IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 10-15

Discovery of Agricultural Patterns Using Parallel Hybrid Clustering Paradigm

P.Arun, M.Phil, Dr.A.Senthilkumar
Research Scholar, Sankara College of Science and Commerce,Saravanampatti,Coimbatore-35
Associate Professor, Department of MCA, Sankara College of Science and Commerce

Abstract: *Data mining model has been adopted in various fields including agriculture. Data mining has been used to interpret specific details and used as process to predict the crop patterns in a clustering form in order to get more precise and accurate information. Crop productivity depends upon the various factors like climate, Type of soil, precipitation and groundwater level in the specific region. Many clustering algorithm has been derived in existing to group the amount of data related to the agriculture. Nowadays the due to employment of information gathering sensor, data about the agriculture is exploiting, in order explore to these data finds to critical using traditional data model. Hence it leads to importance of the big data solution. In this paper, we propose a novel hybrid clustering technique as parallel computing paradigm towards big data mining. Hybrid clustering technique is the combination of Partition around Medoids (PAM) and Clustering for large application (CLARA). Those techniques works in parallel on the data instance to form a cluster based on relationships, Rules and criteria's. The established relations discover the interesting patterns. It is the one of diverse technique which generates rules cooperatively to perform a variety of knowledge discovery. The discovered knowledge provides the suggestion in terms of crop yield based on soil, extent of crop yields in the specific soil and crop planting based on soil and groundwater level. The experimental results of the proposed model on agriculture data has been computed in terms of precision, recall and Fmeasure as part of the research. The result on the current dataset proves that proposed model outperforms state of art approaches in all aspect of prediction accuracy.*

Keywords: *Agricultural data, Clustering, Prediction, Forecasting, Partition around Medoids, Clustering for large Applications*

I. Introduction

Nowadays data mining model has been exposed to agriculture field. Data mining is a process of discovering the hidden information as patterns from the dataset [1]. It is further transformed into user preferred form. Discovering of patterns based on the some specified constraints is named as prediction [2]. Many supervised and unsupervised learning model has employed to extract the knowledge from the agriculture dataset. The extraction of the knowledge is carried out in terms of classification and clustering. The agriculture dataset uses the either classification or clustering to maximize the crop yield[3]. The crop yield depends several factors like soil, groundwater level, climate, precipitation and cultivation. Crop production prediction is carried out using clustering algorithm. The clustering algorithm examines the data instance and groups the data instance into cluster based on distance computation [4]. Due to increase in dimensionality of dataset, many state of clustering algorithm has found insufficient and it leads to importance of solutions through big data analytics [5]. In this work, Hybrid clustering paradigm composed of partition around medoid and Clustering for large application has been proposed. The paradigm analyses the data instance to establish the cluster on basis of data instance relationship, data association rules and distance estimation criteria's. Cluster determines the domain specific predictions [6]. The prediction results indicate crop related categorizations. Additionally it is used for forecasting.

The Rest of the paper is sectioned as, section 2 describes the related work on clustering the agricultural data and section 3 devises the proposed model which followed by section 4 with experimental results on various data performance metrics and dataset. Finally section 5 concludes the work

II. Related Work

In this part, clustering model to examine the data instance of the agriculture field has been carried out on various constraints towards forecasting or prediction of crop related categories.

1.1. Predictive ability of machine learning methods for massive crop yield prediction.

Clustering technique majorly classified into Partitioning clustering, Hierarchical clustering and Density based methods The Machine learning algorithms like naive bayes and decision tree is used to predict the massive crop

yield in the agriculture. Prediction accuracy has been compared with machine learning techniques on different datasets. It process in terms of simple probabilistic assumptions. It is determined based on likelihood estimation. Decision tree uses tree or graph model towards processing of the information for further prediction [7].

1.2. Convex full and DBSCAN clustering

Density based clustering approach is most used to predict the information from the data instances in the cluster.DBSCAN is used to predict future weather conditions [8]. Convex-Hull method is strictly used to convert unstructured data into its corresponding structured form. These structured data is efficiently and effectively used by the DBSCAN clustering algorithm to form resultant clusters for weather derivatives

1.3. Composition Clustering with Numerical Attributes

The Composition clustering with numerical attributes of dataset is carried out. The composition model is combination of PAM and CLARA model. It produces the predict value only numerical data. Considerable information is obtained on the prediction results of this model. PAM and CLARA collaborate with feature selection model for initial sample grouping.

