Experimental Analysis of Machine Learning Algorithms Based on Agricultural Dataset for Improving Crop Yield Prediction

Kusum Lata, Sajidullah S. Khan

Abstract: ***Agriculture is the primary research study area in India as agriculture is the main source of income for various communities. In classification algorithm for agricultural dataset according to production, area, crop and seasons. Here, four classification algorithms are used with the help of WEKA tool. These algorithms are namely the present scenario, there is a call to renovate the enormous agriculture data into diverse technologies and make them accessible to the farmer for improved decision making. The endeavor of this work is to find out the finest Random Tree, J48, Bayes Net and KStar etc. The captured results revealed that Random tree algorithm performed well in terms of error rate and provides slightly better performance than KStar, Bayes Net and J48 classifiers. In this paper, our objective is to apply machine learning techniques to mine constructive information from the agricultural dataset to improve the crop yield prediction for major crops in Nashik district of Maharashtra.***

Keywotrds: WEKA tool, J48, Bayes Net, KStar and Random Tree, Machine learning and Crop Yield Prediction

I. INTRODUCTION

In India population is increasing rapidly as compare to the world population. Hence the demand of food will increase exponentially with limited resources and land. According to the Government data approximately seventy percent of the Indian population is living in villages. In villages, agriculture is a kind of employment for farmers and their growth depend on the agricultural output. But there are many issues untouched in the agricultural domain. In these issues crop yield prediction is the major research area. Early crop yield predictions can help the farmer's to avoid the losses and maximize their crop yield.In broad sense we can say this is the issue originally related to the agriculture planning [8]. Here our primary objective is to experimentally analyze the various machine learning algorithms and suggest the best one to get the higher crop yield accuracy results. Crop yield prediction generally depends on the weather conditions, soil and geography [15]. Secondly, we want to equip the farmers with better technologies in order to make the profit from their agricultural output. In concrete terms machine learning algorithms are classified as supervised and unsupervised machine learning algorithms. In both the algorithms raw data is given in order to predict or find the various data trends to find out the most accurate crop yield prediction.This paper is organized in a sequential manner. Section II is about the related work and Section III is about the core of this research i.e. proposed approach. Section IV describes the Methods and Materials required for the study. Dataset description discussed on Section V. Section VI was discussed with Results and analysis. Section VII concludes the work with possible future enhancement.

Revised Manuscript Received on October 05, 2019

Kusum Lata, Research Scholar, Department of Computing Sciences and Engineering, Sandip University, Nasik.

Sajidullah S. Khan, Associate Professor, Department of Computing Sciences and Engineering, Sandip University, Nasik.

II. RELATED WORK

In the field of crop yield prediction various researchers have contributed. Here, in this section we have tried to communicate the earlier research as mentioned below:Diepeveen and Armstrong [6] discuss about various key attributes which will decide the performance of crop in different areas depending on the season, sowing and soil type etc. This paper mainly concentrates on different data mining techniques to enlighten the crop yield performance.Monali Paul et al. [10] In order to predict the yielding of the crops, the crops are analyzed and based on the categorized analysis. This categorization is done based on data mining algorithms. This paper gives insight into various classification rules like Naive Bayes, K-Nearest Neighbor.Murynin et al. [7] study the dependency between the prediction and the accuracy of the forecast. Linear model is used for yield prediction. This model is extended with non-linear attributes in order to improve the accuracy of the prediction. The accuracy of the model has been expected based on the time period between the moment of the forecast formation and the time of harvest.Hemageetha [8] mainly focuses on using the soil parameters like pH, Nitrogen, moisture etc. for crop yield prediction. Naive Bayes algorithm is used to classify the soil and a77% accuracy is achieved. Apriori algorithm is used to associate the soil with the crops that could provide maximum yield in them. A comparison of accuracy achieved during classification using Naïve Bayes, J48 and JRIP is also presented.

III. RESEARCH METHODS

The main components which contributes to this research are sample study area, agricultural datasets and firm methodology.

A. Study Area

For this particular research, we have selected Nashik district of Maharashtra state as a sample study area. In Maharashtra state there are total thirty six districts, out of these all Nashik is agriculture oriented and catering the agricultural demands of the nearby states. In Nashik, farmers are growing various crops and out of the various crops arhar, bajra, caster seed, cotton, groundnut, maize,

moong, niger seed, ragi, rice, soyabean, sunflower, safflower, gram, jowar, wheat and sugarcane produced in this district we have selected etc.

