Prediction of Crop Production in India Using Data Mining Techniques

Suvidha Jambekar
dept.of Electronics and Communication
Usha Mittal Institute of Technology
Mumbai, Maharashtra, India
suvidhajambekar94@gmail.com

Shikha Nema
H.O.D of Electronics and Communication
Usha Mittal Institute of Technology
Mumbai, Maharashtra, India
shikhanema@gmail.com

Zia Saquib
Reliance Jio Infocomm Ltd
Mumbai, Maharashtra, India
zsaquib@gmail.com

***Abstract*—Agriculture sector is the primary source of food and plays a major role for the Indian economy and employment. India is the top producer of major crops such as Rice, Wheat and Maize. Data mining is the study of extracting useful information from the data sets. The present study focuses on the application of data mining techniques to predict future production of crops such as Rice, Wheat and Maize with respect to various parameters observed during the period (1950-2013). The parameters considered for the study were rainfall, mean temperature, area under irrigation, area, production and yield. The regression algorithms used in the study were Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth). The experimental results show that the performance of Multivariate Adaptive Regression Splines (Earth) is better than Multiple Linear Regression and Random Forest Regression for Rice and Wheat dataset and the performance of Multiple Linear Regression is better than Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) for Maize dataset.**

Keywords: Data Mining, Prediction, Multiple Linear Regression, Random Forest Regression, Multivariate Adaptive Regression Splines (Earth)

I. INTRODUCTION

Data mining is the study of extracting useful and important information from large data sets [1,2,3,4]] Data mining techniques are used for prediction of future crop production which will help farmers to take most appropriate decision for their crops. Data mining and machine learning techniques are used to study the effect of various parameters and make predictions of the crop production.

India has the highest production of many crops and major crops cultivated in India are Rice, Wheat and Maize. On the basis of different cultivation seasons in India, crops can be divided into Kharif, Rabi and zaid crops. In India, Rice is a Kharif and Rabi crop, which is cultivated between (June to December) for Kharif season and (January to June) for Rabi season. Maize is a Kharif and Rabi crop, which is cultivated between (July to October) for Kharif season and (October to April) for Rabi season. Wheat is a Rabi crop, which is a spring harvest or winter crop and cultivated between (October to May)[5].

The present study focuses on the effect of various parameters like rainfall, mean temperature, area and area under irrigation on production of major crops. It also aims to predict crops production by applying various regression algorithms such as Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth).

II. LITERATURE SURVEY

The paper [6] addresses the problem of food insecurity in rural upper Egypt. It proposes a framework which would predict the amount of crops needed to satisfy the Egyptian citizens. This is done by building the prediction using Artificial Neural Network along with multilayer perception (MLP) function in WEKA. It helps to predict annual amount of major crops needed (Rice, Wheat and Beans) up to the year 2020. The paper [7] investigates deep learning techniques for weather forecasting. It makes comparative study of prediction performance of Recurrence Neural Network (RNN), Conditional Restricted Boltzmann Machine (CBRM) and Convolution Network (CN) models. The results will be helpful for weather forecasting for various application including flight navigation, agricultural prediction and tourism.

The paper [8] makes a comparative study of classification algorithms BayesNet and NaiveBayes to predict rice crop yield for Kharif season for Maharashtra state in India. This is done by selecting 27 districts of Maharashtra by considering parameters like minimum temperature, maximum temperature, average temperature, reference crop evapotranspiration, precipitation area, production and yield for the Kharif season (June to November) for the duration of five years from 1998 to 2002.The conclusion drawn at the end is that performance of BayesNet was better compared with NaiveBayes. The paper [9] makes comparative study of classification algorithms to predict rice crop yield for Kharif season of tropical wet and dry climatic zone of India. The classification algorithms have been executed in open source tool WEKA and experimental results include accuracy, sensitivity, specificity, F1 score, mean absolute error, root mean squared error, relative absolute

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

error, root relative squared error and Mathews correlation coefficient. The classification performed J48 and LADTree achieved the highest sensitivity, accuracy and specificity and Classification performed by LWL classifier shows lowest sensitivity, accuracy and specificity results.

