ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 9, Issue 11, November 2022

A Critical Study on the Usefulness of Data Mining Techniques and Advanced Computational Algorithm in Fertilizer Recommendation and Corp Suitability Ensuring Corp Yield Increase

[1] Vaishali Jain

[1] Student, Computer Science, Vellore Institute of Technology, Bhopal, India.
Corresponding Author Email: [1] vaishalijain3456@gmail.com

***Abstract*— There are a number of factors that may be used to forecast the yield of crops, including rainfall and temperature as well as the use of fertilizers and pesticides. Agribusiness is a hotbed for the use of data mining methods. In order to predict crop yields for next year, data mining methods are put to use in agriculture. Farmers and agribusinesses in the agricultural industry face a dizzying array of choices every day, and these decisions are influenced by a wide range of circumstances. It is critical for agricultural planning to accurately estimate the yields of the several crops that will be considered. In order to come up with realistic and efficient answers to this challenge, data mining methods are essential.**

Prior studies have highlighted that Big Data has been a natural fit for the agriculture sector which can increase the productivity rate of the crop. On the other hand, Farmers are increasingly relying on information and assistance to make crucial agricultural choices because of environmental circumstances, soil variability, input quantities, combinations, and commodity pricing. In addition, this research study has utilized data mining methods such as PAM, CLARA, DBSCAN, and Multiple Linear Regression for maximizing harvest yields. A collaborative system of agricultural output prediction, forecasting, and fertilizer recommendation is suggested in the research study. K Nearest Neighbour is used to the agricultural dataset in this study in order to recommend the most acceptable crops. Predicting and projecting crop yields will lead to increased agricultural productivity. Farmer-friendly fertilization decisions are supported by crop rotation, which improves soil fertility.

Keywords: Farmers, Data mining, Agriculture sector, Crops, Yield, Algorithm, Advanced Computational Algorithm and Fertilizer.

I. INTRODUCTION

Presently in the domain of technology data mining is a process incorporated by the companies for turning the raw information or data into essential information for enhancing the performance of an organization. However, large data sets may be analyzed using software to find trends that might help companies better understand their consumers, boost sales, and save expenses. However, in this research study the researcher has tried to detect the significance of data mining techniques and advanced computational algorithms in order to improve the fertilization and crop suitability ensuring crop yield increase [2]. However, according to prior studies it has been determined that the inception of data mining has a significant impact in increasing the efficiency of the agricultural sector. As the inclusion of data mining and advanced computational algorithms can assist the process of crop disease prediction, pest id and detection and classification, crop yield prediction and input management have all been made possible by the application of Data Mining methods in the agricultural area. In addition to that, the inception of data mining and advanced computational algorithms is also highly effective for recommending effective fertilizer to increase the productivity of agricultural sectors. However, the further part of this research study has determined to ensure that the yield of a firm is increased via the use of data mining techniques and advanced computational algorithms in fertilizer recommendation and suitability.

Aim:

The purpose of this study is to investigate the significance of data mining techniques and advanced computational algorithms in fertilizer recommendation and increasing the yield of a firm.

II. REVIEW OF LITERATURE

Significance of Data mining technique in fertilizer recommendation and crop suitability ensuring crop yield increase

Data mining methods can be used to estimate agricultural yields based on the input parameters of average rainfall and field size. The crop production can be predicted with the use of recent advances in information technology in the agricultural sector [3]. In a growing country such as India, agriculture serves as the primary source of income for the vast majority of its citizens. Agricultural data mining is a key area of informatics that may be used effectively. Data mining tools may help agricultural stakeholders anticipate and forecast crop production. Crop yields are reduced to a bare minimum when farmers are ignorant of soil nutrition and

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Vol 9, Issue 11, November 2022

composition. In order to deliver the most appropriate crop recommendations, a method has been created that focuses on macro nutrients (NPK) available in soil [4]. Currently, there is no integrated system for agricultural output prediction, forecasting, and fertilizer recommendation. In addition to that, data mining and analysis of fresh, non-experimental data from vast amounts of existing crop, soil, and climate information maximizes productivity while also strengthening agriculture's resistance to climate change. Apart from that, a well maintained irrigation system is essential for farmers who want to maximize their agricultural yields per acre. A plant's growth and yield are closely related to the quantity of water it receives from the soil.

