AATCC Review

10 January 2022: Received
09 May 2022: Revised
08 July 2022: Accepted
11 September 2022: Available Online

www.aatcc.peerjournals.net

Research Article Open Access

Appraising relation between weather and Red gram yield estimates with Regression Machine Learning

Baby Akula[1*], N. Divya[2], K. Indudhar Reddy[3], and R.S.Parmar[4]

[1]Department ofAgronomy, Professor Jayashankar Telangana State Agricultural University, Hyderabad, India
[2]Chaitanya Bharathi Institute of Technology, Osman Sagar Rd, Kokapet, Gandipet, Hyderabad, Telangana 500075, IndiaIndia
[3] Professor Jayashankar Telangana State Agricultural University, Hyderabad, India
[4]College of Agricultural Information Technology, Anand Agricultural University, Gujarat, India

ABSTRACT

The frequent surge in the price of red gram as compared with other pulses necessitated seed yield estimation to cope up with demand-supply equilibrium by policy-makers and efficient resource utilization by farmers and agronomists. Besides, to achieve another main objective of appraising the relationship between red gram crop yields and weather, fivesupervised regression machine learning algorithms namely,Gaussian processes, Linear regression, Support vector machines, k-Nearest neighbors and Decision tree were used in the study. Among these tested algorithms, the Random Forest algorithm was better with crop yield predictability of 95 % (R^2), lowest Mean Absolute Error (MAE) of 32.7 and Root Mean Squared Error (RMSE) of 40.8 as compared with other fitted regression algorithms. It was further noticed that, the actual yield and the predicted yield based on training data set were close to each other and the residual ranged from -76 to 99, while it ranged from -148 to 111 in case of testing data by the same Random Forest model.

Keywords: *Gaussian Processes, Linear Regression,Support Vector Machines, k-Nearest Neighbors, Decision Tree , Red gram yield estimates*

INTRODUCTION

Red gram is the second mostimportant pulse crop of India after Bengal gram . India accounts for 65 % global seed [4]. Its ability to produce high economic yields even under rainfed conditions and being an indispensable part of Indian meals due to high protein of 22.3 % further assumes significance in yield estimation. Telangana state ranks third in red gram cultivation in an area of 2.3 L ha during 2020-2021 after Maharashtra and Karnataka (*http://www.pjtsau.edu.in*). A steep rise in support price as compared with other pulses further necessitates seed yield estimation to understand fluctuations in its production due to the vagaries in monsoon as being grown mainly as rainfed crop. In this context of meeting the local and global supply chain demand, machine learning algorithms come in handy in estimating weather based yield estimates. Regression Machine learning has been gaining popularity in agricultural applications due to its success in bioinformatics.

Crop yield prediction with machine learning techniques is the latest subject in literature and was considered for various crops like a wheat and rice [1] and groundnut [9]. Machine learning is a subset of artificial intelligence that enables an algorithm to learn from the experiences without being clearly programmed.

*Corresponding Author: **Dr. Baby Akula***
*Email Address: **babys_akula@yahoo.co.in***

DOI: https://doi.org/10.58321/AATCCReview.2022.11.04.01

Basically, machine learning can be categorized into three broadcategories namely supervised learning, unsupervised learning, and reinforcement learning. In this article, fivesupervisedregression machine learning algorithms namelyf Gaussian processes, linear regression,support vector machines, k-Nearest neighbors and decision tree wereuto build the most accurate andeffective model since the learning information occurs with required outputs and also theobjective of the study was to determine a common rule of showing input to output. Moreover, regression machine learning algorithms take a data-driven technique to learn useful models and relationships from input data [10] and provides a best way for improving crop yield predictions. In addition, regression machine learning algorithms have some individual benefits like, they can model non-linear relationships between multiple data sources [3].

