2019 IEEE Pune Section International Conference (PuneCon)
MIT World Peace University, Pune, India. Dec 18-20, 2019

A Machine Learning Approach to Predict Crop Yield and Success Rate

Shivani S. Kale,
RRC, VTU, Belgaum, Karnataka.
shivanikale33@gmail.com

Preeti S. Patil,
IT Department,
D. Y. Patil College of engineering Akurdi,
Pune, Maharashtra
dr.preetipatil.dypa@gmail.com

Abstract- **In India agriculture contributes approximately 23% of GDP and employed workforce percentage is 59%. India is the second-largest producer of agriculture crops. the technological contribution may help the farmer to get more yield. The prediction of the yield of different crops may help the farmer regarding taking the decision about which crop to grow. The research focuses on the prediction of different crops yield using neural network regression modeling. The data of crop cycle for summer, Kharif, rabi, autumn and whole year is used. The dataset is resourced from an Indian government website. The experimental parameters considered for study are cultivation area, crop, state, district, season, year and production or yield for the period of 1998 to 2014. The dataset consists of 2 lakh 40 thousand records. The dataset is filtered using Python Pandas and Pandas Profiling tools to retrieve data for Maharashtra state. The model is developed using a Multilayer perceptron neural network. Initially the result obtained considering optimizer RMS prop with accuracy 45 %, later it will be enhanced to 90% by increasing layers, adjusting weight, bias and changing optimizer to Adam. This research describes the development of a different crop yield prediction model with ANN, with 3 Layer Neural Network. The ANN model develops a formula to ascertain the relationship using a large number of input and output examples, to establish model for yield predictions an Activation function: Rectified Linear activation unit (Relu) is used. The backward and forward propagation techniques are used.**

Keywords-Indian agriculture dataset, Neural network, Machine learning, linear regression, multiple regression, Relu-activation function, Crop Yield.

I. INTRODUCTION

Forecasting yield of crops will surely help the farmer. The farmer can make a decision about crop choice and can contribute more to its profit. There is a large number of crop yield prediction models available which may use weather real parameters or static parameters. Machine learning is found to be a very appealing field that can contribute to the agriculture field. The different models built using machine learning can take different crisp inputs to give some concrete output.

This research proposes the Neural Network model to predict crop yield and success rate of crop depending on the dataset provided by the Indian government. The dataset is huge containing data for all the regions of India which were filtered to get data for Maharashtra state i.e. 12000 records.The crop yield prediction model uses backpropagation algorithm of Artificial neural network. A multilayer perceptron technique is used.

The aim of the research is the development of the crop yield prediction model by considering data for 10 districts of Maharashtra for approximately 20 crops.

II. LITERATURE SURVEY

In this paper, the author has discussed effect of weather conditions on crop yield. The paper focuses on artificial neural network technology. The parameters used are sensor parameters such as type of soil, Ph value, N, P, K values, etc.[1].Multilayer perceptron model is developed by using neural network. The accuracy of the model is validated using cross-validation. The weka tool is used for execution. The accuracy obtained is 97.5%. Performance summarization is shown using ROC FIGURE[2]. Appropriate pesticide and insecticide suggestion are given prior. Main aim is comparison of ANN and CNN algorithm for better forecasting of cropyield[3]. Paper focuses on description of different number of agronomic based models. Models have used artificial neural network algorithm. This model focuses on development of crop [3]. Assessment of Loss and usage of insecticide, nutrient and pesticide [4], Estimation of retention of water by soil [5], similarly prediction of disease [6]. Crop yield prediction using aerial pictures have been utilized for taking decision-related harvesting [7]. Artificial neural network model gives accurate and reliable results for prediction of crop than simple linear regression model. [8]. ANN models compared to traditional statistical methods found better for predication of soybean rust by [6]. The backpropagation network model of ANN is used to predict rice yield by considering weather data. [13]. 14.8% testing error for maize yield prediction is obtained for model executed on parameters like soil, rainfall [15]. Prediction of rice yield by applying used neural networks produce testing error of 17.3% [16].

This paper reports on the use of Artificial Neural Networks to predict the rice crop yield for Maharashtra state, India.

The proposed work focuses on the use of Artificial Neural Network to predict the crop production to help the farmer to make crop choice for harvesting. The aim of the research is

1. Performance evaluation check of Artificial Neural Network multilayer perceptron model used for prediction of crop yield.