The paper [10] investigates predictive Apriori algorithm using open source data mining tool (WEKA) for prediction of paddy yield and to analyze the effect of daily temperature and rainfall on paddy yield. The paper [11] states Multiple Linear Regression technique for estimating the future yield prediction in tea cultivation with climatic change trends observed during the period (1977-2006).The parameters selected for the study were temperature, rainfall, relative humidity, sunshine and evaporation for the four regions (South Bank, North Bank, Upper Assam and Cachar) of Assam. The conclusion drawn at the end is that tea production estimation equations developed for the four regions were validated for the future yield prediction (2007, 2009 and 2010) and developed model can be used to predict tea production for each region with precision.

III. RESEARCH METHODS

This includes details of the data sets, data preprocessing, data mining techniques and building prediction model using Scikit-learn and py-earth

A. Dataset used

All the datasets were collected from the publically available records of the Indian government for the duration of 64 years from 1950 to 2013. It consists of monthly rainfall, monthly mean temperature, area under irrigation, area, production and yield for the (1) Rice-Kharif season (June to December) and Rabi Season (January to June) (2) wheat-Rabi Season (October to May) (3) Maize- Kharif season (July to October) and Rabi season (October to April).

- Rainfall (mm): The monthly rainfall for Rice, Wheat and Maize for Kharif and Rabi season for every year in India was considered for the present study.
- Mean temperature (degree Celsius): The monthly mean temperature for Rice, Wheat and Maize for Kharif and Rabi season for every year in India was considered for the present study.
- Area (Million Hectares): The total cultivated area for Kharif and Rabi season for Rice, Wheat and Maize for every year in India was considered for the present study.
- An area under irrigation (%): An area under irrigation for Rice, Wheat and Maize for Kharif and Rabi season for every year in India was considered for the present study.
- Production (Million Tonnes): The total production for above cultivated area for Kharif and Rabi season for Rice, Wheat and Maize for every year in India was considered for the present study.
- Yield (Tonnes/Hectares): Depending on the Rice, Wheat and Maize production and area cultivated in Kharif and Rabi season, calculated yield for every year was considered for the present study.

B. Dataset preprocessing

Dataset preprocessing was done to collect all the datasets in Microsoft Office Excel. For every crop it consists of following columns: year, area, production, yield, area under irrigation, monthly rainfall and monthly mean temperature. When we divide Production by Area, we get values very close to Yield. we should take either production or yield column. Considering Production and removing Yield the file was saved as .csv.

C. Data mining techniques

Regression analysis is a predictive modeling technique which estimates linear relationship between dependent variable and one or more independent variables. The dependent variable is also called as predictant and independent variables is also called as predictors. Considering production as dependent (target) variable and area, area under irrigation, monthly rainfall and monthly mean temperature as independent (feature) variables, regression analysis was conducted using 1950-2013 (64 years) data for Rice, Wheat and Maize. Regression algorithms used in this research were Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth)

D. Building prediction model using Scikit-learn and py-earth

This section discusses building of prediction model for Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) using scikit-learn and py-earth. Scikit-learn is an open source tool for Data mining and Data analysis via consistent interface in python It is licensed under simplified BSD license and provides classification, clustering, and regression algorithms [12].The py-earth package is a python implementation of Multivariate Adaptive Regression Splines (MARS) algorithm which is introduced by Jerome H. Friedman's [13].The term 'MARS' is licensed and trademarked to salford system. In order to avoid trademark infringement Multivariate Adaptive Regression Splines(MARS) is called as Earth[14].Fig 1shows the steps for building a prediction model for Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines(Earth) using Sci-kit learn and py-earth.