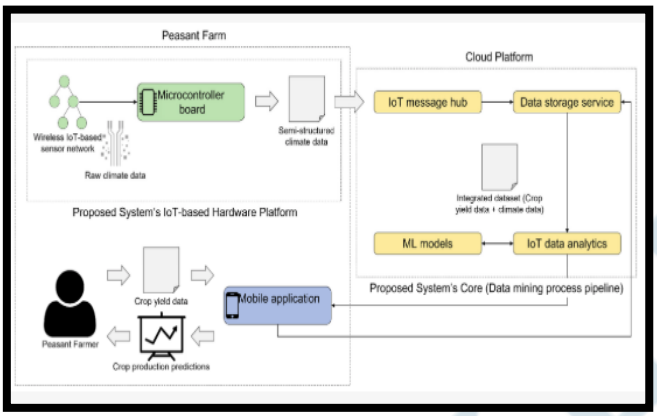

Figure 1: Benefit of Data mining in Agricultural sector (Source: [4])

Today, India is the world's second-largest producer of agricultural goods. Agriculture is India's most populous economic sector and an essential part of the country's cultural structures. Agricultural crop production is unusual in that it is highly reliant on a wide range of climatic and economic conditions. Terrain, temperature, cultivation, irrigation, fertilizers, climate, rainfall, harvest, pesticide weeds, and other elements all have a role in agriculture [1]. A company's supply chain operations are also dependent on historical crop production statistics. All of these businesses rely on agricultural goods as raw materials for anything from livestock to food to animal feed to chemicals to insecticides. In order to manage supply chain decisions including production scheduling, these organizations benefit from an accurate assessment of crop output and risk. Crop production projections guide the production and marketing strategies of companies such as seed, fertilizer, agrochemicals, and agricultural equipment. Once it comes to data mining, algorithms and predetermined rules are used to make it work intelligently. In order to understand patterns in the previous data, it employs a set of predefined rules and algorithms, which it then applies to the current work. Agricultural data mining is a key area of informatics that may be used effectively [5]. In agricultural situations, the use of big data and high-performance computing has opened up new avenues for unravelling the quantification and comprehensiveness of the information heavy process. Throughout the whole growing and harvesting process, data is being mined.

On this note, in order to improvise the process of examination of data, data mining technique plays an essential role. In the examination of data, data mining is essential. Data mining is a computational technique that uses approaches from the areas of AI, ML, analytics, and database systems to find patterns in big data sets. Data mining employs unsupervised and supervised techniques of learning. Since data points in a single cluster are close together, then data points in distinct clusters should be separated by substantial distances. Cluster analysis is a statistical technique that splits large amounts of data into smaller, more manageable chunks. Apart from that, crop production records and forecasts are provided to farmers to assist they better manage their risk [1]. In addition, it aids in the formulation of crop insurances and supply chain management plans by the government. Therefore, it is fair to anticipate that the inception of data mining is effective in enhancing the performance and productivity of the agricultural sector.

Significance of Advanced Computational Algorithm technique in fertilizer recommendation and crop suitability ensuring crop yield increase

There are several ways in which machine learning and artificial neural networks can help the farmers and agricultural sectors to increase productivity of the crops. However, the usage of advanced computation algorithms can lead to the creation of smart farming technologies that can be used by peasant farmers to assist them in making decisions [6]. In addition to that, incorporating low-cost IoT sensors and popular cloud-based data storage and analytics platforms has also facilitated the process of smart farming systems for crop production. On the other hand, advanced computational algorithm approach can help to utilize climatic data crop production data to facilitate the forecasting process of production volume from heterogeneous data sources. Apart from that, Machine learning (ML) and statistical analysis are at the heart of data mining, which is a branch of artificial intelligence [1]. It is feasible to tackle issues involving prediction, classification, and segmentation by applying models derived via the implementation of AI and statistical analytic approaches. Predictive analytics is a method for gleaning information from enormous datasets in order to make accurate predictions about what will happen in the future. It is a stepping stone in the larger data analysis procedure known as business analytics. Data analysis methods that automate the creation of analytical models might be referred to as machine learning in this context.