2. Finding the relationship of parameters to the accuracy and improvement to accuracy by the effect of addition and removing of parameters considered while experimenting.

978-1-7281-1924-3/19/$31.00 ©2019 IEEE

III. STUDY AREA

Maharashtra state was considered for this research as a study area. The longitude and latitude parameters for Maharashtra are 19.7515° N, 75.7139° E. This state stands third with respect to population and area in India [21]. Maharashtra is popularly known for farming and agriculture food products grown in the state. Major agriculture products of Maharashtra constitute groceries goods such as Bajara, Jowar, wheat, rice, and protein-rich pulses [21]. The state is divided into 6 regulatory divisions, Those are after that divided into thirty five districts, next it is divided into 109 sub-divisions and finally forms three hundred fifty-seven talukas [21]. 'Aurangabad', 'Beed', 'Hingoli', 'Jalana', 'Latur', 'Nanded', 'Osmanabad'.

IV. DATASET

The data used for research is resourced from the Indian government website. The dataset is publicly available for research and academic use. The dataset consists of records from the year 1997 to 2014. The present study uses the following parameters for the experiment.

Crop – The dataset contains a number of crops such as Sunflower, Bajara, Jowar, Season, groundnut, rice, cottonseed ,tur etc.

State-Maharashtra

District- 'AURANGABAD', 'BEED', 'HINGOLI', 'JALANA', 'LATUR', 'NANDED', 'OSMANABAD'.

Season- Kharif, Rabi, autumn, whole year

Year- 1997 to 2014

Production- It is given in tons per hector in lakh

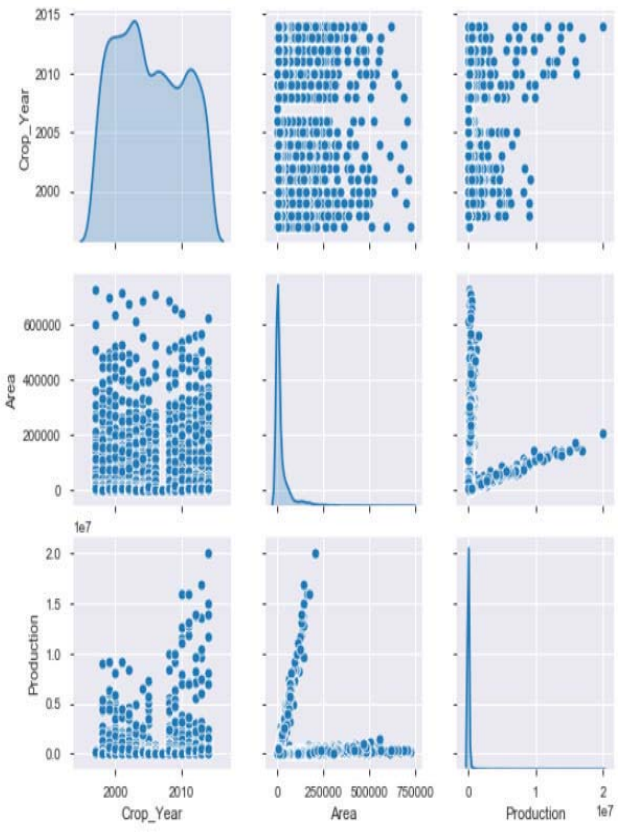

Fig. 1. Skewness FIGURE for parameters production, Area and Crop_year

The dataset given by the government contains data which was analyzed for outlier and noise. The variables also converted to categorical and numerical forms as per requirement for the model.

So the Figure1 gives the skewness of dataset as follows

if the Skewness of the dataset is equal to zero means the data is normally distributed so model will give good accuracy.If skewness is greater than zero the data is skewed more on the left trail of data distribution.If the skewness is less than zero data is distributed to right side of TABLE.

Following are the skewness FIGURE used for the study of each parameter.

V. METHODOLOGY

The Proposed study was conducted using Python **matplotlib** and **Seaborn** which is used for data visualization. Data Pre-processing and Data cleaning processes are performed by **Pandas** library of python. Basically broad five steps are used for experiment that are

1. Data collection
2. data Wrangling
3. data Pre-processing
4. data Visualization (Different Visualization Library Used)
5.explotary data analysis (EDA)

The above steps are further explained in detail as follows which are followed for processing and preparing the data for applying the multilayer perceptron technique.