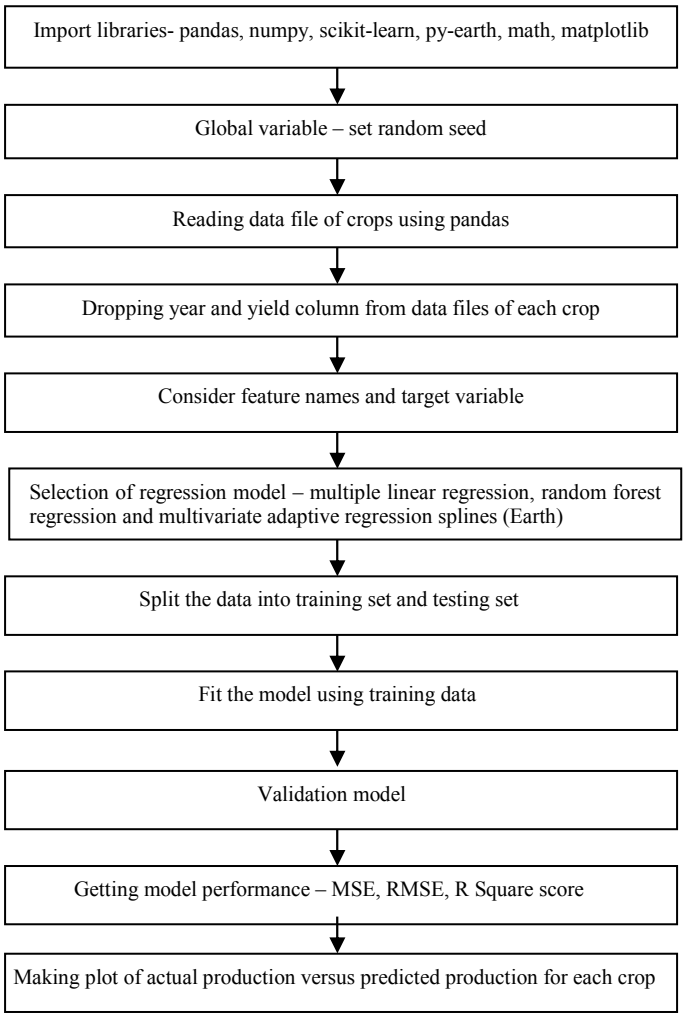

Fig. 1. Steps for building a prediction model

IV. Experimental Results

This section presents the results obtained after performing Multiple Linear Regression, Random Forest Regression and Earth algorithms on the dataset of Rice, Wheat and Maize crop. The result for each regression algorithms for each crop is shown below.

A. Rice

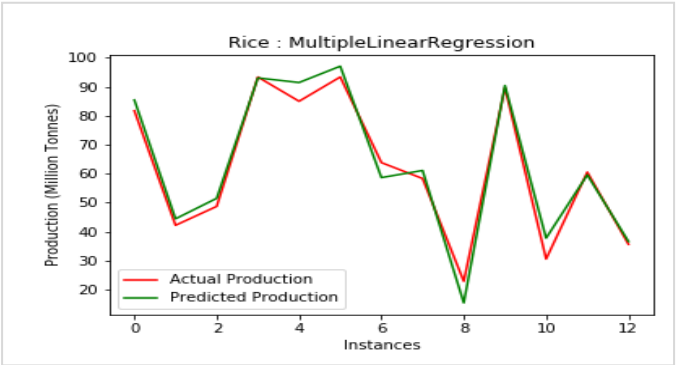

Fig. 2. Multiple Linear Regression - Comparision between Actual Production and Predicted Production

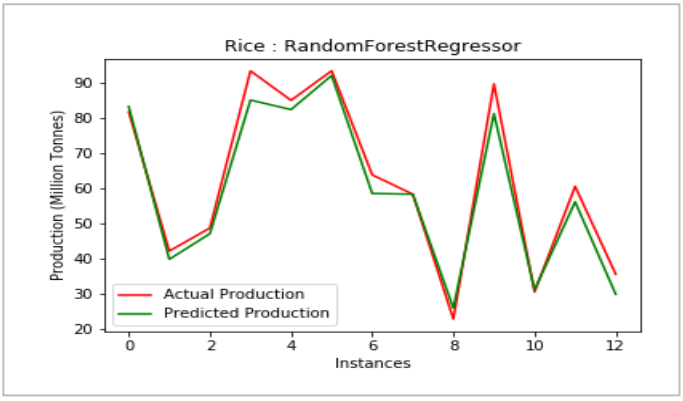

Fig. 3. Random Forest Regression - Comparision between Actual Production and Predicted Production

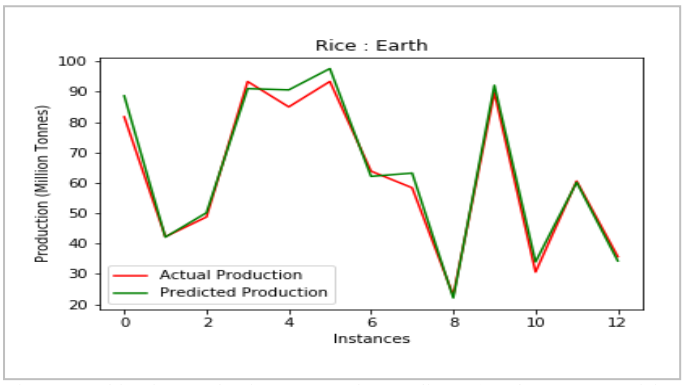

Fig. 4. Multivariate Adaptive Regression Splines (Earth) - Comparision between Actual Production and Predicted Production