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Vol 9, Issue 11, November 2022

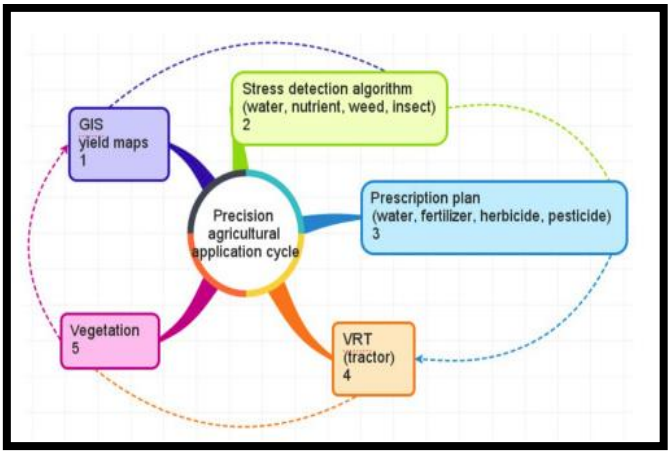

Figure 2: Benefit of Advanced Computation Algorithm in Agricultural sector
(Source: [6])

Recently, the agriculture sector has seen a significant increase in the use of predictive analytics and the Internet of Things (IoT). As a result, ideas including "smart farming" and "precise farming" have been born. Interconnecting virtual and physical "things" through Internet of Things (IoT) technology is a key component of the information society's global infrastructure. IoT technologies include a wide range of prevalent and developing ICTs, which may increase the crop. According to prior researchers, the incorporation of advanced computation can assist the farmers to monitor the health of crops in real time, anticipate future yields and make resource management choices based on established patterns [7]. On this note, several farmers and agricultural sectors are utilizing the clustering method in order to improve the productivity of crops. It is possible to use a variety of clustering algorithms for a variety of reasons. Methods for categorizing data into groups may be broken down into a number of subcategories, including Partitioning, Hierarchical, Density, Grid, and Model-based clustering approaches [1]. In order to increase the quality of clustering findings, partitioning algorithms like K-means, K-medoids PAM and CLARA and CLARANS distribute items to k clusters and repeatedly reallocate objects. In hierarchical clustering methods, items are assigned to tree-structured clusters, although a cluster might include data points that reflect lower-level clusters. Density-based clustering algorithms assume that the vicinity of a given unit distance has a certain number of points for each point in a cluster [8]. The population density of the community should reach a certain level. Clustering algorithms that use density-based clustering assume that each cluster point's neighbourhood must have a minimum number of points at a given unit distance.

Researchers throughout the globe have created and tested a variety of agricultural forecasting methods. The K-means clustering approach was used to divide the rainfall data into four groups. Modelling the linear connection among predictor variables and one or even more independent variables using MLR is a common statistical practice. Year, sowing area, and production are all independent factors that have an impact on rainfall. Finding data models that are accurate and broad enough to accurately forecast yield is the goal of this research [9]. In Bangladesh, consumers may choose from a wide variety of rice cultivars, each with a distinct harvest season. Climate in Bangladesh and its influence on rice production have been studied in the past. After that, the results of this research were put through a regression analysis that included data on temperature and rainfall. Temperature has a negative impact on agricultural yields. The whole dataset was separated into three three-month periods during pre-processing [10]. It has been calculated and correlated with each characteristic during this period of time. Each rice type has been subjected to this pre-processing. As part of clustering, a variety of pre-processed tables were examined in order to identify regions with comparable meteorological characteristics. Hence, it is fair to anticipate that advanced computation algorithms have a significant impact in increasing the crops productivity.

III. METHODS AND MATERIALS

Methods and materials play a significant role in addressing the purpose of a study in a more credible manner. Apart from that, in this research study a secondary source of information has been taken into account for assessing the purpose of the study. However, ***experimental research analysis*** has been performed in order to gain information regarding the significance of data mining and advanced computational algorithms in fertilizer recommendation and increasing the yield of a firm. However, incorporating the DBSCAN algorithm that has been updated, the data are clustered by districts with comparable temperatures, rainfall, and soil types [1]. Clustering the data using PAM and CLARA is done based on the districts that produce the most crops. Moreover, the yearly agricultural production may be predicted using a multiple linear regression technique.

Evaluated DSCAN approach

DBSCAN is a foundational technique for density-based clustering with a lot of noise and outliers in a huge volume of data. In addition to this, Eps and MinPts are the only two DBSCAN settings. It is not possible to get the best Eps value using standard DBSCAN [1]. The DBSCAN's Eps value is one of the most important modifications that have to be made. In the below figure, the researcher has illustrated the DBSCAN method's improved approach.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 9, Issue 11, November 2022