Step 1: **Acquiring each parameter (Area, Crop, State, District, Season, Year and Production**) monthly mean records of each district from 1998 to 2002 from the Indian Government records.

Step 2: Selecting a subset of the **dataset for the state of Maharashtra**

Step 3: Calculating the **Number of records Based on Year with Different categorical Variables**

Step 4: Visualizing a dataset with **matplotlib** and **Seaborn python library**

Step 5: Extracting each districts area, production details of the year 1997 to 2014 from the publicly available Indian Government records for Maharashtra state.

Step 6 : Then row data were passed through different data Pre-processing.

Step 7: Using statistical methods the relation between different variables like are analyzed.

When Variables are Categorical used – Frequency, Count

When variable is Numerical used – mean, median, mode, standard deviation

Exploration of variables or attributes can be done using univariate analysis. Type of variable could be either categorical or numerical. Investigation of each type of variable can be carried by different statistical and visualization techniques. Descretization or Binning process is used for transforming numerical variables to categorical variables. On the other hand , transformation of categorical

variable to numerical variable is executed by a process of encoding.

The study in the paper uses both binning and encoding for analysis. Such as follows for the parameter District Name

```
District_Name
AURANGABAD      934300
BEED           1152900
HINGOLI         516200
JALNA           642101
LATUR           784200
NANDED          896400
OSMANABAD       859500
Name: Area, dtype: int64
```

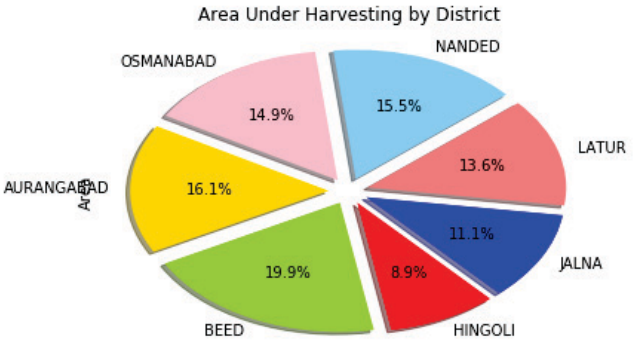

Fig. 2. Area under harvesting by considered district

The formulas used to find out the relationship between variables are used in the study as follows

Linear correlation equation is used to find out relationship between two variables. To check the dependency of variable on each other this correlation is used. This relation is also known as "Pearson's Correlation", The value of linear correlation ranges between 1 to -1 and it is denoted by 'r' relationship between variables is stated by the value of 'r'.

If r is equal to 1, variables are having strong relationship and if value comes near to -1, variables are having negative relationship. If the correlation value is equal to zero, there is no relationship between variables.

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \quad (1)$$

where

Number of observation denoted by "n". Input and output variables are "x_i" and "y_i".

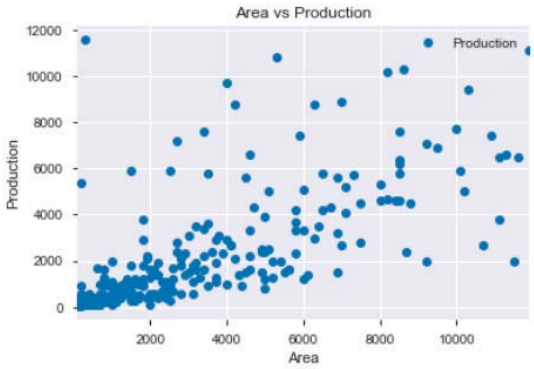

Fig. 3. Analyzing Area parameter Vs Production

Figure 3 shows distribution of production over area given.

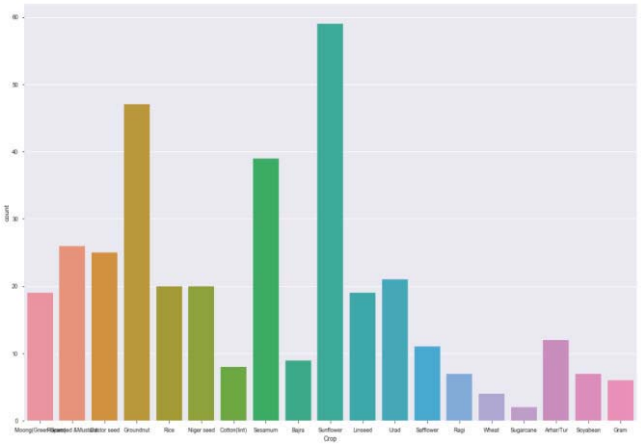

Fig. 4. Analyzing Crops and their count

Crops analysis for number of records for every crop so as to decide parameter priority for accuracy improvement.