As [11] studied the effect of rainfall on crop yield using the regression machine learning algorithm and reported that the Gaussian Processes model explained the good degree of relationship between annual rain fall and wheat yield. As proposed systemby [2] in order to improve crop yield using different machine learning algorithms namely back-propagation, k means clustering and random forest. The results explained that, random forest algorithm works well with small and large experimental datasets and with highprecision on evaluating with other algorithms. [6] integrated five field-based soil properties and topographic data to predict maize yield by applying various regression machine learning algorithms namely random forest, neural network, support vector machine. The result determined that a random forest always better than other fitted models. As [5] used a random forest algorithm for global and regional crop yield such as maize, wheat and environmental parameters such as soil, climate, fertilization

data etc. Results demonstrated random forest is an effective and dynamic algoritm for crop yield prediction with high accuracy and precision [7], [10]. Thus, the main objective of this study was framed to explore the possibility of suggesting a suitable regression machine learning algorithm for predicting of red gram yields in Ranga Reddy district of Telangana.

Materials And Methods

The present investigation was undertaken to appraise the relationship between weather parameters and red gram yield with regression machine learning algorithms.The average yield data for red gram over a period of 31 years i.e. 1988-2019 were collected from the Directorate of Economics and Statistics, Government of Telangana, India. The daily weather data of maximum temperature, minimum temperature, morning relative humidity, evening relative humidity, rainfall , bright sunshine, wind speed and evaporation during the crop season(30^{th} to 47^{th}Meteorological Standard Weeks) were also collected from Agro-climate Research Centre, PJTS Agriculture University, Hyderabad. These daily weather data were compiled as weekly for the purpose of analysis.

Steps involved as suggesting suitable machine learning algorithm for predicting red gram yields:

- Experimental Dataset: It was prepared in an excel sheet with a CSV extension for study by a machine learning system (Weka 3.8.5).
- Normalized Dataset: Min-max algorithm was used to normalize the dataset as it one of the most regular ways to normalize data.
- Attribute Selection: The attribute evaluator namely **"cfsSubsetEval"** and search method namely "**BestFirst**" were used as it selected those feature variables which contribute most to the prediction variable (Table 1).
- Evaluate Algorithms: The five regression machine learning algorithms (Table 2) namely Gaussian Processes, Linear Regression, Support Vector Machines, k –Nearest Neighbors and Decision Tree were then employed over the experimental data set. The results of each regression algorithm were noted and compared with each other.
- Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error and Coefficient of determination values were taken into consideration for each regression algorithm.

Results And Discussion

An open source system Weka 3.8.5 is a collection of regression machine learning algorithms for regression analysis. The regression machine learning algorithms can be applied directly to an experimental dataset. Weka has several useful regression machine learning algorithms to make crop yield predictions. All regression machine learning algorithms are usually driven by the number of feature variables, the shape of the regression line and the type of target variables. From weka regression algorithms, five machine learning algorithms are evaluated namely Gaussian Processes (GP), Linear Regression (LR), Support Vector Machines (SMOReg), k Nearest Neighbors (IBK) and Decision Tree (Random Forest). The performance of each algorithm is checked in terms of MAE, RMSE, RAE, RRSE and R^2.The Fig. 1 explains the graphical distribution of each selected attribute. It reveals that the attributes have differed distribution range.

The characteristics of fitted regression machine learning algorithms in Table 3 indicated that the tree based algorithm exhibited better performance as compared with function based algorithms and lazy based algorithm. In case offunction based algorithms, three algorithms were examined namely, Gaussian processes, linear regression andSMOReg. Among these, SMORegshowcased better performance as compared with other fitted algorithms. However, in general, it could be observed that, the highest R^2 value and lowest MAE value suggests that the fitted random forest algorithm was adequate in predicting the relationship between the weather parameters and red gram yield. Therefore, the random forest algorithm is appropriate to predict Redgram yield algorithmas an efficient and adaptable machine-learning algorithm for crop yield prediction because of its high exactness and precision, ease of use, and usefulness in data analysis..

