JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 825 - 831
https://publishoa.com
ISSN: 1309-3452

Performance Of Soil Prediction Using Machine Learning For Data Clustering Methods

M. Rajeshwari[1], N. Shunmuganathan[2], Dr. R. Sankarasubramanian[3]
[1](Ph.D Research Scholar, Department of Computer Science, Erode Arts and Science College, Erode)
[2](Ph.D Research Scholar, Department of Computer Science, Erode Arts and Science College, Erode)
[3](Principal & Research Supervisor, Erode Arts and Science College, Erode)

ABSTRACT
Objective: Agricultural play a major role in human life. The crop yield prediction is the needed one, because the investment and work process consume high but the yield output going low in every year. **Methods:** Here introduces machine learning (ML), which can be a key differentiator for obtaining real, estimated predictions for yield issues. In ML, we choose the random forest algorithm for the yield predictions. The classifier model used here includes logistic regression, naive Bayes, and random forests, of which extended random forests provide maximum accuracy **Findings:** Based on the dataset provided, we got the yields prediction by RF. The crop yield is different by the crops and usage of fertilizer. The fertilizer also depends upon the soil of the place. **Novelty:** The clustering method considers data-related environmental factors, soil factors and weather, soil fertility, and production over the past year, and recommends profitable plants that remain mature in the expected atmospheric conditions.

Keywords: Crop Yield, Machine Learning, Random Forest Algorithm, Fuzzy K-Means, Yield Prediction.

1. INTRODUCTION

Agriculture is one of the most important elements and plays an important role in personal life. India's flagship presence, agricultural fragmentation, continues to cultivate through community demands as know-how advances. The main profession of people in most rural areas of Tamil Nadu. According to statistics, Tamil Nadu has a total of 123,100 km2 of land cultivated, which occupies 64.60% of the state's total geographic area. Agriculture remains the main activity and livelihood of the state's rural population. It is characterized by a large difference in yield and residual resources that are highly dependent on the whims of the southwest monsoon. Agriculture employed 50% of India's workforce, accounting for 17-18% of the total population of 1.4-1.8 per 10,000 rupees in the decade since 2005. The purpose of this paper is to help farmers get information about the location of previous crop predictions [1]. Knowing the soil and its advantages, they are determined to produce crops with mass R programming [2] using key tools for machine learning techniques, statistics, data analysis and machine learning. It's more than a statistical package, and its programming language creates its own package. R programming is platform independent and can be used with any operating system. The dataset contains parameters such as precipitation, season, temperature, and crop production. Random forest [3] is an ensemble classifier that uses many decision tree models to predict results. To train each tree, spares are used to select different subsets of training data. The collection of trees is a forest, and the trees are trained in a selected subset in a random forest. It can be used for both classification and regression problems. Session responsibilities are completed by the number of divisions in all trees and are used to reduce normal results.

2. METHODOLOGY

Preprocessing is a technique for improving data excellence that is suitable for the mining process. Data preprocessing is used in three main ways: (I) Data cleaning (ii) Feature selection (iii) Change. Data cleaning is the process of modifying incomplete, inconsistent, noisy data. To gain a great advantage; scrubbing is completed by ignoring non-essential criteria. Contributions that further contribute to the mining process are identified in this process. The signature range is completed by eliminating both inappropriate and unwanted features. The process of adapting the data to the procedure intended for extraction is called transformation. Farmers give different qualities and restrictions as input and later use support vector machines to predict possible yields of low, medium and high. The generation of SVM [4] is absorbed by Naive Bayes,

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 825 - 831
https://publishoa.com
ISSN: 1309-3452

which affects the accuracy of the calculation. According to our analysis, temperature, precipitation, and soil type are the most commonly used functions and the most commonly used algorithms. Although the yield prediction model can reasonably estimate the actual yield, it is still desirable to improve the performance of the yield prediction.