Step 8: For preparing the data set for applying multilayer perceptron technique, unwanted columns were removed. They were sr. no, name of the district and year.

Step 9: The data set was then sorted on the basis of area. Area less than 100 hectares were not considered for the present research. So those records were omitted.

Step 10: the data which is present in **label from converted to encoding using sklearn** .

Step 11: The dataset was then sorted on the basis production.

Step 12: we considered production as output parameter and **features like: crop,area,district,season**

Step 13: This data set was then saved in .csv format for further application of the multilayer perceptron technique in **Python TensorFlow**.

Step 14: Model is trained **Using linear Regression with Neural Network** with Adam optimizer and 3 layers.

Step 15: The dataset was **Split** into dev-test and train set using 80:20 assignment.

VI. Training Parameters

Following are key parameters considered for experimentation

- batch_size=10,
- epochs=100
- Layer : 3
- Neuron at each layer : Layer 1, Layer 2 = 20
  Layer 3 = 1
- Oprimizer = adam
- Activation : Relu
- kernel_initializer='uniform
- lr rate : 0.01

VII. Performance Evaluation

The performance of the algorithm is evaluated using three metrics as follows

A. *Mean Absolute Error (MAE)*

Difference between two continuous variable is measured by Mean Absolute Error method. Suppose two variable X and Y are observation value of same phenomenon, so when we compare Y versus X states that predicted versus observed.

Formula for Mean Absolute Error is as follows:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

B. *Mean Squared error*

Mean squared error or mean squared deviation is used to measure average of squared error of estimated value and actual value.MSE is always positive. It is risk function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

C. *Root Mean Squared Error*

Difference between values predicted by model and observed value is measured by RMSE. The equation for RMSE is given as follows.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

VIII. Overview of Artificial Neural Network

Artificial Neural network is the system designed to work like a human brain. The computational system inspired by but not identical to the human brain. The ANN systems or model learn from the facts and experiences feed to them and will react to the input situation depending on training given to it. As we add more training it will work more efficiently. The artificial Neural network for our study comprised of three-layer. The circle represents nodes and arrow represents the relationship between the input layer and the output layer.

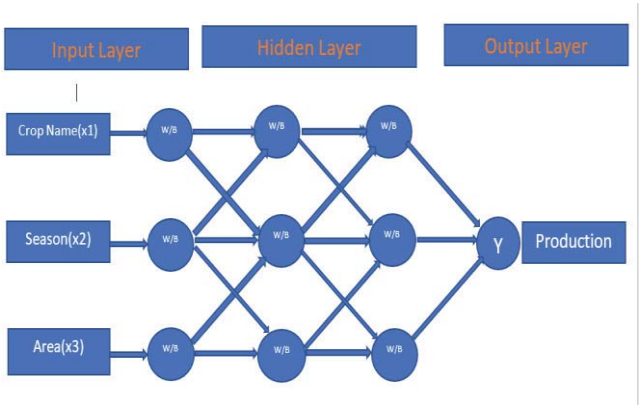

Fig. 5. ANN Model of the Proposed Study

IX. Experimental Results

Using ANN with linear regression with forward and backward propagation the model predicted the dependent variable with 82 % accuracy and very little loss.

The model will able to predict better in the real-time dataset which will be a more effective suggestion to the farmer for making crop choice.

Experimental result of ANN Multiceptron model is shown below.

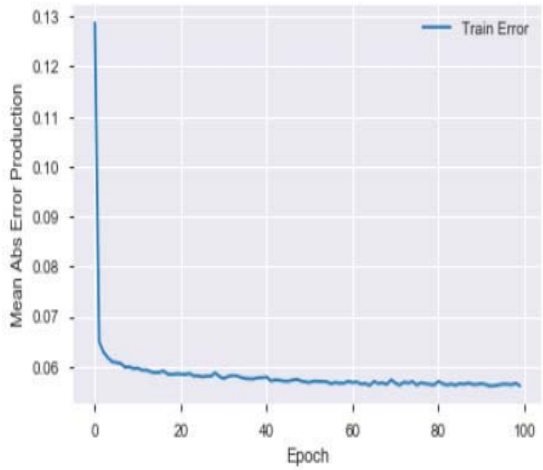

Fig. 6. MAE

The Mean Absolute error with every epoch is changing. The error value is very less so we stopped at 100 epoch.