The results were compared with multiple linear regressions and evaluated using R^2 and MAE. Results demonstrated random forest is an efficient algorithm for crop yield prediction with high accuracy and precision. [6] studied three machine learning algorithms namely random forest, neural network, and support vector machine for maize yield prediction. Their results also corroborated the finding that the random forest algorithm was consistently better as compared with other fitted algorithms. [5], [8] also opined similarly.

The Fig. 2 shows the prediction accuracy of different fitted regression machine learning algorithms. Out of five algorithms used in this research work, the Random Forest algorithm was better in crop yield predictability as compared with other fitted regression algorithms with 95 % (R^2) followed by KNN (89 %), while Gaussian Processes exhibited the lowest predictability (74%).

The Fig. 3 depicts the error results of the different regression machine learning algorithms. Random Forest exhibited lowest Mean Absolute Error (MAE) of 32.7 and Root Mean Squared Error (RMSE) of 40.8. This exposed minimal error estimated during the crop yield prediction processes. In contrast, Gaussian Processes had resulted in the highest error rate with 71.6 and 85.5 of MAE and RMSE, respectively.

A predicted yield error rate of the random forest algorithm for training and testing data set, respectively as shown in the Fig.4 and Fig. 5demonstrated that, the predicted yields were both over and underestimated for different years. In the case of a training data set, the predicted yield was underestimated by 11.9 %, 2.2 %, 8.3 % , 7.9%, 1.3 %,10.6 %, 3.9 %,10.3 %, 6.5 % , 1.8 % and 8.5 % for the years 1994,1995,1999,2001,2003,2004,2005,2006,2009,2013 and 2014 respectively. But, predicted yield were overestimated by 17.8 %, 13.0 %, 4.5 %, 5.4 %, 30.4 %, 4.8 %, 11.0 %, 5.6 %, 18.7 %, 47.8 %, 0.4 %, 15.7 % and 7.2%accordingly for the years 1988, 1989, 1990, 1991, 1992, 1996, 1997, 2000, 2002, 2007, 2010, 2011 and 2012, respectively. For the testing data set, the predicted yield was underestimated by 17.8 % and 29.2 % for the years 2015 and 2016 respectively while, overestimated by 27.9 % and 12.4 % for the years 2017 and 2018, respectively. The predicted yield error rates ranged from –11.9 % to 47.8 % for a training data set and while it ranged from – 29.2 % to 27.9 % in the case of testing data set.

The predicted yield based on the training data set is presented in Table 4. The same is demonstrated in Fig.6. It was noticed that the actual yield and the predicted yield were close to each other. The residual ranged from -76 to 99 while it ranged from -148 to 111 in case of testing data by the same Random Forest model .

Conclusion

Appraising the relationship between red gram crop yields and the weather is an important dimension of its seed yield estimation because the crop is mainly grown as a rainfed crop. Five supervised regression machine learning algorithms namely, Gaussian processes, linear regression,support vector machines, k-nearest neighbors and decision tree were used in the study as these have been gaining popularity in agricultural applications due to its success in yield estimation. Among these, the Random Forest algorithm was found be superior with crop yield predictability of 95 % (R^2), lowest Mean Absolute Error (MAE) of 32.7 and Root Mean Squared Error (RMSE) of 40.8 as compared with other fitted regression algorithms.

Future scope of work: The study can be further extrapolated and make the model robust by interfacing with GIS for better utilization by the stakeholders.

Author statement (Disclaimer): The contents and views expressed in this research paper are the views of the authors and do not necessarily reflect the views of the organizations they belong to".

Conflict of Interest: All the authors declare that there exists no conflict of interest.

Acknowledgement: The authors are thankful to Directorate of Economics and Statistics, Government of Telangana, for sparing redgram yield data and Telangana State Development and Planning Society, Govt of Telangana , Hyderabad for providing weather data.

