to ensure the error rate with Silhouette Coefficient validation, whereas from previous studies several studies did not use the validation method[51], [52].

The results of the study provide a conclusion that for the optimal K value or number of clusters is 3 seen from the difference in the largest sum of square error (SSE) value in cluster 3 with a difference in value of 11.286642. The amount of data in cluster 1 is 106 data, the amount of data in cluster 2 is 15 data, and the amount of data in cluster 3 is 110 data. The greatest regression relationship in cabbage plants is in cluster 3 with a coefficient of determination value (R Square) of 0.408013795 in the rainfall range of 0-0.827863, cabbage production amount of 0- 0.381781, and is between 2000-2019. Mustard plants are in cluster 2 with a coefficient of determination value (R Square) of 0.135921554 in the rainfall range of 0.258331-0.7298923, mustard production is 0.31305-0.337411, and is between 2013-2014. In eggplant plants, it is in cluster 3 with a coefficient of determination value (R Square) of 0.197528516 in the rainfall range of 0-0.827863, the amount of eggplant production is 0-0.40345, and is between 2000-

2019. In chili plants, it is in cluster 1 with a coefficient of determination value (R Square) of 0.127581145 in the rainfall range of 0.00952-1, the total chili production is 0.316471-0.714344, and is between 2002-2020. And in carrot crops are in cluster 3 with a coefficient of determination value (R Square) 0.201155964 in the rainfall range of 0-0.827863, the amount of carrot production is 0-0.418881, and is between 2000-2019. Finally, the most large and influential linear regression relationship between rainfall and 5 plants (cabbage, mustard, eggplant, chili, and carrots) were involved in cluster 3 with the total value of the overall coefficient of determination of 0.989688. And global cluster validation value = 0.58964931822801 with MSE = 0.0692422 and belongs to a viable or compliant cluster.

5. CONCLUSION

This research showed there are tree clusters as the best fit with CLARANS and elbow method to describe two decades of data of crop plantation over Rejang Lebong district. The clusters developed have also been defined by the rule of decision tree that showed the group of significant variables to improve data understanding and description. This research showed a positive regression equation for each of the plantations involved in this research. Each cluster developed has different significant values for each plantation developed by the regression analysis. Although all of the plantations are showing positive, each cluster has a different kind of plantation as the best regression development. In cluster 1, we can see that the best plantation was Chili, in cluster 2 was Mustard, and in cluster 3 was cabbage. The government can use the result of Rejang Lebong district to develop strategic planning of smart agriculture to accelerate crop production to make food security in Bengkulu area become more reliable and stable.

REFERENCES

[1] T. T. H. Tambunan, Perkembangan Sektor Pertanian di Indonesia, Cet. 1. Jakarta : Ghalia Indonesia, 2003. [Online]. Available: http://agris.fao.org/agris-search/search.do?recordID=US201300101636

[2] C. Kubitza, V. V Krishna, K. Urban, Z. Alamsyah, and M. Qaim, "Land Property Rights, Agricultural Intensification, and Deforestation in Indonesia," Ecological Economics, vol. 147, pp. 312–321, 2018, doi: https://doi.org/10.1016/j.ecolecon.2018.01.021.

[3] TNA, "Indonesia Technology Needs Assessment for Climate Change Mitigation," UNEP on behalf of Global Environmental Facility (GEF), 2012.

[4] H. S. Lee, "General Rainfall Patterns in Indonesia and the Potential Impacts of Local Seas on Rainfall Intensity," Water (Switzerland), vol. 7, no. 4, 2015, doi: 10.3390/w7041751.

[5] R. D'Arrigo and R. Wilson, "El Niño and Indian Ocean influences on Indonesian drought: Implications for forecasting rainfall and crop productivity," International Journal of Climatology, vol. 28, no. 5, 2008, doi: 10.1002/joc.1654.

[6] Supari, F. Tangang, E. Salimun, E. Aldrian, A. Sopaheluwakan, and L. Juneng, "ENSO modulation of seasonal rainfall and extremes in Indonesia," Clim Dyn, vol. 51, no. 7–8, 2018, doi: 10.1007/s00382-017-4028-8.

[7] Badan Pusat Statistika, "Statistik Perumahan Dan Permukiman 2019," Katalog BPS, 2019.

[8] N. S. Sani, A. H. A. Rahman, A. Adam, I. Shlash, and M. Aliff, "Ensemble Learning for Rainfall Prediction," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111120.

[9] G. B. Sai Tarun, J. V. Sriram, K. Sairam, K. T. Sreenivas, and M. V. B. T. Santhi, "Rainfall prediction using machine learning techniques," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7, 2019.

[10] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, "Rainfall prediction using data mining techniques: A systematic literature review," International Journal of Advanced Computer Science and Applications, vol. 9, no. 5. 2018. doi: 10.14569/IJACSA.2018.090518.

