Artificial & Computational Intelligence / Published online: Dec 2019
http://acors.org/ijacoi/VOL1_ISSUE1_06.pdf

Artificial & Computational Intelligence

Data Analytics for Predicting Wheat Crop Production

Akanksha Gahoi
Department of Computer Science
Akansha.gupta1995@gmail.com
Madhav Institute of Technology & Science, Gwalior (M.P), India

Abstract

"Agriculture is the art and science of growing plants and other crops and for food, other human needs, or economic gain". India is an agricultural country. Two- third of the population is dependent on agriculture directly or indirectly. Agriculture playas a vital role in the economy of country but now a day's various factor affects the production rate of crop. These factors may be pesticides, rainfall, soil type, nutrients, fertilizer, pH, what type of land is suitable for a particular crop. In the field of agriculture there are many serious problems are people face in growing crop today. These problems are- Erosion, Diseases, Pests, Weeds, Drought, Rainfall. Classification and prediction techniques are applied on metrological-related data and crop related data. Various predictions can be made on the basis of predicted result which can help in increasing crop production rate. In this research work, the main focus is on analysis of prediction of wheat crop production of Madhya Pradesh state on the basis of basic parameter like area harvested, rainfall and soil type by using data mining techniques for those farmers who cannot afford laboratory test. And also build an efficient model for prediction of wheat crop production. For implementation of this framework different types of classification and prediction techniques are used. These techniques are Recursive Partitioning and Regression Tree, Random Forest, Generalized Linear Model, Neural Network and Support Vector Machine. Among these techniques Random Forest is more suitable to predict the Wheat Crop Production.

Keywords: Data Mining, Artificial Neural Network (ANN), Generalized Linear Model (GLM), Random Forest, Recursive Partitioning and regression tree (RPART), Support Vector Machine (SVM), Crop Prediction

1. INTRODUCTION

In agriculture there are many serious problems like- Erosion, Diseases, Pests, Weeds, Drought, Rainfall that people face trying to grow food today.

Classification and prediction techniques are applied on metrological-related data and crop-related data [1]. Various predictions can be made on the basis of predicted result which can help in increasing crop production rate. Data Mining plays a significant role in decision making on several issues related to the agriculture field. Applicability of Data Mining techniques in agriculture, crop planning, and crop management can change the way of farming farmers can yield in a much efficient way [4].

Data Mining is used in various field of Agriculture-

- ➢ Prediction Problem
- ➢ Diseases detection
- ➢ Optimizing the results
- ➢ Forecasting Agriculture water consumption

- Data mining is used to find the appropriate agriculture crop productivity.
- To increase the income of the farmers
- Reduce the transport cost
- To predict the climate change using previously stored data set.

2. LITERATURE REVIEW

D. Rajesh et. al. in [4] have proposed a method to extract a pattern from spatial database using k-means clustering algorithm.

A.T.M Shakil Ahamed et. al in [2] have tested few DM techniques for prediction of the annual yield of major crops in Bangladesh. In this paper, the clustering technique is used to predict the results. There are parameters such as temperature, humidity, minimum temperature; maximum temperature, average sunshine, Soil PH and salinity are used to predict the Annual crop production. K-means clustering is used by the author for recommending plant crops in the districts of Bangladesh.

Belabed Image et. al. in [3] have proposed an approach for extracting information by using data mining approaches in three domains, bioinformatics, medicine, and agriculture industry. Initially, the variables are clustered to increase the functionality, and the association rules are used between the target variables and the previously identified set of variables.

Miss Snehal S.Dahikar et. al. in [6] used ANN for crop production on the basis of soil behavior and author also justified that ANN is better classifier for crop prediction of Bajara, Soya bean, Corn, Wheat, Rice, Groundnut, Cotton, Sugarcane, Jawar.

Gang Liu et. al. in [17] have proposed a BPN model for crop yield prediction on the basis of soil parameters like moisture, N, P, K and SOM. This model also described the relationship between soil behaviours and crop yield.

Snehal S.Dahikar et. al. in [16] have proposed a model for crop yield prediction on the basis of area, production and suggest the fertilizer on the basis of soil parameters (N, P, K) using ANN.

SML Venkata Narsimhamurthy et. al. in [23] have proposed a model for crop yield prediction of Andhra Pradesh state on the basis of temperature, rainfall and production using Random Forest.

P. Priya et. al. in [18] have proposed a model to predict rice crop yield production on the basis of temperature, rainfall and area by using random forest ensemble learning approach.