III. Proposed Model

In this section, proposed hybrid clustering model has been designed and described in details with its process of information prediction, the clustering model is as follows

1.4. Data Collection

In this model, the data is been collected from various source by employing crawling technique. The data collected contain the information about crop, soil and weather. In addition, types of crop, area, production, average rainfall, temperature, pH value etc has been extracted from the sources to predict the maximum crop production on various categories. The data collection is composed of numerical and categorical data.

1.5. Hybrid Clustering Model

The hybrid clustering model composed of two models which is work parallel to categories the data instance into cluster in parallel. The composed models are

- Partition around medoid
- Clustering for large application

The clustering models aids as big data solution for handling high dimensional data on some distance measures. The figure 1 represents the architecture of the proposed model

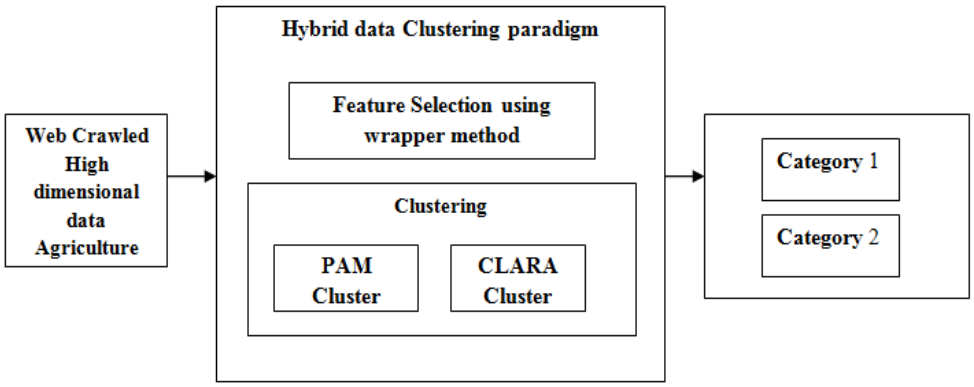


Figure 1: Proposed Architecture of the Research

1.5.1. Partition Around Medoid

Initially the extracted feature subset is partitioned into random clusters. The clustered instances are computed to determine the set of objects. The Set of objects is termed as medoid which has centre value. The medoid helps in determining the nearest data points to form clusters. It uses batchelor Wilkins technique to obtain k values. With K value the nearest data points is mapped using distance graph.

1.5.2. CLARA technique

CLARA techniques use the sampling instead of mapping of set of objects. On extraction of some samples in dataset, CLARA employs PAM and computes the medoid to the sample. CLARA extracts multiple sample and its computes the neighbour accurately. The Distance between the data points is computed in terms of average dissimilarity of all objects in the entire data set.

1.6. Regression Model

Applying regression model is used to forecasting or prediction of information from the clustered records.

Algorithm

Input: Agriculture Data (High Dimensional Data)
Output: Crop related prediction
Process:
Apply PAM ( )
Partition data instance into random groups R1, R2, R3
Select centroids value as M1, M2, M3 in the random group
If (distance < Specific Threshold)
Cluster the instance in Cluster 1
Else
Create new cluster, place the data instance into it
Apply CLARA ( )
Obtain some sample from the dataset
Use PAM( ) to calculate to centroids
Neighbour is determined after computation of multiple samples
Apply Regression ( )
If ((Attribute 1 & Attribute 2)== Constraint 1)
Crop related constraint 1
Else crop related constraint 2
The Regression model forecast the crop yields on several constraints effectively by significant attributes of the dataset.

IV. Experimental Results

The proposed model is computed against several performance measures towards analysis of agriculture data. The measures like precision ,Recall and f measure were used to compare the hybrid clustering models . Precision is measured in terms of set of the instance mapped with high similarity index. Recall is measured as actual instance identified. F measure is mean of precision and Recall. The Figure 2 represents the performance measures of the proposed model against existing approaches.

