International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Amalgamation of Machine Learning Algorithms for Crop Yield Prediction

D.Maghesh Kumar, K.Mohan Kumar

Abstract:- ***Agriculture is India's prime occupation. In Indian economy agriculture plays a major role by means of providing more employment opportunities for the people. In order to provide food to the huge population of India, agriculture sector needs to maximize its crop productivity. This research work presents an approach which uses different Machine learning (ML) techniques by considering the various parameters of cultivated crop to predict the best yield. Further in this Multiple Linear Regression (MLR) technique and artificial neural networks (ANN) are used to make a brief analysis for the prediction crop yield. With the above idea a new model was created, and from this numerical results were obtained. The accuracy and efficiency of the method has been explored and results from the artificial neural network and regression methods were obtained and compared.***

Key Words: Agriculture, Artificial Neural Networks, Crop Yield, Machine Learning, Multiple Linear regression.

I. INTRODUCTION

Expeditious increase in human population and diminished productivity of agricultural fields are major concern worldwide. In this scenario, food security has become a key problem and prediction of crop yield is an important issue in agricultural [1]. All farmers are always endeavouring to know, how much yield will achieve from his anticipation. Farmer's vast experience on crop cultivation was taken into account to analyze and for yield prediction[2].

Various factors such as soil, weather, pest, fertilizers, the geography of the region, and their interactions are used to determine crop yield but it is a highly complex process. The basic knowledge of agriculture and a clear understanding of linkages between the crop yield parameters help to predict the yield accurately. Revealing such linkages needs both exhaustive datasets and robust methods. It's very important to get the past crop yield details from the authorized source supports to make agriculture-related decisions correctly.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
D.Maghesh Kumar*, Research Scholar pursuing Ph.D., in the PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Trichirappalli, Tamil Nadu, India. Email: maghesh.d@gmail.com
Dr. K. Mohan Kumar, Heading the PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Trichirappalli, Tamil Nadu, India. Email: tnjmohankumar@gmail.com

The motivation behind this research work is to investigate appropriate machine learning techniques, which are handy to solve agricultural problems. Machine Learning comes into the picture when problems cannot be solved by means of conventional approaches. This study analyzes the various factors of agriculture, to evolve a new prediction model and helps the farmer's in easy decision making there by increasing the crop yield.

Machine learning uses archive data to gain knowledge and makes deeper insights in the data for output predictions. The excelling in the compilation of good dataset gives the better accuracy in predictions. It has been contemplated that machine learning techniques such as regression and classification gives better results than various other statistical models [3].

A. Multiple linear regression.

In statistics dependent variables are termed as predictant and independent variables are termed as predictors. The correlation between a predictant and a predictor is called linear regression but the correlation between a predictant and one or more predictors is called multiple linear regression (MLR).

The least square based methods are widely used in MLR techniques. Crop yield depends on all these ecological factors so, which is considered as dependent variable. In this crop yield prediction model, Production is the predictant and which is used in the MLR technique[4].

B. Artificial neural network

A human brain encompasses millions of neurons, which are used to communicate with each other, and send information. Neural networks replicate the central nervous system of humans [1]. Artificial neural networks (ANN), in fact, simulate a part of brain functions. One of the important concepts in machine learning is Artificial Neural Networks. Configure and train the ANN, is cumbersome and time consuming process because of its complex nature but once trained, it becomes very fast in executing application.

C. ANN structure

The artificial neurons have capability to process colossal amounts of data, and deduce to conclusions by deriving a pattern from it. Due to the speedy competence of neural networks in processing the training datasets, makes them to use in classification. Input, Hidden and Output are the three layers, which constitute the artificial neural network [5].

- **Input layer**: Get the raw data that has been fed to the network as input.

A desired output is also given along with input to verify the correctness of the neural network.

- **Hidden layers**: In this layer, weights and attributes multiplied then all such resultants are aggregated to generate the output.
- **Output layer**: When dissimilarity between the desired and the actual output are found, then the feedback is given to the hidden layer from output layer to obtain better accuracy.

The structure of ANN has been displayed in Fig 1.