B. Data Set

For this experiment, dataset is extracted from link https://data.gov.in. In this government website agricultural datasets are shared containing data of past few years. This dataset contains various columns in Microsoft Excel like production area, crop, Crop year and season etc. This CSV file has the hundred records of eight districts of Maharashtra state of India. Moreover this file contains data used for twenty different crops.

C. Major Factors Used

The major factors used in this sample dataset are described as below:

Year: This will contains in which year how much production takes place. This is further classified for every crop in various districts of the state.

Season: In India the crop seasons are basically classified in two types. These are scheduled as below:

- **Kharif** (July- Oct.)
- **Rabi** (Oct.-March) _Winter crops and (March-June)_Summer crops

Area: It contains cultivation (Hectare) area according to the crop and district wise.

Production: It contains total production (Tons) for according to the crop and district wise.

IV. METHODS AND METHODOLOGY

In this paper, we are focusing to preprocess the data and using weka tool to analyze the simulation results obtained after applying the machine learning classifiers. Comprehensive depiction of weka tool is described as below:

WEKA Tool: The acronym WEKA can be elaborated as Waikato Environment for Knowledge Analysis. This tool was come into existence at University of Waikato, New Zealand. Here in this tool a variety of machine learning algorithms realized that can be implemented directly to agricultural data set. Graphical User Interface is well designed to compare the various experimental results. Machine learning supports the artificial intelligence to learn or train the data without any external assistance. Weka is a Preprocess: This step contains the preprocessing of data set in various ways.

- Classify: The training and testing of data set occurs which performs the classification and regression and also evaluate them.
- Cluster: The clustering of dataset takes place.
- Associate: Association rules are applied for the data set and evaluation takes place.
- Select attributes: Selection of the most relevant attributes of the given dataset.
- Visualize: Visualization of data set occurs in the different two-dimensional plot and interact with them.

The main features of weka are as listed below:

- Platform independent
- Open source and free
- Different machine learning algorithms
- Easy to use
- Data prepossessing tools
- Flexibility for scripting experiments
- Graphical user interface

Weka tool consists of four keys which are described as below:

- Explorer: It is the location where you can explore the data
- Experimenter: It is the location where you can execute the experiments and can bear statistical tests between learning outlines.
- Knowledge flow: It is the location which upkeep essentially the similar functions as the EXPLORER nevertheless with a drag and drop interface.
- Simply CLI: It offers an easy command-line interface that permits implementation of WEKA commands for operating systems and does not permit their own command line interface.

V. EXPERIMENT RESULTS ANALYSIS

To find the experimental results dataset is obtained as discussed above. This dataset contains seven columns. The columns are indentified as state, district, crop year, crop area, season and annual production. The stated crop dataset is explored in Weka tool and the initial preprocessing of the data was completed. The Weka tool is capable of handling various data formats. The files can be ARFF (Attribute relation file format) [RemcoR.Bouckaert 2010], Comma Separated value 51 format (CSV), URL, Decision induction algorithm acceptable format and SQL database etc. Before you apply the algorithm to your data, you need to convert your data into comma-separated file (.csv) into ARFF format (.arff extension). The sample agricultural dataset for Nashik district is illustrated in Fig.1 robust tool which can be instrumental to work out the real time data mining problems. It contains tools for:

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Maharashtra	NASHIK	2011	Kharif	Arhar	10900	8300
Maharashtra	NASHIK	2011	Kharif	Arhar	165900	170100
Maharashtra	NASHIK	2011	Kharif	Arhar	4400	2000
Maharashtra	NASHIK	2011	Kharif	Arhar	47000	96500
Maharashtra	NASHIK	2011	Kharif	Arhar	27400	28100
Maharashtra	NASHIK	2011	Kharif	Arhar	5900	7900
Maharashtra	NASHIK	2011	Kharif	Arhar	173800	554000
Maharashtra	NASHIK	2011	Kharif	Arhar	8800	7000
Maharashtra	NASHIK	2011	Kharif	Arhar	15800	5000
Maharashtra	NASHIK	2011	Kharif	Arhar	12100	6700
Maharashtra	NASHIK	2011	Kharif	Arhar	40000	36300
Maharashtra	NASHIK	2011	Kharif	Arhar	62900	77800
Maharashtra	NASHIK	2011	Kharif	Arhar	100	100
Maharashtra	NASHIK	2011	Kharif	Arhar	58600	81500
Maharashtra	NASHIK	2011	Kharif	Arhar	100	50
Maharashtra	NASHIK	2011	Kharif	Arhar	14900	14500
Maharashtra	NASHIK	2011	Rabi	Arhar	37900	27000
Maharashtra	NASHIK	2011	Rabi	Arhar	7000	5000