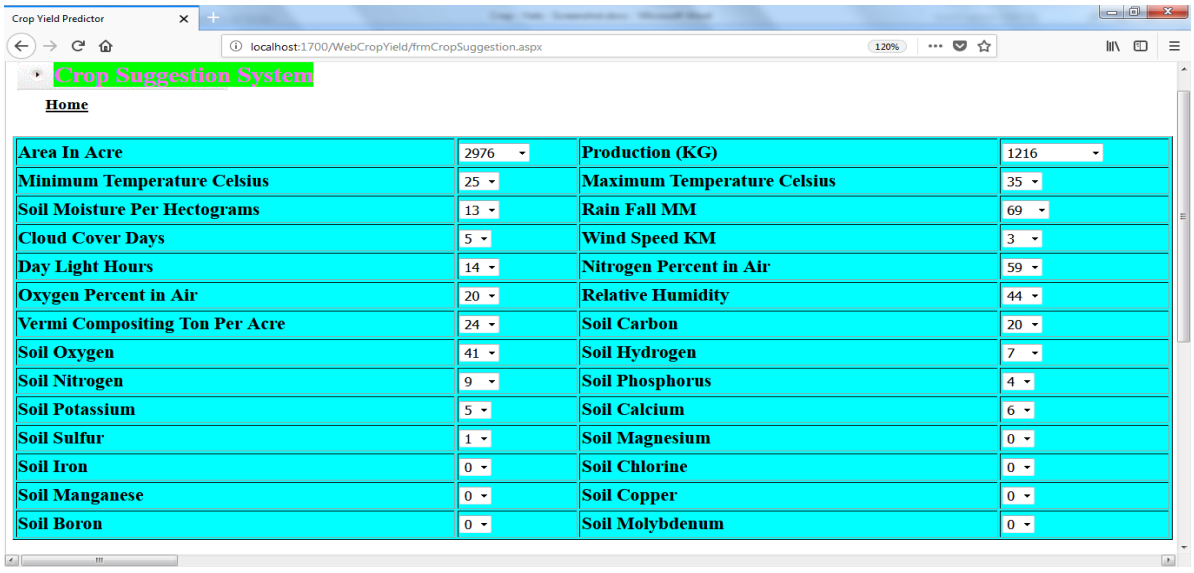


Figure 2: Input for Numerical Attributes only

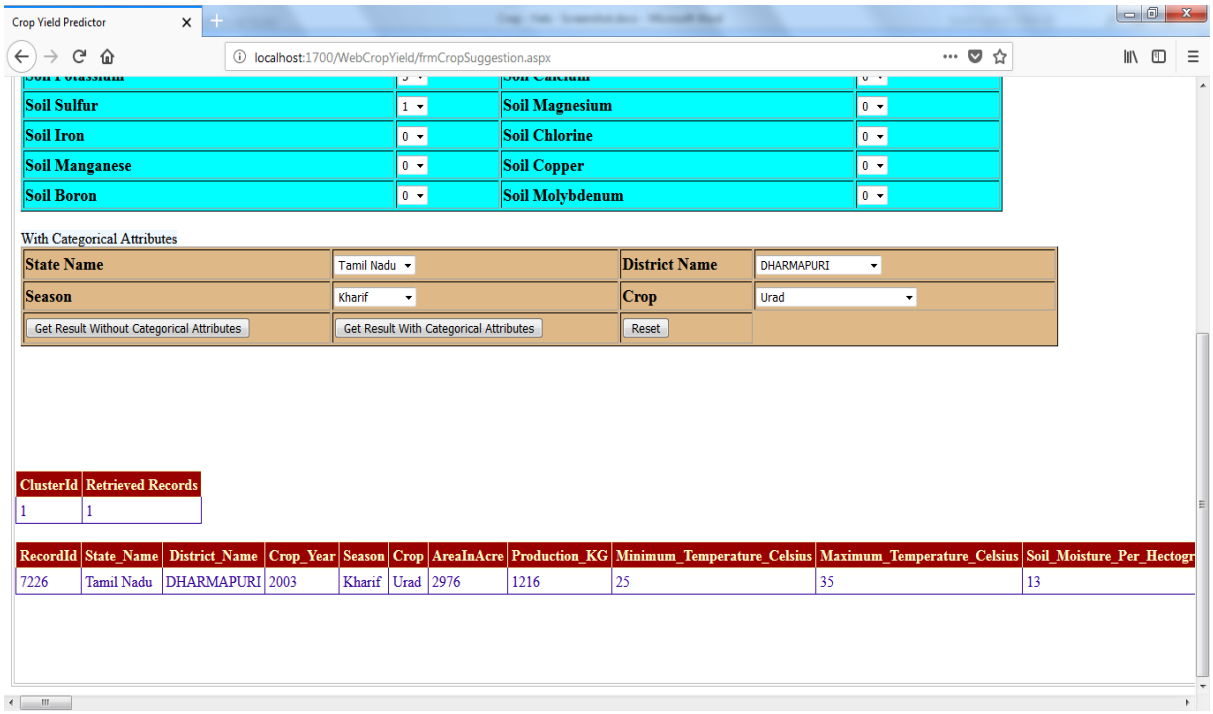