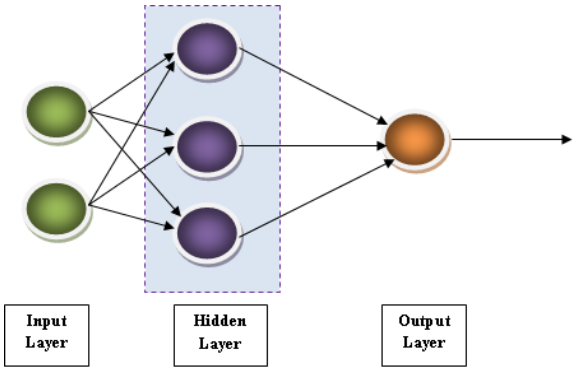

Fig. 1: ANN Structure

D. Neural networks training mechanism

The flow of data is propagated in forward direction and is always in the direction of output. There is no feedback mechanism, available in forward propagation. So, verifying the precision of output is extremely tedious.

Actual and expected output was repeatedly compared to train the back propagation. The difference between actual and expected output is given to the error function as input. The error functions are responsible for the changing of weights in the hidden layer. This process will be repeated again and again to reduce the difference to minimum.

Feed forward and back propagated artificial neural networks are used to get accurate crop yield forecast and helps the farmers in devising better decisions. ANN has been enforced for predicting better crop yield on the basis of various predictor variables [6].

The blue print of this paper is given below: Section II describes the literature survey in the field of crop yield prediction. Section III discusses about the proposed methodology for predicting the yield of crops. In Section IV, evaluates the proposed methodology and discuss the experimental results. In Section V, concludes the methodology for the yield of crops**.**

II. LITERATURE REVIEW

ANN is more powerful, when compared to conventional linear and simple nonlinear analyses. For yield estimation, ANN uses self adaptive method. In a special network structure a nonlinear response function has been used to study the cumbersome functional connection between input and output training data. Once the ANN model has been trained, it remembers functional connection and uses it in the later part of the calculations. ANN has the ability of self learning, so it has been widely used to create nonlinear and complicated models [6] [7].

During the preparation process agriculturalist wants easy and meticulous assessment methods to forecast the crop yields [8]. It is necessary to study the effectiveness of ANN models, before forecasting the crop yield in the usual weather conditions of mountain regions [9]. The effectiveness of MLR models are compared with ANN models by evaluating its performance in relation to various parameters [10].

Hemageetha discuss on the crop yield prediction by using soil parameters like organic matter, Soil pH, Electrical Conductivity and moisture. Naive Bayes algorithm achieves 77% of accuracy in soil classification. Apriori algorithm was used by combining the soil and crop details to attain the highest yield. Naïve Bayes, J48 and JRIP algorithm`s accuracy during classification were compared and listed [11].

Ankalaki, Chandra & Majumdar discuss about the clustering algorithms like DBSCAN and AGNES. Crop yields are forecasted by using MLR technique and for each crops a new formula has been devised [12].

Raorane & Kulkarni appraises several data mining approaches like ANN, Decision Tree, Regression Tree, Bayesian network, SVM, k-means to improve the agricultural crop productivity. Data mining techniques are mainly used to make classification of data[13].

Gonzalez-Sanchez, Frausto-Solis & Ojeda-Bustamante raises a substantial problem in agricultural planning is accurate yield estimation method. To achieve realistic solutions for this problem is Machine learning (ML) approach. A survey was conducted on ML methods used for yield prediction and to find the most meticulous method out of it. In this paper author noted the dissimilarity in the accuracy of crop yield prediction between ML with linear regression techniques [14].

III. METHODOLOGY

By analyzing the large datasets, Machine Learning algorithms classify and establish patterns from it. Based on available data, the crop yield production was analyzed. Crop yield Prediction by using Machine Learning technique will increase the crop productivity.

In this research the statistical methods were applied to get the predictive accuracy. For the estimation of crop yield analysis, MLR technique and ANN were taken up.

A. Overview of dataset

In any kind of research the data should collect properly, which makes the model more efficient; otherwise it can cause imprecise results. The dataset considered in the present study is collected from various sources like Department of Statistics and Agriculture, National Informatics Centre, data.gov.in and merged into one, such that a dataset sufficiently large enough for the study is created.In order to increase the purity of collected data is to be pre-processed by means of removing the outliers from it.