Fig.1 Sample Agricultural Dataset (Nashik)

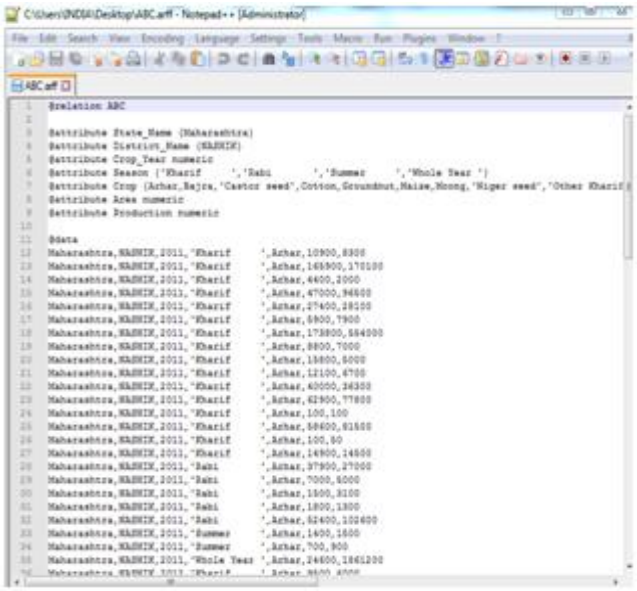

Fig.2 Attribute relation file format (ARFF)

A. Data Set and Attribute Selection

We have collected the crop data set which contains 95 instances and 7 attributes. The data file has to be in either in 'CSV' format or 'ARFF' format. In order to explore dataset the following steps need to be integrated.

- Load data
- Pre-process data
- Analyze attributes

The pre-process section enables you to load data into weka tool. The dataset may be imported from a CSV file as well as from numerous data formats like Binary, ARFF and C4.5.Moreover, Weka tool provides the feature to read the data from a database server and from a web address Sample ARFF file is shown in Fig 2. After importing the data the preprocessing tools are applied to get the output in terms of discretization, re-sampling, normalization, attribute selection etc. The pictorial representation of preprocessing tool is shown in Fig 3. and Fig 4.