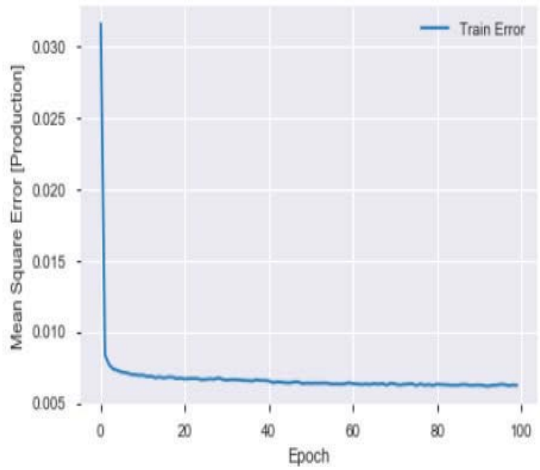

Fig. 7. MSE

The Mean squared error is very less which fit the regression line very close. The optimum regression line equation generated.

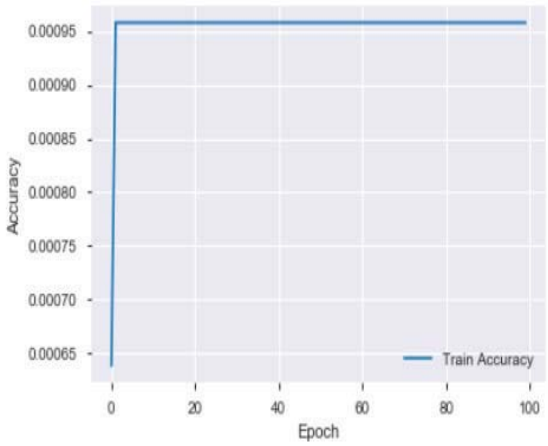

Fig. 8. Accuracy

The FIGURE shows that with every epoch the error is decreasing and the accuracy is increasing.

Discussion and Conclusion

The proposed model with backpropagation is trying to reduce MSE by using RELU activation function and gradient decesent. The Learning rate for each layer is kept constant i.e. 0.001.As we increase number of epochs error will get reduced. This result is used as input for deciding success rate of the crop over another crop. The best crop will be suggested to the farmer depending on the district and weather.

A non-linear technique is requisite to understand the association to find out the interactions between different parameters which are directly or indirectly affecting the cropyield. Due to complexity and severity of crop parameters a linear methodology is insufficient to conclude relationship between factors and crop yield. Traditional linear regression can be replaced with Artificial Neural Network methods to give better accuracy for crop prediction.

crop yield prediction for Maharashtra, India can be performed more accurately by ANN models. The ANN algorithm can be improved further to give more accurate prediction by adding layers, and adding some more parameters. This yield prediction surely going to help farmers for better decision making regarding crop harvesting.

The model also suggest success rate for the crops as per input given by the farmer. So model suggest best possible crop with highest success rate.

REFERENCES

[1] Y.A. Pachepsky, D. Timlin, and G. Várallyay, "Artificial neural networks to estimate soil water retention from easily measurable data", Soil Science Society of America Journal, vol. 60(3), pp.727-733, 1996.

[2] D.A. Elizondo, R.W. McClendon and G. Hoogenboom, "Neural network models for predicting flowering and physiological maturity of soybean", Transactions of the ASAE, vol. 37(3), pp.981-988, 1994.

[3] D.A. Elizondo, R.W. McClendon, and G. Hoogenboom, "Neural network models for predicting flowering and physiological maturity of soybean", Transactions of the ASAE, vol. 37, pp. 981–988, 1994.

[4] C.C. Yang, S.O. Prasher, S. Sreekanth, N.K. Patni and L. Masse, "An artificial neural network model for simulating pesticide concentrations in soil", Transactions of the ASAE, vol. 40, pp. 1285–1294, 1997.

[5] M.G. Schaap and W. Bouten, "Modeling water retention curves of sandy soils using neural networks", Water Resources Research, vol. 32, pp. 3033–3040, 1996.