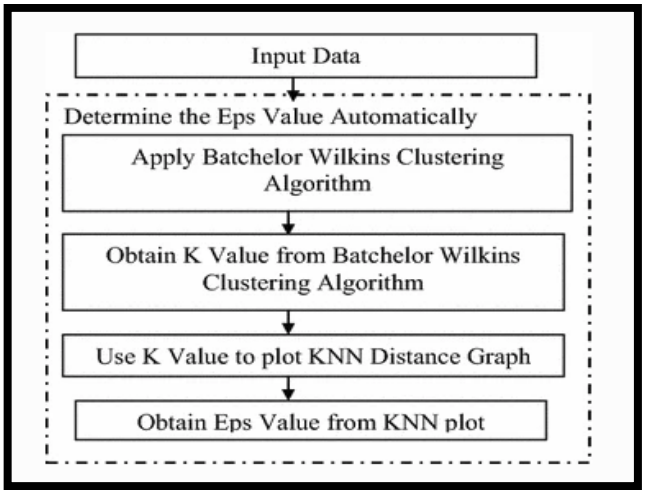

Figure 3: The model of DBSCAN
(Source: [1])

New DBSCAN suggests a way for automatically determining epsilon, the minimal points. User-defined input to the KNN plot is used to determine the epsilon value. Batchelor Wilkins clustering method is applied on the database and the K value and its related cluster centres are obtained without the user defining K as input to the KNN display. The KNN Plot requires this value as an input.

Crop yields are predicted using multiple linear regression

A variation on "linear regression" analysis is "multiple linear regressions". Modelling the link between a single dependent variable and two or more independent variables is the goal of this approach. Multiple linear regressions fit a dataset to a model for which x1, xk are independent variables and Y is a dependent variable.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

However, the matrix form is given below:

$$Y = XB + E$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} \quad E = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

Knowing the most important data points in a database before using multiple linear regressions to predict crop production is essential [1]. It is unlikely that any of the database's qualities will have an impact on the dependent variables, even while their values are altered. Multiple linear regressions are used to anticipate crop yields using just the significant features in the database after a P value test is done on the data.

Methods for evaluating

Data mining algorithms use a variety of techniques, each of which may be impacted by various types of data relationships. This study investigates the best clustering strategy for agricultural data analysis in order to maintain fairness in the assessment process. External quality measurements such as purity and homogeneity are used to compare the clustering algorithms of PAM, CLARA or DBSCAN in the research study. It is calculated by putting each cluster into a class that is most common in that cluster. Homogeneity signifies the presence of just one kind of individual in each cluster. Whenever the class is considered complete, all of its members are placed in the same cluster [1]. An individual's homogeneity and completeness scores are used to calculate the harmonic mean of the V-measure. Once it comes to making good judgments, the Rand Index is a great way to gauge your progress. The percentage of pairings successfully clustered together is known as precision. Recall is a measure of how many real pairings of objects were able to be correctly recognised throughout the search. The higher the quality metrics value, the better the cluster quality is represented.

IV. RESULT AND DISCUSSION

Experimental results

Adjusted approach of DBSCAN

```
.....................................
Total number of clusters : 7
.....................................
cluster point 1 is : BIJAPUR
.....................................
cluster point 2 is : DAKSHINAKANNADA
.....................................
cluster point 3 is : KOPPAL
.....................................
cluster point 4 is : SHIMOGA
.....................................
cluster point 5 is : UTTARAKANNADA
.....................................
cluster point 6 is : BAGALKOT
.....................................
cluster point 7 is : CHIKMAGALUR
```

Figure 4: Using the Batchelor Wilkins method, cluster centres has been founded
(Source: [1])

The Minpts and Eps values must be determined before the DBSCAN algorithm can be applied to the dataset. The dataset is subjected to the Batchelor Wilkins method in order to automatically compute the K value [1]. The dataset utilized in this study has a K value of 7, with the following districts serving as cluster centres, as determined by the Batchelor Wilkins. However, figure number 4 has interpreted the findings of the ***Batchelor Wilkins algorithm.***

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 9, Issue 11, November 2022

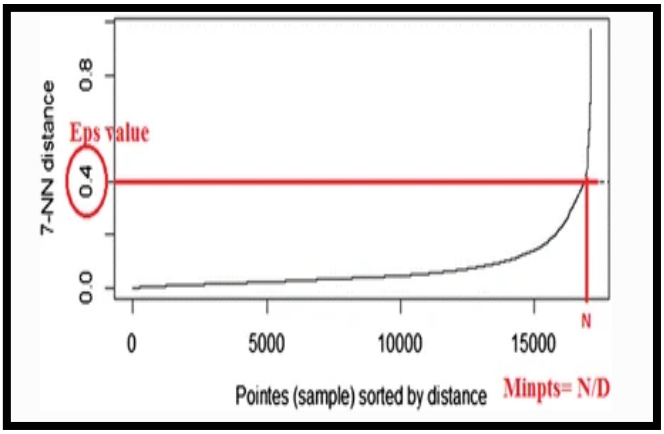

Figure 5: KNN plot
(Source: [1])