P. Surya et. al. in [14] have proposed a model for North western zone of Tamilnadu state to predict the area which have highest crop yield production rate. This model predicted the production rate on the basis of harvested area and production using predictive analytic techniques.

Lei Shi et. al. in [36] have proposed a model for classification of agriculture data by using support vector machine classification algorithm. And author also compares this algorithm with nave Bayes and neural network.

3. PROPOSED METHODOLOGY

In this research paper the data set has been collected from thehttp://www.mp.gov.in.The attributes in this data set are: - District of Madhya Pradesh, Year, Area, Production, Rainfall, Soil type. An efficient model for the prediction of wheat crop production on the basis of basic parameters like as Area, Rainfall and Soil type.

For the classification and prediction of wheat crop production various types of classification techniques are used namely- RPART, GLM, ANN, SVM, and Random Forest. These classification techniques are used with multivariate regression to predict the wheat crop production.

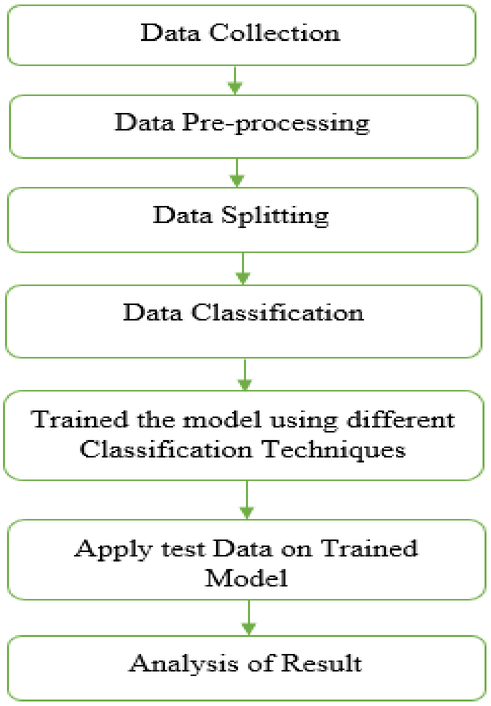
Data Collection
Data Pre-processing
Data Splitting
Data Classification
Trained the model using different Classification Techniques
Apply test Data on Trained Model
Analysis of Result

Figure 1: Work flow of proposed methodology

The classification and predictive approaches tested for wheat crop data are as follows:

a) RPART: Recursive partitioning is an analytical method for multivariable analysis purpose. Recursive partitioning generates a decision tree and also generates prediction rules by dividing data into subparts of data. [30].

Regression techniques consist a single response variable and one or more input variables. Input variable may consist categorical and continuous both types of variables. A regression tree is alternative of decision trees used for find out real valued functions. A regression tree is created by using the process of recursive partitioning [31].

b) Random Forest: Belongs to the family of supervised learning approaches, suitable for the classification and regression problems as well. Basic working ideas behind this approach are multiple collections of tree-structured classifiers. Random Forest is ensemble learning method. It is used when size of dataset is large and the very large number of input variables approximately hundreds or thousands [29].

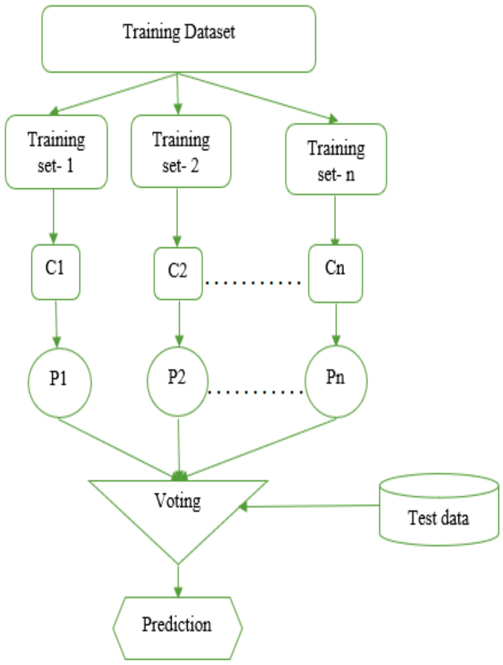

Figure 2:Random Forest

Random Forest Pseudo code-

```
1. Randomly select "k" features from total "m" feature.
        Where k<<m.
2. Among the "k" feature calculate the node "d" using the best split point.
3. Split the node into daughter node using the best split.
4. Repeat 1 to 3 steps until "1" number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.
```

Random Forest Prediction Pseudo code-

```
1. Take the next feature and use the rules of each randomly create decision tree to predict the outcome and store the predicted outcome.
2. Calculate the votes for each predicted target.
3. Consider the highest voted predicted target as the final prediction from the random forest algorithm.
```

Where K = selecting feature; m = total number of features; l = leaf node; n = randomly create tree and d = daughter node which is calculated by Gini Index.