Figure 3: Output for Numerical Attributes only

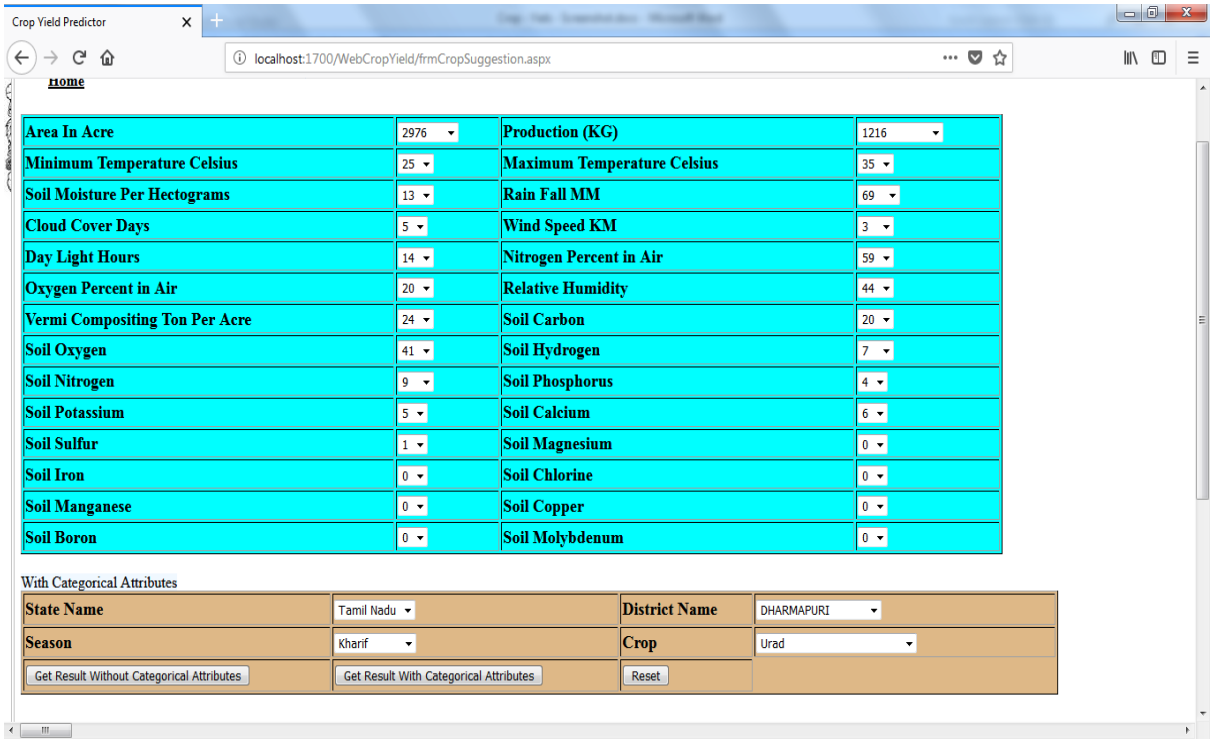
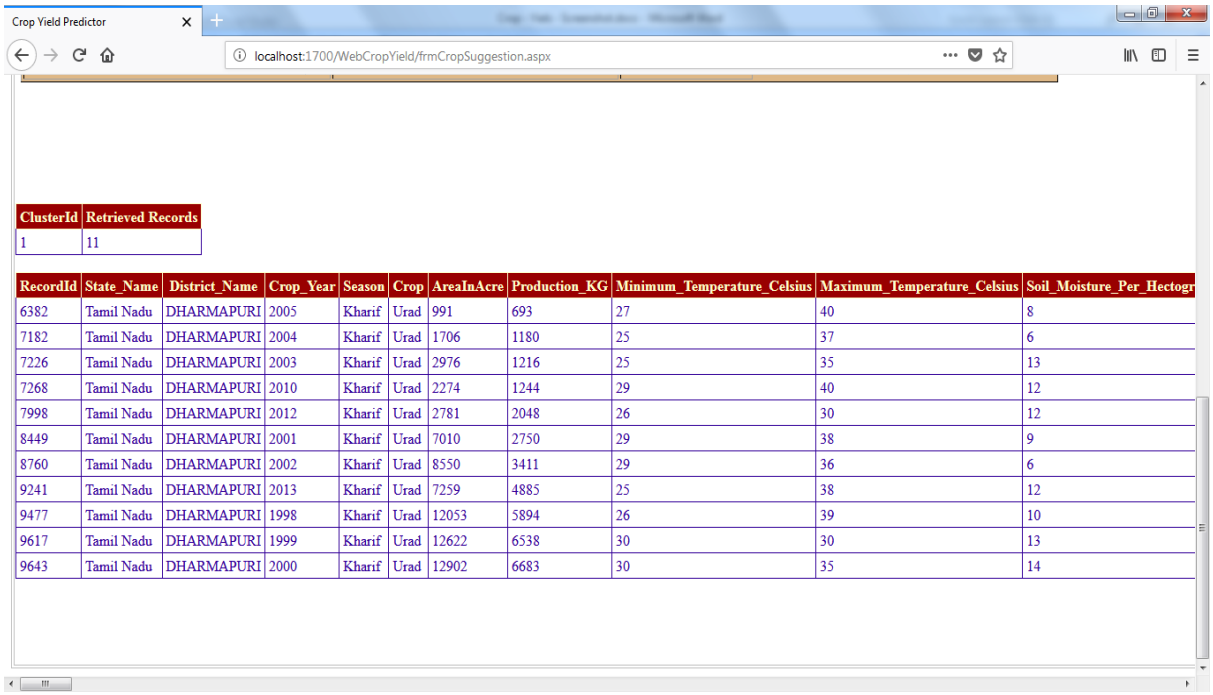