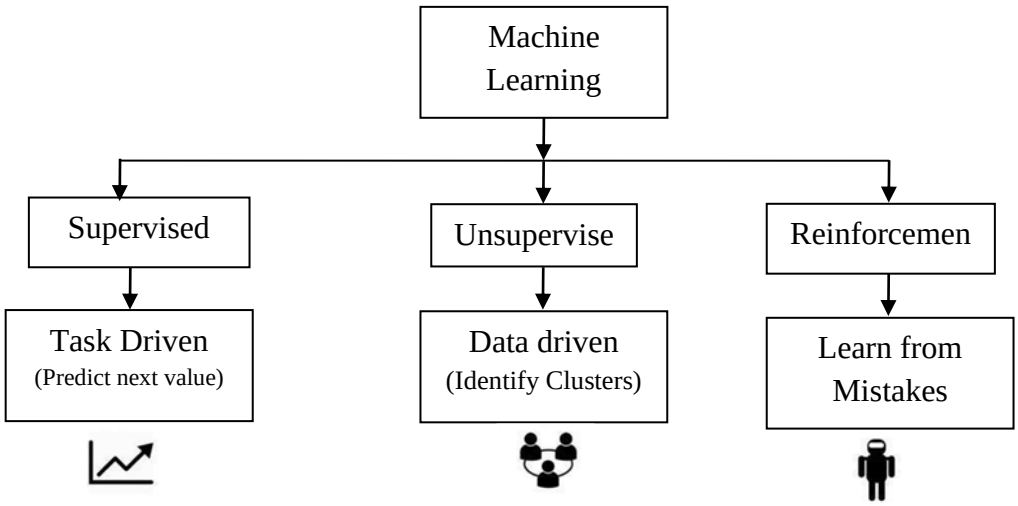

Fig 1 Types of Machine Learning

Prediction is one of the difficult problems in precision agriculture, and many models have been proposed and validated so far. Crop yields depend on a variety of factors, including climate, weather, soil, fertilizer use, and seed type, so multiple datasets should be used for this issue [5].

RANDOM FOREST ALGORITHM

Random Forest builds a couple of random Forest bushes in addition to merges them collectively to get extra correct in addition to strong prediction. One massive gain of random Forest is that it could be applied for each clustering in addition to regression problems, which paperwork the mainstream of cutting-edge device studying systems. The random Forest set of rules stays that its miles equal casual to amount the relative function of every characteristic at the prediction. A notable device for this that measures a capability significance via way of means of searching at how tons the tree nodes that use the characteristic lessen impurity throughout all bushes within side the Forest. It computes this rating mechanically for every characteristic after schooling and scales the outcomes the sum of all significance is same to one.

3. RESULT

The energetic parameters in random forest are either utilized to increase the predictive power of the model to make faster.

3.1 To increase the predictive power:

Energy parameter n the job indicates the number of processors available to the engine. It consumes a value of 1 and can only use a single procedure. Random state energy parameters make the output of the model reproducible.

3.2 To increase the model's speed:

The n jobs energetic parameter tells the engine how many processors it is allowed to use. It consumes a value of one, it canister only utilize single procedure. The random state energetic parameter makes the model's output replicable.

Step 1: Start by collecting a random model from the expected dataset.

Step 2: This algorithm creates an extended random forest tree for each sample. Get the prediction results from each extended random forest tree.

Step 3: You can make a voting decision for each predicted result.

Step 4: The maximum calculation result selected will be awarded as the final estimation result.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 825 - 831
https://publishoa.com
ISSN: 1309-3452