Gini Index is calculated by: -

$$_{\text{Gini Index}}= 1- \sum_{i=1}^{C} (P_i)^2$$

c) Generalized Linear Model: Generalized linear model is a data mining classification tool. In Generalized Linear model various types of non-linear models may be tested according to the regression theory. It is supervised learning algorithm used for classification and regression [32].

d) Artificial Neural Network: An ANN often just called a neural network. ANN is an efficient and most used approach in the field of the computing system. The idea behind the ANN is based on Biological Neurons in the human brain [16].

It comprised of an interrelated group of neurons and process instruction using connectionist approach to computation. ANN has an adaptive property because neural network changed their structure according to internal information or external information passes through this network [17].

e) Support Vector Machine: SVM is a discriminatory classifier. SVM generates the hyperplane of data. SVM is a classification technique which come under the supervised learning. For generating the rules from the data, the classification and regression, SVM is a training algorithm [33]. In support vector machine classifier generates the support vectors which are data points which are lie at the boundary of an area, that distinguish one class from another. If the data can be separate into two distinguish classes or linearly separable then a unique global minimum value exist. In support vector machine various kinds of kernels can be used [34].

4. RESULTS

Figure 3 show the comparison between the actual product and predicted by using Random Forest and Figure 4 show the comparison between Actual production which is based on area harvested and the predicted production of the wheat crop on the basis of area, rainfall, and soil type by using four classifiers- RPART, Random Forest, Generalized Linear Model, Neural Network and Support Vector Machine. Figure 5 shows the accuracy between RPART, Random Forest, Generalized Linear Model, Neural Network and Support Vector Machine.

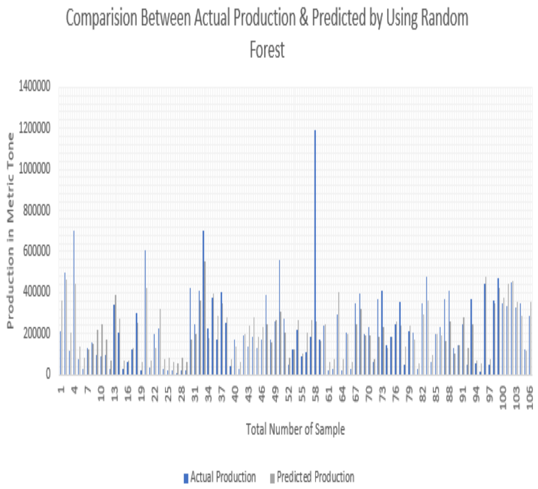

Figure 3: Comparison between Actual Production & Predicted Production by using Random Forest

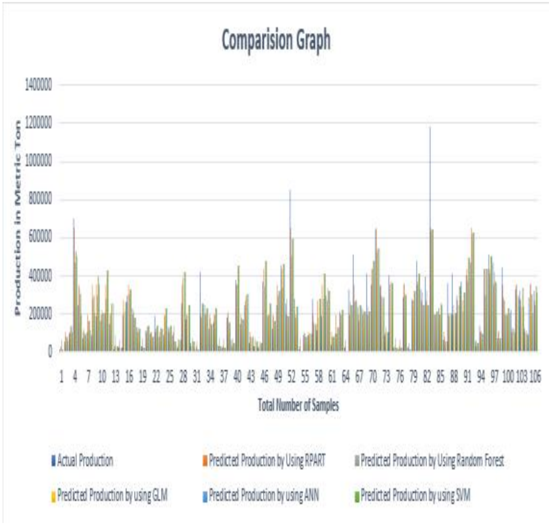

Figure 4: Comparison Graph between Actual Production and Predicted Production

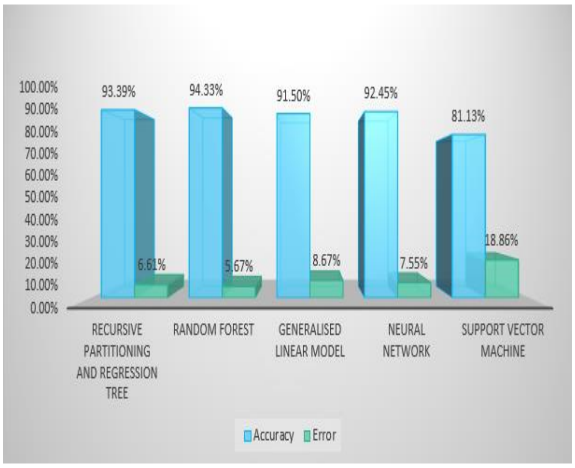

Figure 5: Accuracy Comparison

5. RESULT ANALYSIS

Table 1 shows the classification of the dataset on the basis of soil type by using the Random Forest and table 2 shows the resultant classification. Table 3 shows the Precision, Specificity and Recall of the Random Forest.