Table 1: Detail of Selected Feature Variables

Sr. No.	Variables	Descriptions
1	MAXT47	Weekly Average of Maximum Temperature for 47th MSW
2	MINT38	Weekly Average of Minimum Temperature for 38th MSW
3	MINT39	Weekly Average of Minimum Temperature for 39th MSW
4	MINT44	Weekly Average of Minimum Temperature for 44th MSW
5	MINT47	Weekly Average of Minimum Temperature for 47th MSW
6	RHI34	Weekly Average of Morning Relative Humidity for 34th MSW
7	RHI40	Weekly Average of Morning Relative Humidity for 40th MSW
8	RHI47	Weekly Average of Morning Relative Humidity for 47th MSW
9	RHII46	Weekly Average of Afternoon Relative Humidity for 46th MSW
10	RF31	Weekly Total Rainfall for 31th MSW
11	BSS32	Weekly Average of Bright Sunshine for 32nd MSW
12	WS32	Weekly Average of Wind Speed for 32nd MSW
13	EVP36	Weekly Average of Evaporation for 36th MSW
14	EVP37	Weekly Average of Evaporation for 37th MSW
15	EVP46	Weekly Average of Evaporation for 46th MSW

Table 2: List of Regression Machine Learning Algorithms

Category	Algorithm	Description
Function Based	Gaussian Processes	It is a very useful regression machine learning algorithm non-linear multiple variate interpolation. It can be comprehensive in the future to aid in both supervised and unsupervised learning models (Rasmussen, 2004).
	Linear Regression	It is a regression algorithm model the correlations between the predicted variable and one or more predictor variables. It is a commonlyused statistical model for predicting crop yield(Sheehy *et al.*, 2006).
	Support Vector Machines (SMO Reg)	It is most powerful algorithm with strong theoretical foundations. It has strong regularization and can be used both for classification or regression challenges. It can be used under both linear and non-linear types of regression in machine learning (Saranya *et al.*, 2020).
Lazy Based	k-Nearest Neighbors (IBK)	It is widely used for non-linear regression in machine learning. It assumes that the new data point is like to the presented data points. The new data point is compared to the presented categories and is placed under a relevant category. The average value of the KNN is taken as the input in this algorithm. The neighbors in KNN algorithms are given a meticulous weight that identifies their input to the average value (Saranya *et al.*, 2020).
Tree Based	Decision Tree (Random Forest)	It is widely used for non-linear regression in machine learning.It is split the dataset into smaller sets. The splitting of the data set by this algorithm results in a decision tree that has decision and leaf nodes. Experts prefer this algorithm where there is not enough change in the data set. It has also been usedin studies to estimate crop yields (Johnson, 2014, Everingham *et al.*, 2016).

Table 3: Characteristics of Fitted Regression Machine Learning Algorithms

Parameters	Regression Machine Learning Algorithms				
	Functions Based			Lazy Based	Tree Based
	Gaussian Processes	Linear Regression	SMO Reg	k-Nearest Neighbors	Random Forest
Mean Absolute Error (MAE)	71.60	58.20	33.30	54.00	**32.70**
Root Mean Squared Error (RMSE)	85.50	72.80	67.60	70.80	**40.80**
Relative Absolute Error (RAE)	64.37 %	52.28 %	29.91 %	48.53 %	**29.36 %**
Root Relative Squared Error (RRSE)	54.75 %	46.60 %	43.31 %	45.34 %	**26.10 %**
Coefficient of Determination (R^2)	74 %	78 %	81 %	89 %	**95 %**

Table 4: Actual and Predicted Yield Using Random Forest Algorithm (Training Data Set)