Figure 4: Input for Categorical Attributes & Numerical Attributes



ClusterId	Retrieved Records
1	11

RecordId	State_Name	District_Name	Crop_Year	Season	Crop	AreaInAcre	Production_KG	Minimum_Temperature_Celsius	Maximum_Temperature_Celsius	Soil_Moisture_Per_Hectogr
6382	Tamil Nadu	DHARMAPURI	2005	Kharif	Urad	991	693	27	40	8
7182	Tamil Nadu	DHARMAPURI	2004	Kharif	Urad	1706	1180	25	37	6
7226	Tamil Nadu	DHARMAPURI	2003	Kharif	Urad	2976	1216	25	35	13
7268	Tamil Nadu	DHARMAPURI	2010	Kharif	Urad	2274	1244	29	40	12
7998	Tamil Nadu	DHARMAPURI	2012	Kharif	Urad	2781	2048	26	30	12
8449	Tamil Nadu	DHARMAPURI	2001	Kharif	Urad	7010	2750	29	38	9
8760	Tamil Nadu	DHARMAPURI	2002	Kharif	Urad	8550	3411	29	36	6
9241	Tamil Nadu	DHARMAPURI	2013	Kharif	Urad	7259	4885	25	38	12
9477	Tamil Nadu	DHARMAPURI	1998	Kharif	Urad	12053	5894	26	39	10
9617	Tamil Nadu	DHARMAPURI	1999	Kharif	Urad	12622	6538	30	30	13
9643	Tamil Nadu	DHARMAPURI	2000	Kharif	Urad	12902	6683	30	35	14