Table 1:Confusion Matrix on the basis of Soil Type

Class	Positive	Negative	Total
Positive	52	5	57
Negative	1	48	49
Total	53	53	106

Table 2: Resultant Confusion Matrix

Performance Measure:

To evaluate the performance of classifier commonly used measures are accuracy, precision, specificity and recall or sensitivity which are being described below.

a) Accuracy of Result: Accuracy defines the degree of closeness of quantitative measures to the actual value of this quantity [1]. In the context of statistics, in the context of statistics, it is defined as the relationship between the number of truly classified sample sand the actual number of samples in the class. It is given by –

$$\boldsymbol{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{TP+TN+FP+FN}$$

With the help of given confusion matrix obtained after experiment accuracy can be calculated as: -

$$\boldsymbol{Accuracy} = \frac{52+48}{52+48+5+1} \quad \mathbf{=94.33\%}$$

b) Precision: Precision is defined by the relationship as the proportion of the TP to the sum of TP and FP [1]. It is expressed as follows:

$$\boldsymbol{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

c) Recall or Sensitivity: The recall is defined by the relationship as the proportion between the number of TP and the actual number of samples which are truly exist to the positive class. It is termed as *"TPR"* [1]. It is expressed as follows:

$$\boldsymbol{Sensitivity\ or\ Recall}x = \frac{TP}{\mathrm{TP}+\mathrm{FN}}$$

Deep Medium Black	**Shallow & Medium Black**	**Alluvial Soil**	**Mixed Red & Black**	
74	0	0	0	**Deep Medium Black**
1	16	0	0	**Shallow & Medium Black**
2	0	9	0	**Alluvial Soil**
3	0	0	1	**Mixed Red & Black**

d) *Specificity:* The specificity is defined by the relationship as the ratio among the number of TN and the sum of FP and TN. It is the percentage of incorrect predicted instance. It is termed as *"TNR"* [1]. It is expressed as follows:

$$\boldsymbol{Specificity} = \frac{TN}{\mathrm{TN + FP}}$$

	Precision	TRN	TPR
RF	91%	90%	98%

Table 3: Value of Precision, Specificity or Recall

6. CONCLUSION

In this research work the research methodology and framework for wheat crop production of Madhya Pradesh. The classification technique used in this research work are Recursive partitioning and regression tree, Random forest, Generalized linear model and neural network and support vector machine. Data pre-processing steps helps to prepare the data for further processing. Also describes about the dataset for wheat crop prediction in agriculture sector i.e. used in implementation purpose with the help of random forest algorithm, and dataset is divided into four types of soil classes i.e., Deep Medium Black, Shallow &Medium Black, Alluvial Soil. Mixed Red & Black.

The result shows that random forest is more accurate in comparison of other techniques. And precision, recall and specificity also calculated.

7. FUTURE WORK

The efficiency of this model can be increased by integrating Ensemble learning algorithms in the future to get more accurate results. These classifiers can be applied for prediction of other crops produced in India and other countries as well. For improving efficiency, existing optimization techniques can also be applied after slight parameter tuning.

REFERENCES

[1] S. Mishra, P. Paygude, S. Chaudhary and S. Idate, "Use of data mining in crop yield prediction," *2nd International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, 2018, pp. 796-802.

[2] A. T. M. S. Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman, "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," *IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Takamatsu, 2015, pp. 1-6.

[3] B. Imane, B. Abdelmajid, T. A. Mohammed, T. A. Mohammed and T. A. Youssef, "Data mining approach based on clustering and association rules applicable to different fields," *International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Kenitra, 2018, pp. 1-5.

[4] D.Rajesh, “Application of Spatial Data Mining for Agriculture”, International Journal of Computer Applications (0975 – 8887) Volume 15– No.2, February 2011.

[5] Vrushali Y Kulkarni, Pradeep K Sinha, “Effective Learning and Classification using Random Forest Algorithm”, IJEIT, Volume 3, Issue 11, may 2014.