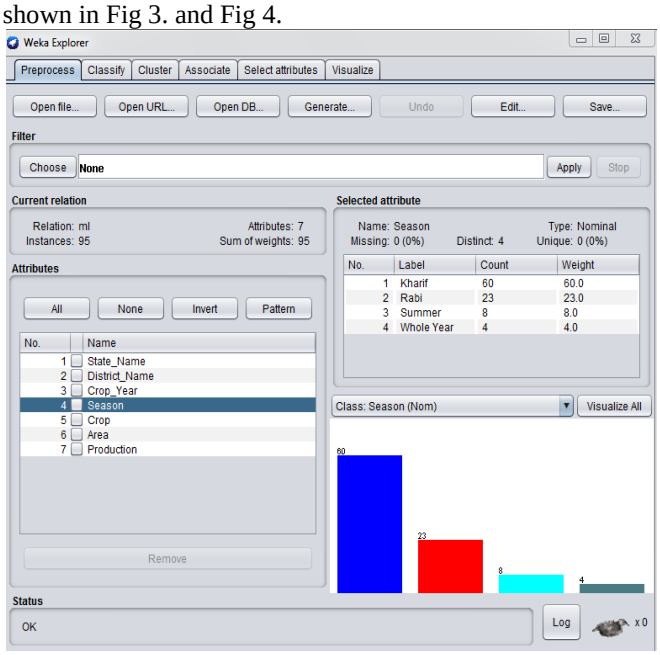

Fig. 3Weka Explorer window

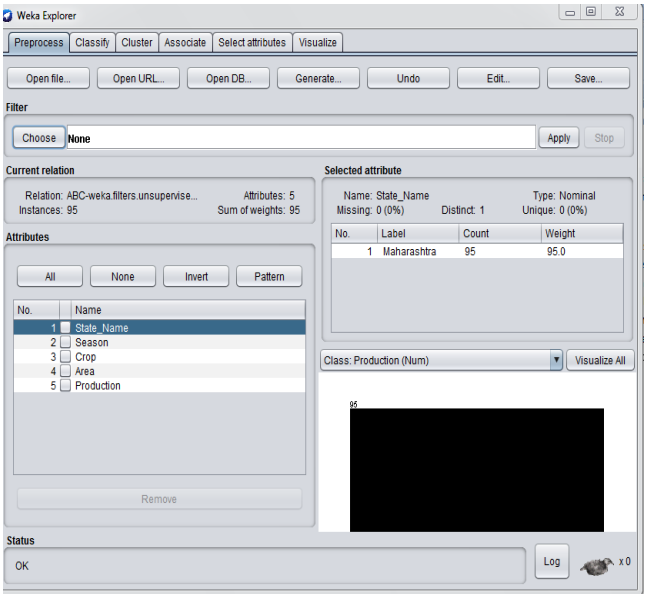

Fig. 4Pre-processing window

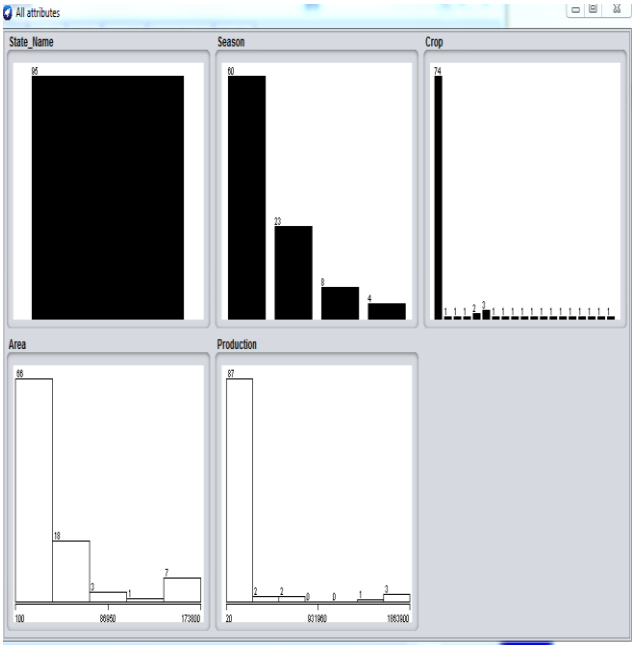

Fig. 5 Visualization of attributes

In Fig 5,we can visualize the attributes based on the selected class. By clicking on the "Visualize All" button, one can visualize the selected attributes.

B. Filters

The preprocessing tools in weka are called filters. As per our requirement, one can select or remove the attributes and also apply filter on our dataset. Weka filters can be used to modify the data sets, so they are called data preprocessing tools.

C. Explore: Building classifiers:

In WEKA tool, classifiers are the models used for classification and regression.

- Choosing a classifier: Once you have loaded your data set in Weka tool you have to click on the classifier button, and choose the appropriate classifier for your data set and one can start

analyzing the data by making the use of provided algorithms. Here, I have analyzed my data with the Random Tree algorithm.

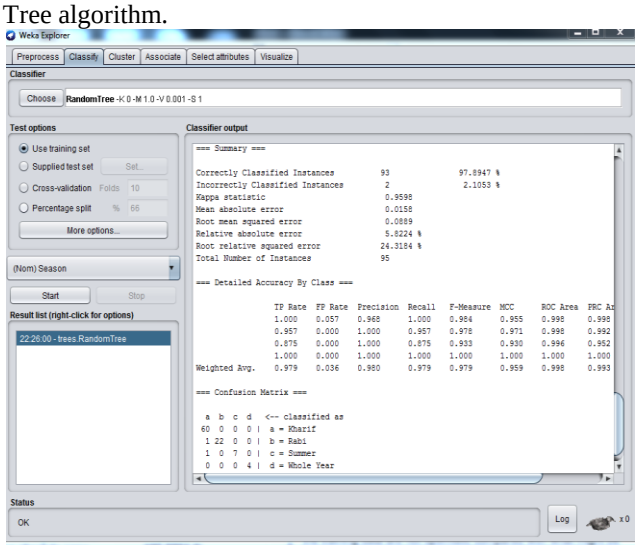

Fig.6 Result window of classification (Random Tree)