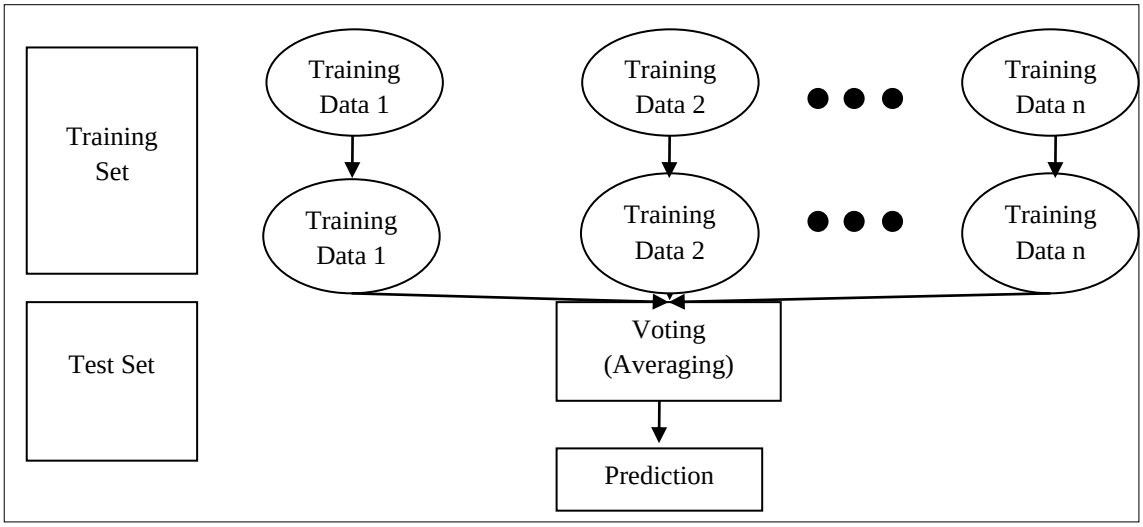

Fig 2 Random Forest Algorithm Flow

Random forest is based on the concept of ensemble learning, which combines multiple classifiers to solve complex problems and improve model performance. Crop predictive analytics is used to predict the right crop using various soil parameters such as nitrogen, fertility, pH, phosphate, precipitation, and moisture. These factors directly affect crop yields and their complexity. Large amounts of data depend on many factors. The proposed system aims to help farmers predict crop yields and rely on many factors such as temperature, rainfall, humidity, soil nutrients and soil erosion. Random forest classification uses an ensemble technique to get the results. Training data is provided to train a large number of random forest trees. Every random forest tree consists of a decision node, a leaf node, and a root node. The leaves in each figure are the final work formed by the very decision tree. The final production assortment overlooks the design chosen by the majority.

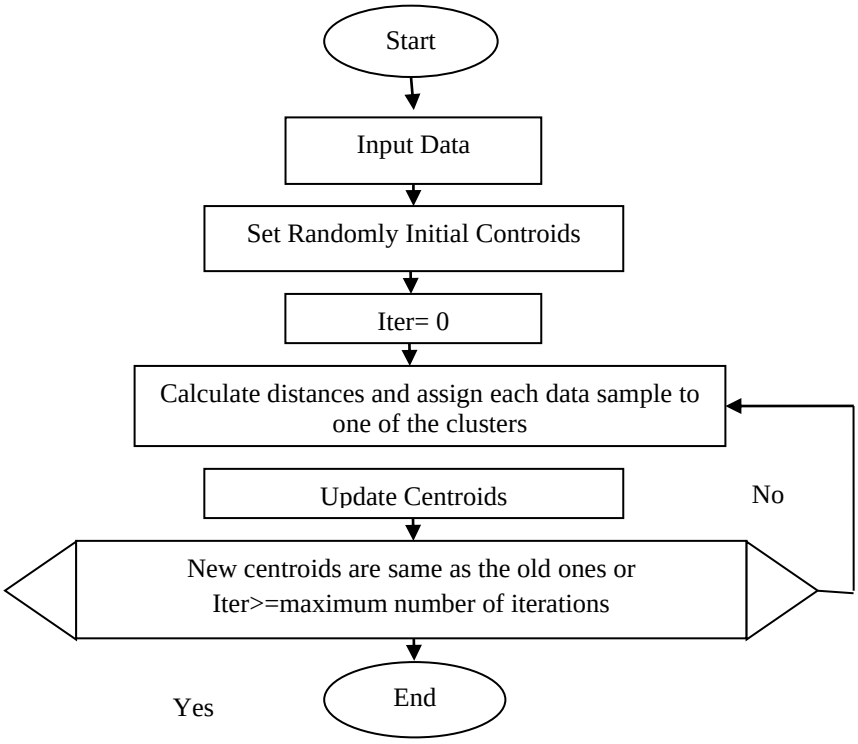

Fig 3 Proposed Enhanced Radom Forest Work Flow algorithm

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 825 - 831
https://publishoa.com
ISSN: 1309-3452

Figure 3 shows the workflow model. The workflow model consists of two phases, a training phase and a test phase. During the training phase, the input data is fed to the input layer and pre-processing is performed using the hidden layer. The training phase contains the input layer and the hidden layer, and the test phase contains the input layer. A hidden layer containing the test phase from the input layer and a hidden layer containing the training phase from the output layer.

A hidden layer that contains only the output layer in the test phase. Data pre-processing includes redundancy checking, data filtering, outlier detection, correlation testing, feature extraction, and more. When the training phase is complete, the data will flow to the test phase. It categorizes the data based on the relevant algorithms and uses predictions to predict the values present in the data.