Year	Actual yield (kg/ha)	Predicted Yield (kg/ha)	Residual	Year	Actual yield (kg/ha)	Estimated Yield (kg/ha)	Residual
1988	228	269	41	2003	415	410	-5
1989	243	275	32	2004	399	357	-42
1990	320	334	14	2005	369	355	-14
1991	291	307	16	2006	740	664	-76
1992	241	314	73	2007	208	307	99
1994	468	412	-56	2008	887	887	0
1995	417	408	-9	2009	620	580	-40
1996	315	330	15	2010	419	421	2
1997	309	343	34	2011	266	308	42
1999	431	395	-36	2012	284	305	21
2000	330	348	18	2013	408	401	-7
2001	433	399	-34	2014	441	404	-37
2002	283	336	53				

Table 5: Actual and Predicted Yield Using Random Forest Algorithm (Testing Data Set)

Year	Actual yield (kg/ha)	Predicted Yield (kg/ha)	Residual
2015	401	488	87
2016	268	379	111
2017	677	529	-148
2018	694	617	-77

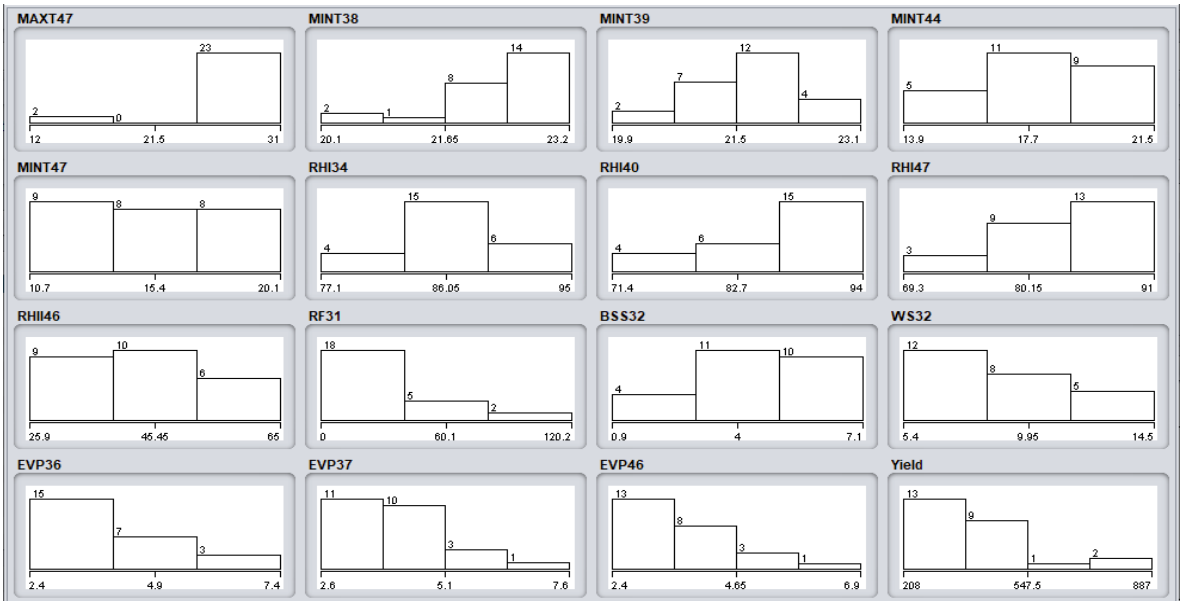

Fig. 1: Univariate Attribute Distributions

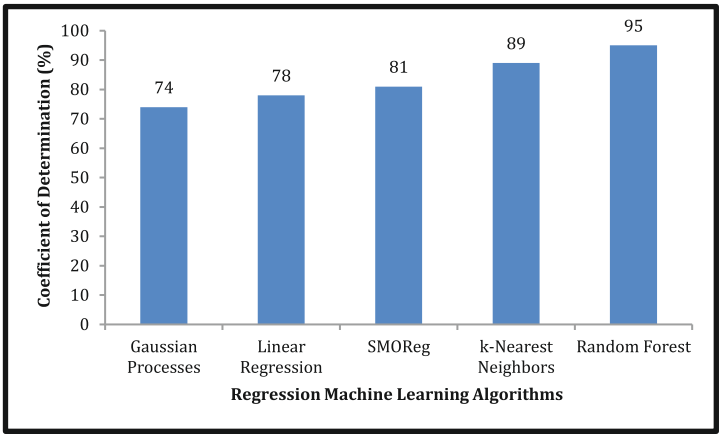

Fig.2: Coefficient of Determination (Predication Accuracy) of Fitted Regression Machine Learning Algorithms