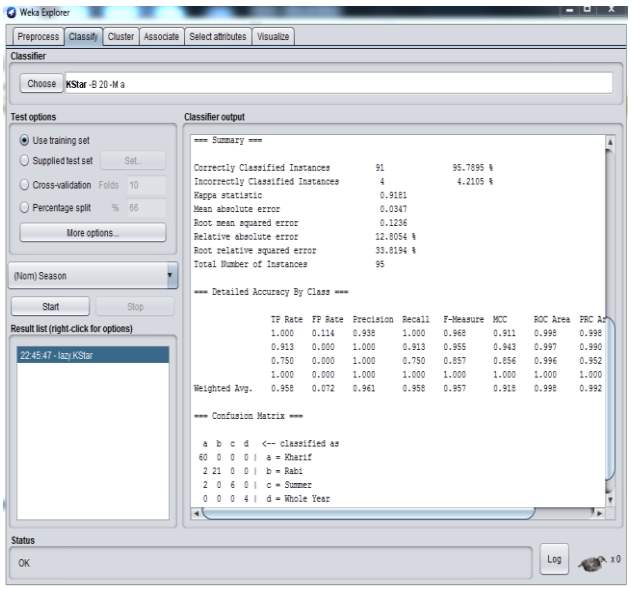

Fig. 7KStar Classifier Results

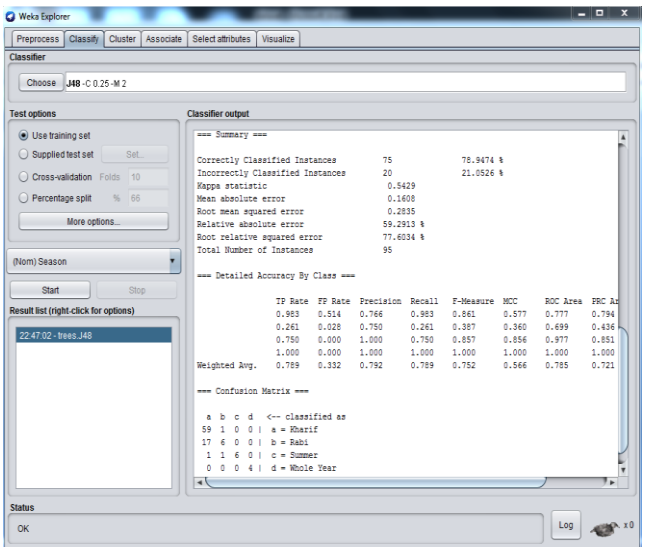

Fig. 8J48 Classifier

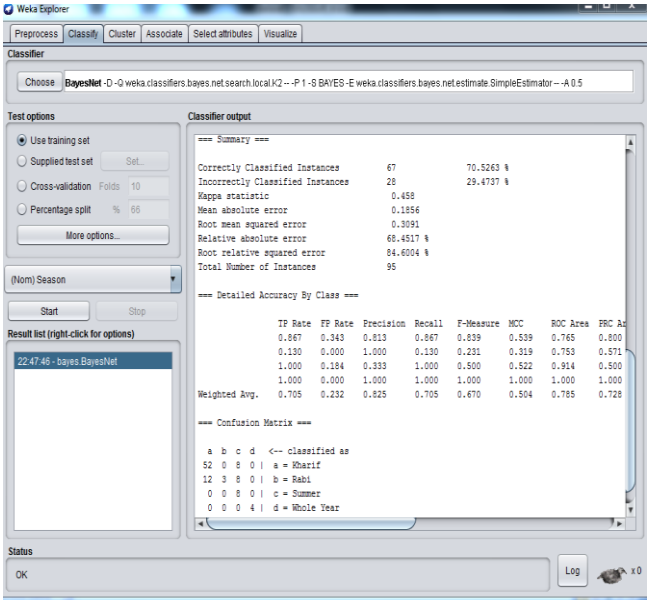

Fig. 9 Bayes Net Classifier

To describe the performance of used various classifiers; a confusion matrix is formed containing the values inside it. Confusion matrix depicts the performance of the classifiers.