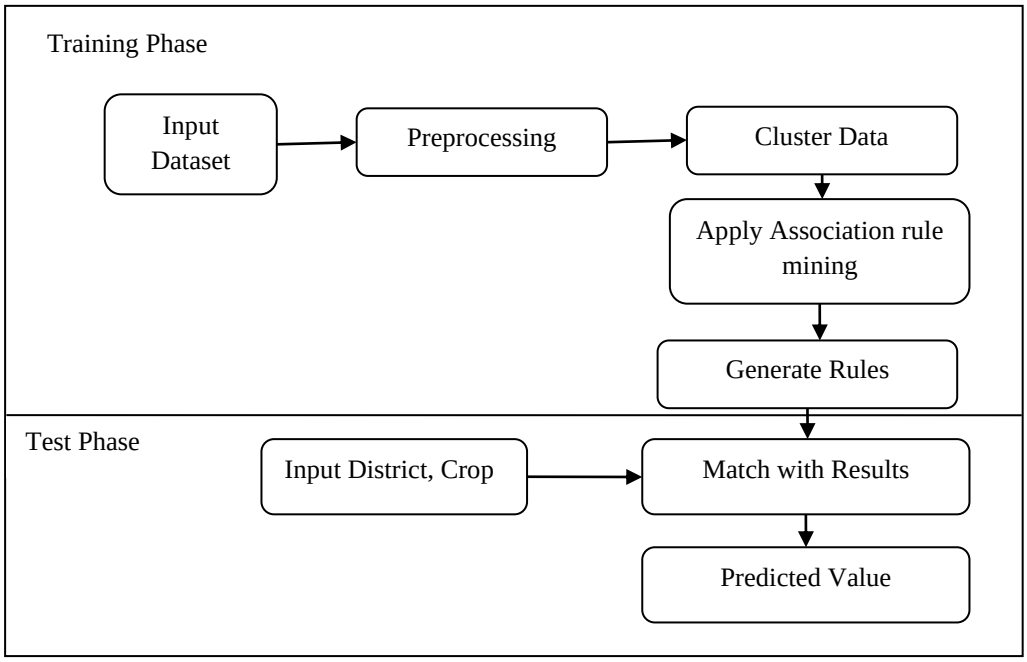

Fig.4 Proposed Radom Forest Flow algorithm

A graphic representation of a system developed to predict yields based on weather conditions. You can understand how the system works by following the steps below. (As depicted in the flowchart):
Step1: The user logs in to the system.
Step2: If login is successful, the location of the user is tracked.
Step3: If it returns —yes the user is suggested not to use the fertilizer.
Step4: Else the user is suggested to use the fertilizer.
Step5: The selection of random samples since a assumed dataset.
Step6: Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
Step7: Voting will be achieved for each prophesied result.
Step6: select the most voted estimate consequence as the ultimate forecast result.

The system now provides user with the following the paths:

Prediction Module: The user can choose to know the prediction of a particular crop or know the list of crops with their corresponding productions.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 825 - 831
https://publishoa.com
ISSN: 1309-3452

Yield Prediction: The utilize requirements to afford crop, soil type as well as area as inputs. The system precedes the production of the crop given.

Crops Prediction: The utilize requirements to provide soil type as well as area as inputs. The system returns a list of crops along with their production values.

Fertilizer Module: The utilize can choose this module to know if it is the right time to utilize the fertilizer.

$$\boldsymbol{X}\mathbf{ms}: \frac{\sqrt{(x_1^2 + x_2^2 + x_3^2 \ldots . x\ x_n^2}}{N}$$

This is completed by forecasting the precipitation aimed on the next one year. In Table1 Discuss the Performance Metrics for Random Forest Algorithm helps to find accuracy, Precision, Recall values compare with existing algorithms like k-mean[6], k-medoids[7], Enhanced Random Forest Algorithm. The accuracy formula assists to distinguish the errors in the capacity of standards.