B. Wheat

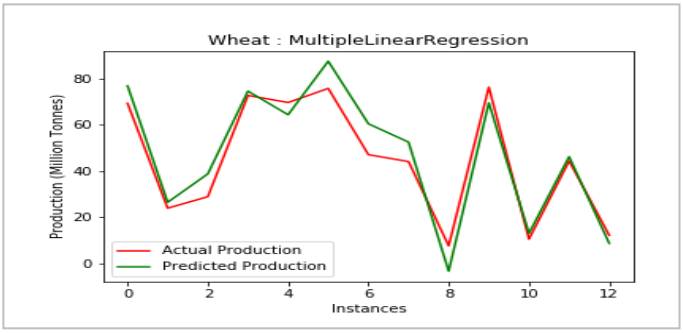

Fig. 5. Multiple Linear Regression - Comparision between Actual Production and Predicted Production

Fig. 6. Random Forest Regression - Comparision between Actual Production and Predicted Production

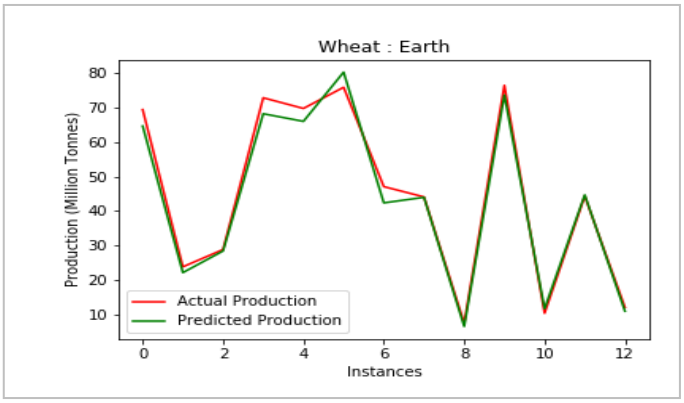

Fig. 7. Multivariate Adaptive Regression Splines (Earth) - Comparision between Actual Production and Predicted Production

C. *Maize*

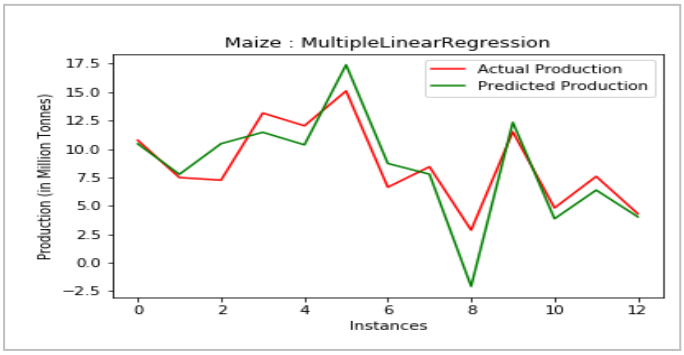

Fig. 8. Multiple Linear Regression - Comparision between Actual Production and Predicted Production

Fig. 9. Random Foresst Regression - Comparision between Actual Production and Predicted Production

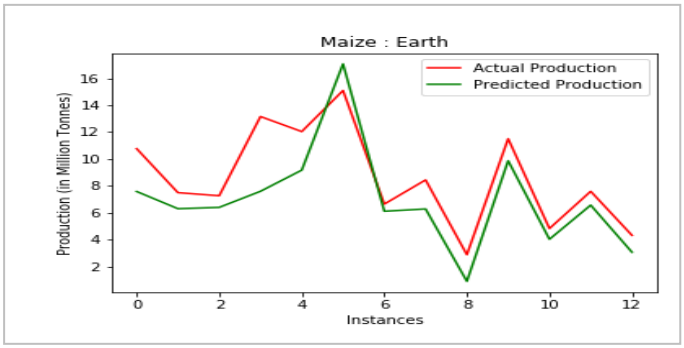

Fig. 10. Multivariate Adaptive Regression Splines (Earth) - Comparision between Actual Production and Predicted Production

Mean Squared Error (MSE) and Root mean squared Error (RMSE) are the evaluation metrics used in regression analysis. R square score is a statistical measure which is also called as the coefficient of determination. The range of r square score is always between 0 to 1 and the higher score is an indicator of better goodness of fit for the data points. Table 1 below shows Mean Squared Error (MSE) for Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) for each crop. Table 2 below shows Root Mean Squared Error (RMSE) for Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) for each crop Table 3 below shows R square score for Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) for each crop.