Batchelor & Wilkins' Algorithm K is used to compute the delta value and the minimum points for DBSCAN in KNN plots. Employing K derived via Batchelor & Wilkins' Algorithm, the KNN plot is produced using the K value of 7 [1]. It is determined by calculating the slope of a line from any place and a combination of sites that have the highest slope to pinpoint a specific location. Eps values an ideal slope for this line is at 0.4, where the line's slope is placed.

Multiple linear regression

Regression analysis has been taken into account as an independent variable whose "p value" is less than 0.05, indicating that the "null-hypothesis" may be rejected. As a result, the model may be expanded to include these additional parameters. In contrast, if the p-value is greater than the standard alpha threshold of 0.05, the variable will be deemed insignificant to the model [1]. On the other hand, figure number 6, has highlighted the usage of **multiple linear regression equations** for enhancing the productivity of crops.

Crop	Yield forecast equation
Cotton	Yield = (7.149372) + (−0.14468)pH + (−0.00131) Nitrogen + (−0.00405) Potassium + (−0.00405) Water Required
Groundnut	Yield = (2.79115) + (0.029217) Temperature + (5.78e−05) Rainfall + (−0.05681) pH + (−0.00127) Phosphorus + (−0.00492) Water Required
Jowar	Yield = (−1.62694) + (−5.35e−02) Temperature + (0.051512) pH + (−0.00113) Potassium + (0.01685436) Water Required
Rice	Yield = (−0.18503) + (0.041593) Temperature + (0.172042) pH + (−8.27e−04) Nitrogen + (−4.28e−03) Phosphorus + (−0.00264) Potassium + (9.15e−04) Water Required
Wheat	Yield = (112) + (−4.14e−02) Temperature + (1.34e−04) Rainfall + (0.079153) pH + (−1.31e−03) Nitrogen + (−0.00167) Potassium + (−0.28125) Water Required

Figure 6: Multiple regression analysis for various crop yields
(Source: [1])

There is a yield of 112 if all independent variables are zero in a wheat harvest. Increasing water requirements reduces yield by 0.28125 units for every 1 unit increase in temperature. Increasing precipitation increases yield by 1.34e04, while increasing pH by 0.079153. Reducing nitrogen decreases output significantly by 1.31e03 and increasing potassium decreases yield by 0.00167 units for every 1 unit increase in water requirement.

CLARA

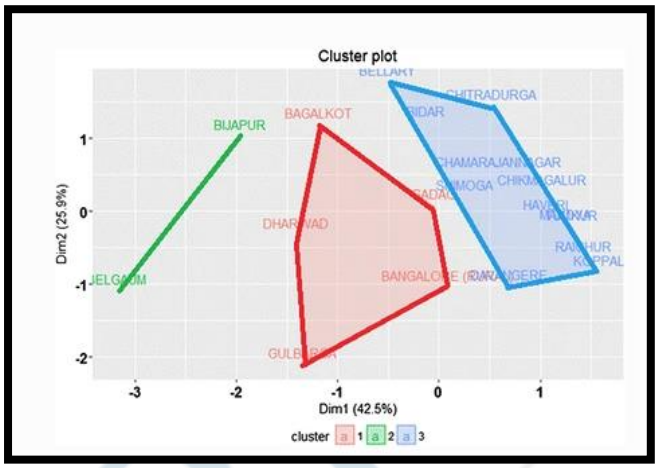

Figure 7: Result of Cluster plot in R language
(Source: [1])

Through the incorporation of the CLARA method, the districts in the data set are grouped into three groups [1]. However, figure number 7 displays areas with comparable characteristics, such as land area, production, precipitation, and temperature, in a graphical way.

Large area, production and moderate rainfall, temperature (24–26)	Moderate area, production and high rainfall, temperature (27–29)	Low area, production moderate rainfall, temperature (29–30)
Bijapur, Belgaum	Gadag, Gulbarga, Dharwad, Bangalore, Bagalkote	Koppal, Davangere, Shimoga, Haveri, Chikmagalur, Bidar, Chamarajannagar, Tumkur, Mandya, Raichur, Bellary

Figure 8: Result of CLARA algorithm
(Source: [1])

On the other hand, figure 8 displays the results of the CLARA algorithm [1]. In addition to that, the figure number 9 has depicts the temperature and wheat crop productivity in Karnataka's several districts.