PERFORMANCE METRICS	ACCURACY	PRECISION	RECALL
K-Mean	88%	75%	73%
K-Medoids	84%	79%	79%
ABC with weighted based CLARA Algorithm	96.71%	92.67%	91.3%
Advanced PAM Algorithm	96.98%	93.78%	92.3%
Effective Fuzzy K-Means Algorithm	96.99%	93.89%	93.2%
Enhanced Random Forest Algorithm	97.02%	93.99%	94.06%

Table.1Performance Metrics for Enhanced Random Forest Algorithm

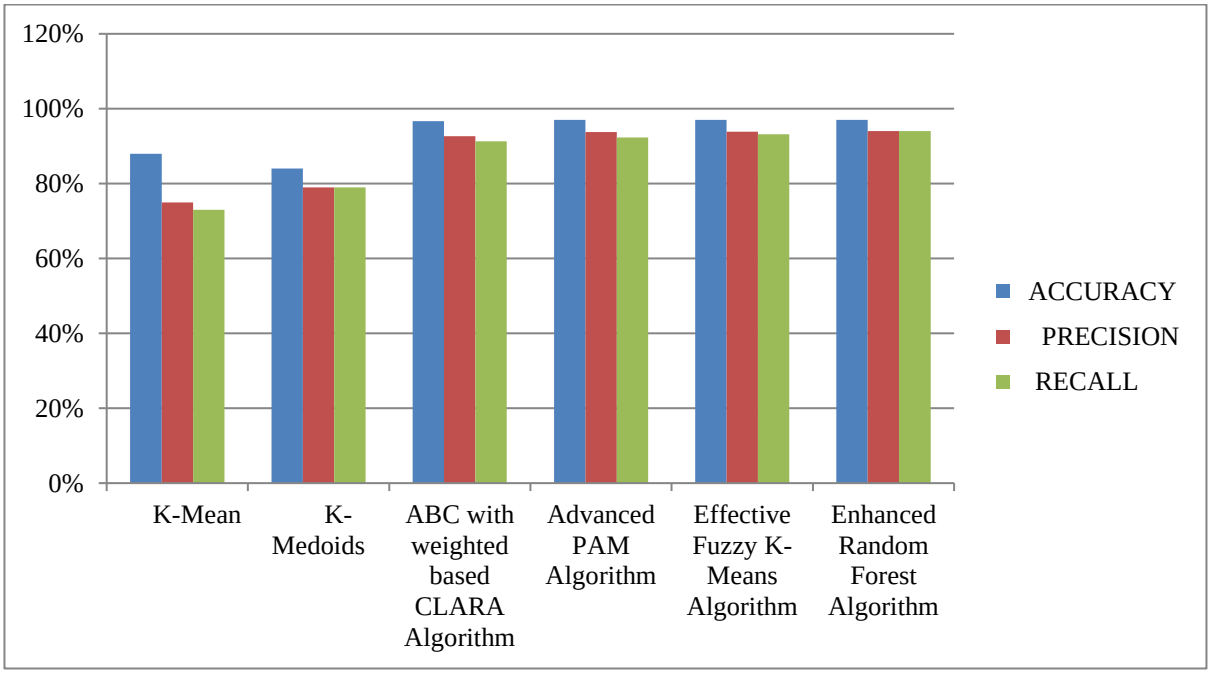

Fig.5 Proposed Metric values in Enhanced Random Forest Algorithms

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 825 - 831
https://publishoa.com
ISSN: 1309-3452

In Figure 5, the proposed metric values are explained using an extended random forest algorithm and you can find the accuracy, accuracy, and recall values for KMean, KMedoids, and the three proposed methods. If a good rating corresponds to the actual value, it is more accurate and less error-prone. Accuracy and error rate are inversely proportional. High accuracy means low error rate, and high error rate means low accuracy. Precision formulation expresses precision as a percentage. Here, the sum of precision and error is equal to 100 percent.

PERFORMANCE METRICS	**MAE**	**RMSE**
K-Mean	1.37	0.97
K-Medoids	1.73	1.96
Enhanced Random Forest Algorithm	2.32	3.45

Table 2: Performance Metrics- MAE, RMSE for Advanced PAM Algorithm

Commonly used planning methods such as recall, precision, FMeasure, and land accuracy are biased, with a clear understanding of the bias and statistical probabilities or base case levels.

Recall = True Positives / (True Positives + False Negatives)

The Mean Absolute Error (MAE) is the normal of entirely complete errors. The formula is:

$$\mathbf{MAE = 1/n \sum_{i=1}^{n} |x_i - x|}$$

n = the number of errors,
Σ = summation symbol
$|x_i – x|$ = the absolute errors.