[11] W. H. H. Wischmeier and D. D. D. Smith, "Predicting rainfall erosion losses," Agriculture handbook no. 537, no. 537, pp. 285–291, 1978, doi: 10.1029/TR039i002p00285.

[12] A. Kurniadi, E. Weller, S. K. Min, and M. G. Seong, "Independent ENSO and IOD impacts on rainfall extremes over Indonesia," International Journal of Climatology, vol. 41, no. 6, 2021, doi: 10.1002/joc.7040.

[13] Supriyono, F. Wira Citra, B. Sulistyo, and M. Faiz Barchia, "Mapping Erosivity Rain And Spatial Distribution Of Rainfall In Catchment Area Bengkulu River Watershed," Journal of Environment and Earth Science, vol. 7, no. 10, 2017.

[14] M. Wang, A. Wang, and A. Li, "Mining spatial-temporal clusters from geo-databases," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006. doi: 10.1007/11811305_29.

[15] M. Bertolotto, S. Di Martino, F. Ferrucci, and T. Kechadi, “Towards a framework for mining and analyzing spatio-temporal datasets,” International Journal of Geographical Information Science, vol. 21, no. 8, 2007, doi: 10.1080/13658810701349052.

[16] G. Atluri, A. Karpatne, and V. Kumar, “Spatio-temporal data mining: A survey of problems and methods,” ACM Computing Surveys, vol. 51, no. 4. 2018. doi: 10.1145/3161602.

[17] M. S. M. Ariff, N. M., Bakar, M. A. A., Mahbar, S. F. S., & Nadzir, “Clustering Of Rainfall Distribution Patterns Using Time Series Clustering Method,” Malaysian Journal of Science, vol. 38, no. Sp2, 2019.

[18] V. Tobar and G. Wyseure, “Seasonal rainfall patterns classification, relationship to ENSO and rainfall trends in Ecuador,” International Journal of Climatology, vol. 38, no. 4, 2018, doi: 10.1002/joc.5297.

[19] S. M. C. M. Nor, S. M. Shaharudin, S. Ismail, S. A. M. Najib, M. L. Tan, and N. Ahmad, “Statistical Modeling of RPCA-FCM in Spatiotemporal Rainfall Patterns Recognition,” Atmosphere (Basel), vol. 13, no. 1, 2022, doi: 10.3390/atmos13010145.

[20] F. Liu and Y. Deng, “Determine the Number of Unknown Targets in Open World Based on Elbow Method,” IEEE Transactions on Fuzzy Systems, vol. 29, no. 5, 2021, doi: 10.1109/TFUZZ.2020.2966182.

[21] B. Purnima, K. Arvind, P. Bholowalia, and A. Kumar, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” Int J Comput Appl, vol. 105, no. 9, 2014.

[22] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” EURASIP J Wirel Commun Netw, vol. 2021, no. 1, 2021, doi: 10.1186/s13638-021-01910-w.

[23] V. Sagvekar, V. Sagvekar, and K. Deorukhkar, “Performance assessment of CLARANS: A Method for Clustering Objects for Spatial Data Mining,” Global Journal of Engineering, Design & Technology/Global Institute for Research & Education, vol. 2, no. 6, 2013.

[24] A. Azizah, R. Cahyandari, A. F. Huda, Sukono, Subiyanto, and A. T. Bon, “Application of spatial weighting matrix of GSTAR by using CLARANS clustering on farmer exchange rates in 32 provinces in Indonesia,” in Proceedings of the International Conference on Industrial Engineering and Operations Management, 2019.

[25] R. T. Ng and J. Han, “CLARANS: A method for clustering objects for spatial data mining,” IEEE Trans Knowl Data Eng, vol. 14, no. 5, 2002, doi: 10.1109/TKDE.2002.1033770.

[26] M. B. Al-Zoubi and M. Al Rawi, “An efficient approach for computing silhouette coefficients,” Journal of Computer Science, vol. 4, no. 3, 2008, doi: 10.3844/jcssp.2008.252.255.

[27] H. Řezanková, “Different approaches to the silhouette coefficient calculation in cluster evaluation,” 21st International Scientific Conference AMSE, no. September 2018.

[28] R. Hidayati, A. Zubair, A. Hidayat Pratama, L. Indana, P. Studi Sistem Informasi, and F. Teknologi Informasi, “Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering,” Techno.Com, vol. 20, no. 2, 2021.

[29] R. D. Jujjuri and M. Venkateswara Rao, “Evaluation of enhanced subspace clustering validity using silhouette coefficient internal measure,” Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 1, 2019.

[30] D. Bera, N. Das Chatterjee, and S. Bera, “Comparative performance of linear regression, polynomial regression and generalized additive model for canopy cover estimation in the dry deciduous forest of West Bengal, Remote Sensing Applications: Society and Environment,” vol. 22, p. 100502, Dec. 2021.