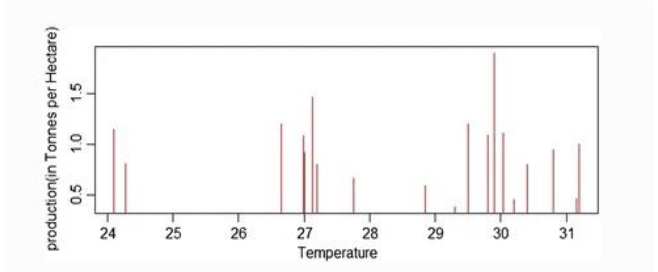

Figure 9: Production versus plot temperature
(Source: [1])

Moreover, the figure number 9 has highlighted that 29.9 degree Celsius is the ideal temperature for growing wheat.

V. DISCUSSION

The PAM, CLARA, and Modified DBSCAN clustering algorithms, as well as the multiple linear regression approach, are all discussed in this paper. Batchelor Wilkins clustering technique is used to find the 'k' value in PAM and CLARA

whereas DBSCAN introduces the KNN clustering method to automatically identify minimal points and radius values. These technologies are used to analyze agricultural data sets and identify the most effective settings for growing wheat. Multiple linear regressions are employed to identify the most important variables and build an equation for predicting yields [1]. The effectiveness of the clustering algorithms is measured using internal quality measures and external quality metrics, respectively. However, in these studies, the external quality metrics, which are a composite of multiple measures, such as set matching metrics, measurements based on counting pairs, and metrics based on Entropy, are the only metrics that are examined.

VI. CONCLUSION

The input data is subjected to a variety of data mining methods in order to determine which approach provides the highest results. In this study, data mining methods PAM, CLARA, and DBSCAN and advanced computation methods were utilized in order to determine the ideal wheat climatic requirements, such as the best and worst temperature ranges and rainfall amounts, in order to boost wheat crop output. Quality measures are used to compare different clustering approaches. CLARA provides better clustering than PAM, whereas DBSCAN provides better clustering than PAM and CLARA, according to the analysis of quality measures. Additional study might be done by analyzing the soil and other crop-related variables in order to boost crop productivity under diverse climatic circumstances. On the other hand, the findings of the study has highlighted that the inclusion of data mining and advanced computation processes has also induced a positive format of impact on the growth of the agricultural field.

REFERENCE

[1] Majumdar J, Naraseeyappa S, Ankalaki S. Analysis of agriculture data using data mining techniques: application of big data. Journal of Big data. 2017 Dec;4(1):1-5.

[2] Issad, H.A., Aoudjit, R. and Rodrigues, J.J., 2019. A comprehensive review of Data Mining techniques in smart agriculture. *Engineering in Agriculture, Environment and Food*, *12*(4), pp.511-525.

[3] Kale, S.S. and Patil, P.S., 2019. Data mining technology with fuzzy logic, neural networks and machine learning for agriculture. In *Data management, analytics and innovation*(pp. 79-87). Springer, Singapore.

[4] Archana, K. and Saranya, K.G., 2020. Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. *SSRG Int. J. Comput. Sci. Eng*, *7*, pp.1-4.

[5] Bhanumathi, S., Vineeth, M. and Rohit, N., 2019, April. Crop yield prediction and efficient use of fertilizers. In *2019 International Conference on Communication and Signal Processing (ICCSP)* (pp. 0769-0773). IEEE.

[6] Nath, B.D. and Sarkar, L., 2021. A Random Forest Algorithm for Predicting Crop Yield in Hilly Regions of North East India. *International Journal of Modern Agriculture*, *10*(2), pp.3415-3433.

[7] Ahmed, U., Lin, J.C.W., Srivastava, G. and Djenouri, Y., 2021. A nutrient recommendation system for soil fertilization based on evolutionary computation. *Computers and Electronics in Agriculture*, *189*, p.106407.

[8] Rajeswari, S. and Suthendran, K., 2019. C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. *Computers and Electronics in Agriculture*, *156*, pp.530-539.

[9] Kulkarni, M.A., 2021. Automatic Agriculture Crop Yield Production Maintenance System Based On Remote Monitoring Techniques In Cloud Environment. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(11), pp.4409-4416.

[10] Jha, K., Doshi, A., Patel, P. and Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture*, *2*, pp.1-12.