VI. PERFORMANCE EVALUATION

The basic evaluation measures from confusion matrix are described in this section. Confusion matrix is a combination of four outcomes such as accuracy, sensitivity, specificity and precision. This also contains the ROC and precision-recall etc. Four outcomes of classification:

- True positive (TP): As the classifier name appears, TP is just the representation of number of correct classifiers reserved to positive class.
- True negative (TN): As the classifier name appears, TN is just the representation of number of correct classifiers reserved to negative class.
- False positive (FP):As the classifier name appears, FP is just the representation of number of classifiers reserved to positive class but in realism fit in to negative class.
- False negative (FN):As the classifier name appears, FN is just the representation of number of classifiers reserved to negative class but in realism fit in to positive class.

Sensitivity (REC/TPR): This is also called as true positive rate and recall. It can be determined as the no. of correct positive predictions divided by the total no. of positives.

Specificity (SP/TNR): This is also called as true negative rate. It can be determined as the no. of correct negative predictions divided by the total no. of negative.

Precision (PPV): This is also called positive predictive value. It can be determined as the no. of correct positive predictions divided by the total no. of positive.

Accuracy: It can be determined as the no. of all correct predictions divided by the total no. of dataset.

RAE (Relative absolute error): Relative absolute

error is just the replica of relative squared error.

RRSE (Relative root mean square error): It is basically applied to calculate the differences between the different values estimated by a model, estimator and the observed values. Sometimes it is also termed as root mean square deviation (RMSD).

MAE (Mean absolute error): The difference between the two continuous variables is called as mean absolute error. For the mean absolute error subtract mean value from each term and calculate the mean of modulus i.e. positive of those values.

The formulas used to calculate the performances are described in Table I. as shown below:

Table I: Model Performance

S.No.	Performance Evaluation	Formula used
1	Sensitivity	$SN = \frac{TP}{TP + FN}$
2	Specificity	$SP = \frac{TN}{TN + FP}$
3	Precision	$PREC = \frac{TP}{TP + FP}$
4	Accuracy	$Accuracy = \frac{TN + TP}{(TN + TP + FN + FP)}$

VII. EXPERIMENTAL RESULTS AND DISCUSSIONS

The experimental results are shown with 10 fold cross validation to avoid overlapping. Four classifiers are studied here namely Random Tree, KStar, Bayes Net and J48 etc. The performances of the classifiers are analyzed in TableII.

A. Performance error of classification algorithms

Accuracy of different classification algorithms are also shown in Table II according to area harvested production, season and crop. From table it is observed that Random tree and KStar higher accuracy than Bayes Net and J48 classification algorithms.

Table II. Performance Error for Classification Algorithms

S.No	Classifiers	RAE	RRSE	MAE	Accuracy
1	Random Tree	5.8224	24.31	0.015	97.89
2	KStar	12.8054	33.81	0.034	95.7
3	Bayes net	68.4517	84.60	0.1856	70.5
4	J48	59.2913	77.60	0.1608	78.9

The results are captured from the WEKA tool and it is observed that Random tree classifier has higher values of accuracy and less error rate. These results are purely depends on the dataset.

VIII. CONCLUSION – MACHINE LEARNING VS. STATISTICS

The recent growing population draws attention to meet the requirements of the people to assure the food security and protect the environment. In this connection crop yield production and its early prediction is very crucial. Crop yield prediction using intelligent machine learning techniques may improve the crop planning decisions. Here in this proposed research many comparisons between the various machine learning algorithms like Random tree, KStar, Bayes Net and J48 using the dataset are performed to get the most accurate technique for crop yield prediction. In our research, it is predicted that the experimental comparisons of various algorithms have diverse accuracy results which further may be instrumental to the farmers. The essence of this result throws light that choosing accurate algorithm and multidimensional dataset can entrust farmers with early crop yield predictions and recommendations. In future we will apply these algorithms for large datasets to suggest one seasonal decision support system using weather and non-weather information for Indian farmers.