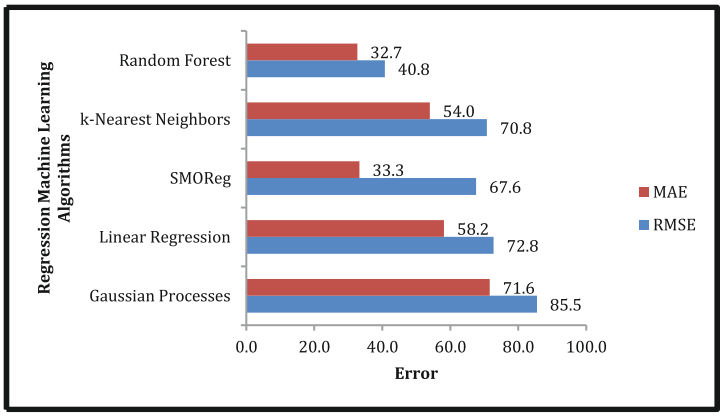

Fig.3: Error Results of FittedRegression Machine Learning Algorithms

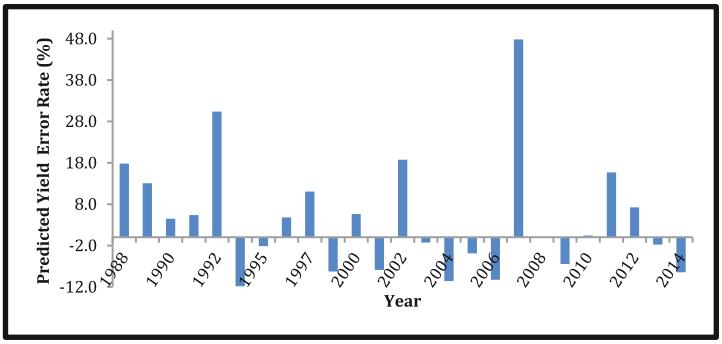

Fig.4: Predicted Yield Error Rate of Random Forest Algorithm(Training Data Set)

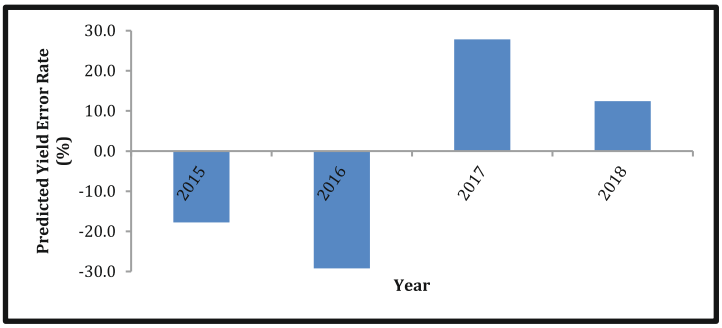

Fig .5Predicted Yield Error Rate of Random Forest Algorithm(Testing Data Set)

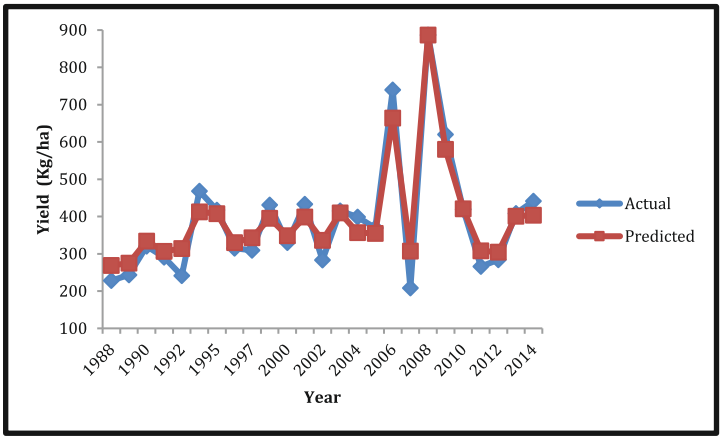

Fig .6 Actual and Predicted Yield Using Random Forest Algorithm (Training Data Set)