[6] Miss. Snehal S. Dahikar, Dr. Sandeep V. Rode, “Agricultural Crop Yield Prediction Using Artificial Neural Network Approach”, IJIREEICE, Vol. 2, Issue 1, January 2014.

[7] Manish Sahane, Balaji Aglave, Razaullah Khan and SanjaynSirsat, “An Overview of Data Mining Techniques Applied to Agriculture Soil Data”, International Journal of Agriculture Innovations and Research, Volume 3, No. 2, September 2014.

[8] M. Paul, S. K. Vishwakarma and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," *International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, 2015, pp. 766-771.

[9] N. Gandhi and L. Armstrong, "Applying data mining techniques to predict yield of rice in humid subtropical climatic zone of India," *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 1901-1906.

[10] Q. Ding, Q. Ding and W. Perrizo, "PARM—An Efficient Algorithm to Mine Association Rules from Spatial Data," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 6, pp. 1513-1524, Dec. 2008.

[11] N. Hemageetha, "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)*, New Delhi, 2016, pp. 3112-3117.

[12] A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada and S. Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture," *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2015, pp. 1-7.

[13] Yogesh Gandge, Sandhya, “A Study on Various Data Mining Techniques for Crop Yield Prediction”, International Conference on Electronics, Communication, Computer and Organization Techniques, 2017.

[14] P. Surya, Dr. I. Laurence Aroquiaraj, “Crop Yield Prediction in Agriculture Using Data Mining Predictive Analytic Techniques”, IJRAR December 2018.

[15] K. Pavya, Dr. B. Srinivasan, “Feature Selection Techniques in Data Mining: A Study”, IJSDR, June 2017.

[16] Snehal S. Dahikar, “An Artificial Neural Network Approach for Agriculture Crop Prediction Based on Various Parameters”, IJARECS, Volume 4, Issue 1, January 2015.

[17] Gang Liu, Xuehong Yang and Minzan Li, “An Artificial Neural Network Model for Crop Responding to Soil Parameters”, Springer 2005.

[18] P. Priya, U. Muthaiah and M. Balamurugan, “Predicting Yield of The Crop Using Machine Learning Algorithm”, IJESRT, April 2018.

[19] Georg Ruß, “Data Mining of Agricultural Yield Data: A Comparison of Regression Models”, Springer 2009.

[20] Sunita Beniwal, Jitender Arora, “Classification and Feature Selection Techniques in Data Mining”, IJERT, August 2012.

[21] Ms. Sonali. B. Maind and Ms. Priyanka Wankar, “Research Paper on Basic of Artificial Neural Network”, IJRITCC, January 2014.

[22] Eesha Goel, Er. Abhilasha, “Random Forest: A Review”, IJARCSSE, 2017.

[23] SML Venkata Narasimhamurthy, AVS Pavan Kumar, “Rice Crop Yield Forecasting Using Random Forest Algorithm”, IJRASET, October 2017.

[24] K.Bharatha Krishna and S.S.Suganva, “Application of Data Mining in Agriculture”, IJRCAR, August 2017.

[25] Dai, Qing-yun& Zhang, Chun-ping & Wu, Hao. Research of Decision Tree Classification Algorithm in Data Mining. International Journal of Database Theory and Application, 2016.

[26] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora and Naresh Dhami. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications 163(8):15-19, April 2017.

[27] Priyanka Gaur, “Neural Networks in Data Mining”, International Journal of Electronics and Computer Science Engineering, ISSN- 2277-1956.

[28] Dr. Yashpal Singh, Alok Singh Chauhan, “Neural Networks in Data mining”, Journal of Theoretical and Applied Information Technology.

[29] Eesha Goel, Er. Abhilasha, “Random Forest: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 1, January 2017.

[30] Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods, 2009.

[31] Terry M. Therneau Elizabeth J. Atkinson Mayo Foundation, “An Introduction to Recursive Partitioning Using the RPART Routines”, April 2019.

[32] Kolyshkina, Inna & Wong, Sylvia & Lim, Steven. Enhancing Generalised Linear Models with Data Mining, 2004.

[33] Nayak, Janmenjoy& Naik, Bighnaraj& Behera, Dr. H. A comprehensive survey on support vector machine in data mining tasks: Applications & challenges, 2015.

[34] Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," *12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2015.

[35] Bharati, M &, Ramageri., DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering, 2010.Han, J., &Kamber, M., Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2001.

[36] Lei Shi, QiquoDuan, Xinming Ma and Mei Weng, “The Research of Support Vector Machine in Africultural Data Classification”, IFIP, 2012.