REFERENCES

1. J. Liu, C. E. Goering, Lei Tian, 2001. "A neural network for setting target corn yields". Transactions of the American Society of Agricultural Engineers44 (3):705-713.
2. SamuelA.L,"Some Studies in Machine Learning Using the Game of Checkers". IBM J.Res. Dev. 1959, 44,206-226.
3. Marcello Donatelli, Amit Kumar Srivastava, Gregory Duveiller, Stefan Niemeyer and Davide Fumagalli," Climate change impact and potential adaptation strategies under alternate realizations of climate scenarios for three major crops in Europe", Environmental Research Letters, vol. 10, no. 7, Jul 2015, Art. No. 075005.
4. Rakesh Kumar, M.P. Singh, Prabhat Kumar, J.P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique",2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM),27 August 2015
5. Report on Economic Survey of Maharashtra 2012-2013, Directorate of Economics and Statistics, Planning Department, Government of Maharashtra, Mumbai (2013).
6. D. Diepeveen and L. Armstrong, "Identifying key crop performance traits using data mining" World Conference on Agriculture, Information and IT, 2008.
7. Alexander Murynin, Konstantin Gorokhovskiy and Vladimir Ignatie,"Efficiency of crop yield forecasting depending on the moment of prediction based on large remote sensing data set" retrievedfromhttp://worldcompproceedings.com/proc/p2013/DMI8036.pdf.
8. Hemageetha, N., "A survey on application of data mining techniques to analyze the soil for agricultural purpose", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp.3112-3117, 2016.
9. Wu Fan, ChenChong, GuoXiaoling, Yu Hua, Wang Juyun. Prediction of crop yield using big data. 8th International Symposium on Computational Intelligence and Design (ISCID).2015; 1, 255-260.
10. Monali Paul, Santosh K. Vishwakarma, Ashok Verma. Analysis of soil behavior and prediction of crop yield using data mining approach. Computational Intelligence and Communication Networks (CICN). 2015; 766-771.
11. Subhadra Mishra, Debahuti Mishra, GourHariSantra,"Applications of machine learning techniques in agricultural crop production: a review paper. Indian Journal of Science and Technology.2016, 9(38), 1-14
12. Kushwaha, A.K., SwetaBhattachrya, "Crop yield prediction using Agro Algorithm in Hadoop", International Journal of Computer Science and Information

Technology & Security (IJCSITS), Vol. 5- No2, pp.271-274, 2015.

13. Sujatha, R., Isakki, P., "A study on crop yield forecasting using classification techniques", International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE), pp.1-4, 2016.
14. N.Gandhi and L.J. Armstrong, "Applying data mining techniques to predict yield of rice in Humid Subtropical Climatic Zone of India", Proceedings of the 10th INDIACom-2016, 3rd 2016 IEEE International Conference on Computing for Sustainable Global Development, New Delhi, India, 16th to 18th March 2016.
15. N. Gandhi and L. Armstrong, "Rice Crop Yield forecasting of Tropical Wet and Dry climatic zone of India using data mining techniques", IEEE International Conference on Advances in Computer Applications (ICACA), pp. 357-363, 2016.
16. Shweta Srivastava, Diwakar Yagysen,"Implementaion of Genetic Algorithm for Agriculture System", International Journal of New Innovations in Engineering and Technology Volume 5 Issue 1-May 2016.
17. Shruti Mishra, Priyanka Paygude,Snehal Chaudhary, Sonali Idate, "Use of data mining in crop yield prediction",2018 2nd International Conference on Inventive Systems and Control (ICISC).
18. Rossana MC, L. D. (2013). A Prediction Model Framework for Crop Yield Prediction. Asia Pacific Industrial Engineering and Management Society Conference Proceedings Cebu, Philippines, 185.
19. R.Kalpana, N.Shanti and S.Arumugam, "A survey on data mining techniques in Agriculture", International Journal of advances in Computer Science and Technology, vol. 3, No. 8,426- 431, 2014.
20. Dr Shirin Bhanu Koduri, Loshma Gunisetti, Ch Raja Ramesh, K V Mutyalu and D. Ganesh," Prediction of crop production using AdaBoost regression Method", International conference on computer vision and machine learning, Conf. Series 1228 (2019) 012005.

AUTHORS PROFILE

Kusum Lata, Research Scholar, Department of Computing Sciences and Engineering, Sandip University, Nasik.

Dr. Sajidullah S. Khan, Associate Professor, Department of Computing Sciences and Engineering, Sandip University, Nasik.