[31] Y. W. Park and D. Klabjan, “Subset selection for multiple linear regression via optimization,” Journal of Global Optimization, vol. 77, no. 3, 2020, doi: 10.1007/s10898-020-00876-1.

[32] B. Dhaval and A. Deshpande, “Short-term load forecasting with using multiple linear regression,” International Journal of Electrical and Computer Engineering, vol. 10, no. 4, 2020, doi: 10.11591/ijece.v10i4.pp3911-3917.

[33] B. Zerouali, M. Chettih, Z. Abda, M. Mesbah, C. A. G. Santos, and R. M. Brasil Neto, “A new regionalization of rainfall patterns based on wavelet transform information and hierarchical cluster analysis in northeastern Algeria,” Theor Appl Climatol, vol. 147, no. 3–4, 2022, doi: 10.1007/s00704-021-03883-8.

[34] W. Y. Ayele, “Adapting CRISP-DM for Idea Mining,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 6, pp. 20–32, 2020.

[35] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, no. 24959, 2000.

[36] C. Schr�er, F. Kruse, and J. M. G�mez, “A systematic literature review on applying CRISP-DM process model, Procedia Computer Science,” vol. 181, pp. 526–534, 2021.

[37] BNPB, “Infografis Bencana Banjir dan Longsor Bengkulu,” 2023. https://bnpb.go.id/infografis/infografis-bencana-banjir-dan-longsor-bengkulu (accessed Feb. 14, 2023).

[38] S. Supriyono, S. Utaya, D. Taryana, and B. Handoyo, “Spatial-Temporal Trend Analysis of Rainfall Erosivity and Erosivity Density of Tropical Area in Air Bengkulu Watershed, Indonesia, Quaestiones Geographicae,” vol. 40, no. 3, pp. 125–142, 2021.

[39] C. Shearer et al., “The CRISP-DM model: The New Blueprint for Data Mining,” Journal of Data Warehousing, 2000.

[40] J. Wu, Advances in K-means Clustering: a data mining thinking. 2012.

[41] E. Biabiany, D. C. Bernard, V. Page, and H. Paugam-Moisy, “Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles,” Comput Geosci, vol. 145, 2020, doi: 10.1016/j.cageo.2020.104612.

[42] M. Senožetnik, L. Bradeško, B. Kažič, D. Mladeni, and T. Šubic, “Spatio-temporal clustering methods,” http://optimumproject.eu/news/44/67/Spatio-temporal-Clustering-Methods.html, 2016.

[43] H. F. Tork, “Spatio-Temporal Clustering Methods Classification,” Doctoral Symposium on Informatics Engineering (DSIE’2012), no. December 2012.

[44] V. V. D. M. S. Takalikar, “Survey on Spatio-Temporal Clustering,” International Journal of Science and Research (IJSR), vol. 5, no. 7, 2016.

[45] Y. Ren, N. Wang, M. Li, and Z. Xu, “Deep density-based image clustering,” Knowl Based Syst, vol. 197, 2020, doi: 10.1016/j.knosys.2020.105841.

[46] S. Rath, A. Tripathy, and A. R. Tripathy, “Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model,” Diabetes and Metabolic Syndrome: Clinical Research and Reviews, vol. 14, no. 5, 2020, doi: 10.1016/j.dsx.2020.07.045.

[47] G. Mulyasari, “KAJIAN KETAHANAN PANGAN DAN KERAWANAN PANGAN DI PROVINSI BENGKULU,” Jurnal AGRISEP, vol. 15, no. 1, 2016, doi: 10.31186/jagrisep.15.1.83-90.

[48] A. Sutoyo, “Implementasi Program Aksi Ketahanan Pangan Di Propinsi Bengkulu,” Jurnal Administrasi Publik, vol. 11, no. 1, 2013.

[49] G. Su, “Analysis of optimization method for online education data mining based on big data assessment technology,” Int J Contin Eng Educ Life Long Learn, vol. 29, no. 4, 2019, doi: 10.1504/IJCEELL.2019.102768.

[50] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan, “Identifying patterns in spatial information: A survey of methods,” Wiley Interdiscip Rev Data Min Knowl Discov, vol. 1, no. 3, pp. 193–214, 2011, doi: 10.1002/widm.25.

[51] C. Fischer et al., “Mining Big Data in Education: Affordances and Challenges,” Review of Research in Education, vol. 44, no. 1, 2020, doi: 10.3102/0091732X20903304.

[52] 김정희, “Exploratory Analysis and Visualization of Spatio-Temporal Data Using Data Mining,” Journal of the Association of Korean Photo-Geographers, vol. 29, no. 4, 2019, doi: 10.35149/jakpg.2019.29.4.011.