[6] W.D. Batchelor, X.B. Yang and A.T. Tshanz, "Development of a neural network for soybean rust epidemics", Transactions of the ASAE, vol. 40, pp. 247–252, 1997.

[7] G.A. Thomas, G. Taylor and J.C.Wood, "Mapping yield potential with remote sensing", Precision Agriculture, vol. 1, pp.713–720, 1997.

[8] M. Kaul, R.L. Hill and C. Walthall, "Artificial neural network for corn and soybean prediction", Agricultural System, vol. 85, pp. 1-18, 2005.

[9] B. Ji, "Artificial neural networks for rice yield prediction in mountainous regions", Journal of Agricultural Science, vol. 145, pp. 249–261, 2007.

[10] B.A. Smith, G. Hoogenboom and R.W. McClendon, "Artificial Neural Networks for Automated Year round Temperature Prediction", Computers and Electronics in Agriculture, vol. 68, pp. 52–6, 2009.

[11] M. Schaap and W. Bouten, "Modeling water retention curves of sandy soils using neural networks", Water Resour. Res., vol.32, pp. 3033–3040, 1996

[12] S.K. Starrett and G.L. Adams, "Using artifi - cial neural networks and regression to predict percentage of applied nitrogen leached under turfgrass", Commun. Soil Sci. Plant Anal., vol. 28, pp.497–507, 1997.

[13] J. Liu, C. Goering and L. Tian, "A neural network for setting target corn yields", Transaction of the ASAE, 44(3), pp. 705-713, 2001.

[14] M. O'Neal, B. Engel, D. Ess, J. Frankenberger, "Neural network prediction of maize yield using alternative data coding algorithms", Biosystems Engineering, 83(1), pp. 31-45, 2002.

[15] S. Puteh, M. Rizon, M. Juhari, J. Nor Khairah, S. Siti Kamarudin, B. Aryati, R. Nursalasawati, "Back propagation algorithm for rice yield prediction", Proceedings of 9th of the Ninth International Symposium on Artificial Life and Robotics, Beppu, Japan, Oita, pp. 586-589, 2004.

[16] W. Ji and J. Cui, "Application of GeoTABLEical Information System (GIS) in Agricultural Land classification and Grading", 2011 International Conference on Agricultural and Natural Resources and Engineering Advances in Biomedical Engineering, 3(5), pp. 201-205, 2011.

[17] B. Jietal, "Artificial neural networks for rice yield prediction in mountainous regions", Journal of Agricultural Science, 145, pp. 249–261, 2007.

[18] T. Ranjeet and L. Armstrong, "An architecture of a decision support system for Western Australian Agriculture Industry", Proceedings of the 9th Conference of the Asian Federation for Information Technology in Agriculture "ICT's for future Economic and Sustainable Agricultural Systems", Perth, Australia, 2014.

[19] S. Jabjone and S. Wannasang, "Decision Support System Using Artificial Neural Network to Predict Rice Production in Phimai District, Thailand", International Journal of Computer and Electrical Engineering, 6(2), pp. 162-166, 2014.

[20] S. Dahikar and S. Rode, "Agricultural crop yield prediction using artificial neural network approach", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 2(1), pp. 683-686, 2014.

[21] Report on Economic Survey of Maharashtra 2012-2013, Directorate of Economics and Statistics, Planning Department, Government of Maharashtra, Mumbai (2013).

[22] Weka 3:Data Mining Software in Java, Machine Learning Group at the University of Waikato, Official Web:

http://www.cs.waikato.ac.nz/ml/weka/index.html, accessed on 26th March 2016.

[23] A. Abraham, Artificial Neural Networks: Handbook of Measuring System Design, John Wiley & Sons, Ltd., 2005,ISBN: 0-470-02143-8.

[24] N.Gandhi, L.J. Armstrong and O. Petkar, "Predicting Rice Crop Yield Using Bayesian Networks", communicated, 2016.

[25] N.Gandhi, L.J. Armstrong and O. Petkar, "Rice Crop Yield Prediction in India using Machine Learning Techniques", 2016.

[26] Mrs. Shivani S. Kale and Dr. Preeti Patil , "Data mining technology with fuzzy logic ,neural networks and machine learning for agriculture", at ASIC book series "Data management ,analytics and innovation by springer 2018.

[27] Mrs. Shivani S. Kale and Dr. Preeti Patil,"Use of data mining technology in agriculture sustainable development",IJCRT,Volume 6,Issue 2. April 2018.