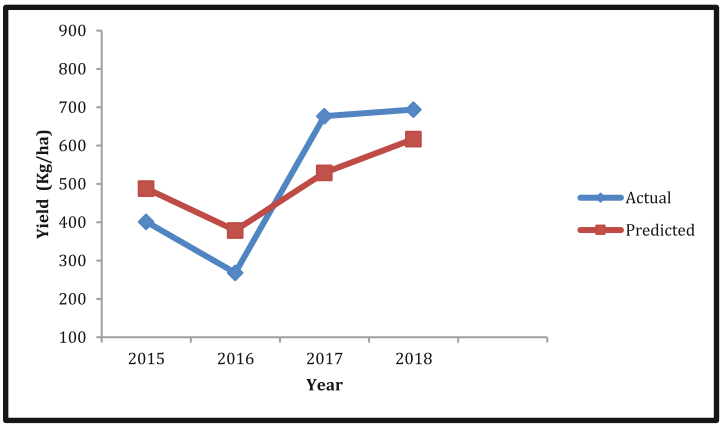

Fig .7 Actual and Predicted Yield Using Random Forest Algorithm (Testing Data Set)

References

1. Baby Akula, RS Parmar, M P.Raj, and K. Indudhar Reddy. (2021). Prediction for rice yield using data mining approach in Ranga Reddy district of Telangana. *J of Agrometeorol.* 23(2):242-248.

2. Bhanumathi, S., Vineeth, M and Rohit, N. (2019). "Crop Yield Prediction and Efficient use of Fertilizers," *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019, pp. 0769-0773, doi: 10.1109/ICCSP.2019.8698087.

3. Chlingaryan, Anna, Sukkarieh, Salah, Whelan and Brett. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture. 151. 61-69.

4. Food and Agriculture Organization Of the United States 2019. FAOSTAT Statistical Database. (http://www.fao.org/).

5. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, et al. (2016) Random Forests for Global and Regional Crop Yield Predictions. PLOS ONE 11(6): e015657.

6. Khanal, Uttam, Wilson, Clevo, Lee, Boon L., Hoang, Viet-Ngu (2018). Climate change adaptation strategies and food productivity in Nepal: a counterfactual analysis. Climatic Change. 148(4): 575–590.

7. Konstantinos G. Liakos, PatriziaBusato, DimitriosMoshou, Simon Pearson and DionysisBochtis, 2018. Machine Learning in Agriculture: A Review, Sensors: 2018,18: (8), 2674; https://doi.org/10.3390/s1808

8. Maya Gopal, P,S., Bhargavi, R. 2019. A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture.165:1-9.

9. Shah Vinita, Shah Prachi. 2018 Groundnut Crop Yield Prediction Using Machine Learning Techniques, International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307 pages 1093-1097

10. Simon Willcock Javier Martínez-López Danny A.P. Hooftman Kenneth J. Bagstad Stefano Balbi Alessia Marzo Carlo Prato Saverio Sciandrello Giovanni Signorello Brian Voigt Ferdinando Villa James M. Bullock Ioannis N. Athanasiadis. Machine learning for ecosystem services.Ecosystem Services: 33 (2018) 165–174

11. Vagh, Y. (2012). An Investigation into the effect of stochastic annual rainfall on crop yields in South Western Australia. International Journal of Information and Education Technology, 2(3), 227-232.