Figure 5: Output for Categorical Attributes & Numerical Attributes

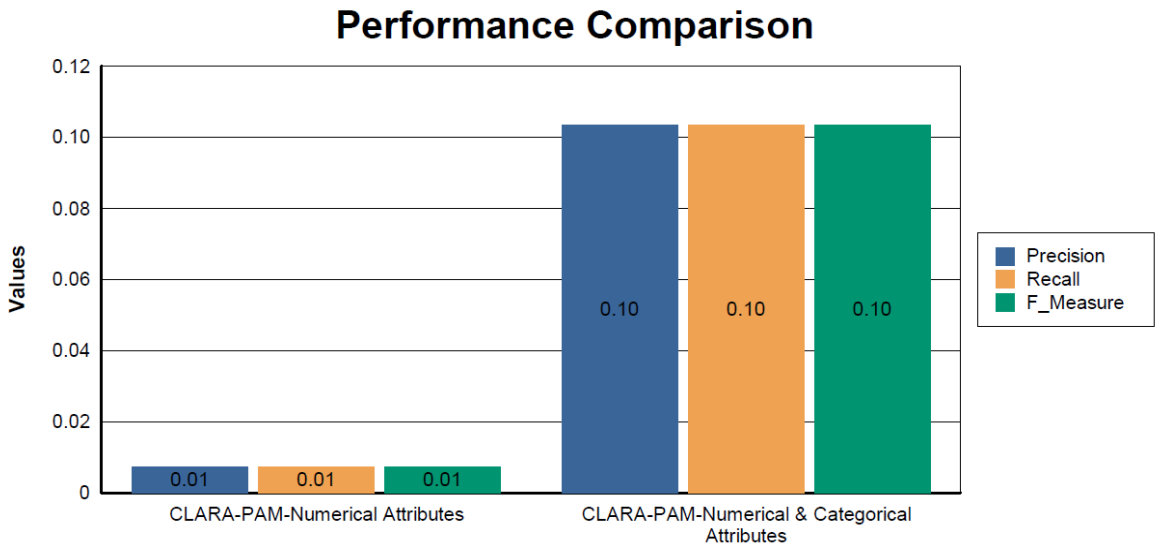


Figure 6: Performance Evaluation of clustering model against various measures

The Medoid and neighbour value is computed using Batchelor Wilkins algorithm. The table 1 provide the performance value of proposed model and existing value for different medoid values.

Table 1: Performance Values of the proposed model against Existing model

Technique	Precision	Recall	Fmeasure
PAM +CLARA with Numerical and Categorical Attributes	0.1033	0.1034	0.1033
PAM+CLARA with Numerical Attributes	0.0073	0.0074	0.0074

V. Conclusion

The hybrid clustering paradigm towards agriculture data has been designed and implemented. The combined clustering of PAM and CLARA on the numerical and categorical data of the dataset provides the valuable prediction based on the relationship, rule and criteria's employed on the data instances of the dataset. The proposed explore large dimensional data without dimensionality reduction effectively. According the experimental results, proposed model gives better prediction accuracy compared with various states of art

approaches. The proposed model provides the predictions in terms of the crop yield based on the soil, Maximum crop yield on the particular soil and crop classification based on the soil, climate and groundwater level.

References

[1]. Shweta Taneja, Rashmi Arora, Savneet Kaur, "Mining of Soil Data Using Unsupervised Learning Technique", International Journal of Applied Engineering Research, Vol. 7 No.11, 2012.

[2]. D Ramesh , B Vishnu Vardhan, "Data mining technique and applications to agriculture yield data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013 .

[3]. Chandrakanth Biradar, Chatura S Nigudgi,"Statistical Based Agriculture Data Analysis", in International Journal of Emerging Technology and Advanced Engineering, 2012.

[4]. M. Divya, T.N. Manjunath, R.S. Hegadi, "A Study on Developing Analytical Model for Groundnut Pest Management using Data Mining Techniques", In IEEE International Conference on Computational Intelligence and Communication Networks (pp. 691-696).,2014

[5]. A. Raorane, R. Kulkarni, "Data Mining: An effective tool for yield estimation in the agricultural sector", In International Journal of Emerging Trends and Technology in Computer Science, vol.1, issue 2, pp. 75-79, 2012.

[6]. G. Fathima, R. Geetha, "Agriculture crop pattern using data mining techniques", in International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, issue.5, pp. 781-786, 2014.

[7]. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28. Science 2016

[8]. Ratul Dey,Sanjay Chakraborty "Convex-hull & DBSCAN clustering to predict future weather"in International Conference and Workshop on Computing and Communication,2015