TABLE 1 COMPARISON OF MULTIPLE LINEAR REGRESSION (MLR), RANDOM FOREST REGRESSION (RFR), MARS (EARTH) BASED ON MEAN SAQURED ERROR (MSE)

Name of the Crop	MLR	RFR	MARS
Rice	17.38	23.32	11.60
Wheat	59.48	14.97	8.75
Maize	4.13	6.10	5.41

TABLE 2 COMPARISON OF MULTIPLE LINEAR REGRESSION (MLR), RANDOM FOREST REGRESSION (RFR), MARS (EARTH) BASED ON ROOT MEAN SAQURED ERROR (RMSE)

Name of the Crop	MLR	RFR	MARS
Rice	4.16	4.82	3.40
Wheat	7.71	3.87	2.95
Maize	2.03	2.46	2.32

TABLE 3 COMPARISON OF MULTIPLE LINEAR REGRESSION (MLR), RANDOM FOREST REGRESSION (RFR), MARS (EARTH) BASED ON R SQUARE SCORE

Name of the Crop	MLR	RFR	MARS
Rice	0.97	0.96	0.98
Wheat	0.92	0.97	0.98
Maize	0.80	0.59	0.60

V. CONCLUSION

Regression analysis was used as a predictive modeling technique for prediction of crop production. Three regression algorithms namely Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) have been used. The experimental results showed that the performance of Multivariate Adaptive Regression Splines (Earth) was better compared to Multiple Linear Regression and Random Forest Regression on the Rice and Wheat dataset and the performance of Multiple Linear Regression was better compared to Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) on the maize dataset

The results of Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) show that the regression analysis can be used to predict production of Rice, Wheat and Maize with precision. Accurate forecasts of these parameters would result in accurate production forecast in the future. Hence Data mining

techniques will be used for decision making in the agriculture sector.

REFERENCES

[1] J. Abello, P.M. Pardalos, M. Resende, Handbook of massive data sets,Kluwer, New York, 2002

[2] W. Klosgen, J.M. Zytkow, Handbook of data mining and knowledge discovery,Oxford University Press, 2002.

[3] P.M. Pardalos, L.V. Boginski, A. Vazacopoulos, Data mining in biomedicine, Springer, New York, 2007.

[4] P.M. Pardalos, P. Hansen, Data mining and mathematical programming, American Mathematical Society, USA, 2008.

[5] Crop cultivation season retrieved from Agriculture Market Information System (AMIS) crop calender

[6] Aymen E Khedr, Mona Kadry, Ghada Walid(2015), "Proposed framework for Implementing Data Mining Techniques to enhance Decisions in Agriculture sector Applied case on Food Security Information Center Ministry of Agriculture,Egypt",international conference on Communication, management and Information technology (ICCM

[7] Afan Galih Salman, Bayu Kanigoro, Yaya Heryadi, "Wheather forecasting using deep learning Techniques", 2015 International Conference on Advance Computer science and Information Systems(ICACSIS)

[8] Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong, "Predicting Rice crop yield using Bayesian Networks", 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India.

[9] Niketa Gandhi, Leisa J. Armstrong "Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India using data mining Techniques", 2016 IEEE International Conference on Advances in Computer Applications (ICACA)

[10] Kuljit kaur, Kalnwalpreet singh Attwal "Effect of Temperature and Rainfall on Paddy Yield using data mining ", 2017 7th international conference on Cloud Computing, Data Science and Engineering-Confluence

[11] Rupanjali D. Baruah, R.M. Bhagat, Sudipta Roy, L.N. Sethi "Use of Data Mining Technique for Prediction of tea yield in the face of Climate Change of Assam, India ", 2016 International Conference on Information Technology

[12] Scikit-learn, Machine learning in python retrieved from http: //scikit-learn. org

[13] Py-earth retrieved from https: //contrib.scikit-learn. org/py-earth

[14] Multivariate adaptive regression spines retrieved from https: //en.wikipedia. org/wiki/Multivariate_adaptive_regression_splines