3.3 Root Mean squared error (RMSE)

Underlying square error or RMSE[8] is a periodically realistic amount of realistic amount between the statistical population value and the samples predicted by the estimator or mode. RMSE refers to an intelligent anomaly of deformation of expected values and observations. Single these changes are detected as residues as residues when calculations are estimated as prediction errors and calculations are performed through data samples known as prediction errors. Route Means

$$\mathbf{RMSE=\sqrt{\sum^{n}_{n-1} (X_{objecti}- X_{modeli})/n}}$$

The square is quantified below to obtain the RMS value of a set of data values. RMSE from prediction that is best for approved adjustable XMODELIS distinguishes expressions with words, corrections under the calculation, and the detected standards are crushed and then closer to the samples.

4. DISCUSSION

The crop yield prediction in agriculture is difficult one in nature. But we apply the Machine learning with the existing data. The Random forest and their cluster techniques provides the accurate result of the process. The soil of the land place major role in the yield. If soil health is good means we put fertilizer based on the crop grow only otherwise it’s

not needed . Because the soil have almost fertilizers in nature. In modern agriculture fertilizer not necessary to put every time in the soil. So treat the soil in good manner then yield automatically increase. The proposed extended random forest, and other algorithms that are said to give the best results for these crop prediction methods. About the specific analysis process of soil classification in Tamil Nadu.

5. CONCLUSION

The crop yield prediction in agriculture based on the machine learning algorithm. Its helps Agriculture to reach, reaching the decision to stimulate most yield by stimulating factors such as temperature, precipitation, area etc. Grouping the agriculture of artificial intelligence will benefit from most farmers in the future. An important role in the application of ML in the prediction of cultivation based on a particular important element, and the evaluation of crop yields created with current / past data. A new clustering algorithm, called the Extended Random Forest Algorithm for text data segmentation, is adaptive and iteratively processed to run the CLARA algorithm, advanced PAM algorithm, EFKM, and ERFA algorithms before the actual and predicted values. The difference between is very large. Extended Random Forest continued to maintain an accuracy of 90 or higher.

REFERENCES

[1] Thomas Van Klompenburga, AyalewKassahuna, CagatayCatal," Crop Yield Prediction using Machine Learning: A Systematic Literature Review", Journal Homepage: www.elsevier.com/locate/compag, received 29 January 2020; Received in revised form 21 July 2020, Computers and electronics in agriculture 177 (2020) 105709, Accepted 9 August 2020, Available online 18 august 2020,0168-1699/ © 2020 elsevierb.v. All rights reserved.
https://doi.org/10.1016/j.compag.2020.105709

[2] M. Prakash, G. Padmapriy, M. Vinoth Kumar "A Review on Machine Learning Big Data using R ",2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 27 September 2018, DOI: 10.1109/ICICCT.2018.8473342

[3] M. Abarna, M. Akshaya, S. Illakiya , A.P. Janani , "An Effective Crop Prediction Using Random Forest Algorithm " , 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 30 November 2020, DOI: 10.1109/ICSCAN49426.2020.9262311

[4] Vaibhavi Vanarase, Vrushali Mane, Harsha Bhute, Ankita Tate, Simran Dhar, "Crop Prediction Using Data Mining and Machine Learning Techniques", 01 October 2021, DOI**:** 10.1109/ICIRCA51532.2021.9544724

[5] S. Bhanumathi, M. Vineeth, N. Rohit, "Crop Yield Prediction and Efficient use of Fertilizers", 25 April 2019, **DOI:** 10.1109/ICCSP.2019.8698087

[6] Hassan Ibrahim Hayatu, Abdullahi Mohammed, Ahmad Barroon Ismaeel, Yusuf Sahabi Ali, "K-means Clustering Algorithm Based Classification of Soil Fertility in North West Nigeria", Fudma Journal of Sciences (FJS), ISSN online: 2616-1370, ISSN Print: 2645 – 2944, Vol. 4 no. 2, June, 2020, Pp:780–787, DOI: https://doi.org/10.33003/fjs-2020-0402-363

[7] P.Surya, I.Laurence Aroquiaraj, "Crop Prediction Analysis in North western Zone of Tamilnadu using Artificial Bee Colony with Weighted based Fuzzy Clustering", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

[8] Ansarifar, J., Wang, L. & Archontoulis, S.V. An interaction regression model for crop yield prediction. *Sci Rep* **11,** 17754 (2021). https://doi.org/10.1038/s41598-021-97221-7,