3.2 Variables used for crop yield prediction:

In this study the variables depicted for predicting the crop yield are described by us are Yearly Rainfall (YR), the cultivated Area (AR) is specified in Million Hectare, Food Price Index(FPI), Production (PRO) is mentioned in Million Tons and Crop Yield(CY) is specified in kilograms per hectare. Crop Production details in India are presented in the table I from 2001 – 17.

Table I: Crop production details of India

Year	YR	AR	FPI	PR	CY
2000-01	1120.2	44.71	92.4	84.98	1900.7
2001-02	981.4	44.9	101.0	93.34	2079
2002-03	1278	41.18	96.2	71.82	1744
2003-04	1085.9	42.59	98.1	88.53	2077
2004-05	1185.4	41.91	105.0	83.13	1984
2005-06	1133	43.66	106.8	91.79	2102
2006-07	1180.2	43.81	112.7	93.36	2131
2007-08	1075	43.91	134.6	96.69	2202
2008-09	972.8	45.54	155.7	99.18	2178
2009-10	1212.3	41.85	132.8	89.13	2129.7
2010-11	1213	36.95	150.7	80.41	2177
2011-12	1090.2	41.59	151.1	90.56	1977
2012-13	1270	40.18	154.2	76.86	1846
2013-14	1182.4	41.82	150.6	83.36	2131
2014-15	1078.2	43.76	152.4	92.02	2106
2015-16	990.8	42.52	154.7	99.18	2178
2016-17	1212.3	41.85	156.8	75.11	1806.7

B. Multiple Linear Regression

If the variables used in regression function is linear then that model is called as linear regression model otherwise it is non-linear model. The relationship between one dependent variable and more than one independent variables is called multiple linear regression. Let us take dependent variable as y, independent variables as x_i and set of unknown parameters as b_i which is associated by the regression function.

Multiple linear model General formula is as in (1)

$$y = b_0 + b_1x_1 + b_2x_2 + \dots b_px_p \quad \text{------- (1)}$$

In this research work, the independent variables YR, AR, FPI, PRO and dependent variable CY are used for the MLR analysis to determine the relationship among them. Here the crop yield is influenced by environmental factors such as rainfall etc.,

The dissimilarity between some data points and the regression line is called residuals otherwise known as "errors". The residual (e) is calculated by subtracting dependent variable`s observed value (y) from the predicted value (y`) is shown in the equation (2). In general each data point used has one residual.

$$e = y - y` \quad \text{------- (2)}$$

The deviation of the predicted values from the observed values are squared and then summed is called residual sum of squares. The squared value of correlation coefficient is taken as R^2 and is used to evaluate the strength of any linear model. The R^2 value corresponds the dependent variable`s variability, which should always lie within the range 0 and 1. The data in the Table 1 is showing the association between the independent variables YR, AR, FPI, PRO and dependent variable CY after applying the MLR process.

C. Artificial Neural Network

Inputs to the ANN are represented by x0, x1, x2, x3...x(n) and each of these inputs are multiplied by a connection weight or synapse which are represented as w0, w1, w2, w3….w(n) . The Weight represents the particular nodes strength. A bias value b allows moving the activation function in either direction. To generate a result, these products are summed and fed to an activation function. Finally this result is sent as output. This mathematically represented in (3).

$$x_1.w_1 + x_2.w_2 + x_3.w_3 \dots\dots x_n.w_n = \sum x_i.w_i \quad \text{---(3)}$$

Later apply activation function $(\sum x_i.w_i)$

IV. RESULTS AND DISCUSSION

The various regression statistics obtained from the tool XLMiner are listed in table II. Obtained value of R^2 through the regression analysis is larger than 0.5, means the bonding between the dependent variable and the independent variable is more.

Table II: MLR Statistics

MLR Statistics	
Multiple R	0.881948
R^2	0.9012232
Adjusted R^2	0.697043
Standard Error	78.21197
Observations	16

The crop production was influenced by increase or decrease in the value of independent variables such as YR, AR, FPI and PRO. By the implementation Multiple Linear Regression Analysis the influenced value of R^2 = 0.91 is obtained is shown in Table I. Obtained R^2 value clearly states that an average of 91% influence by the environmental factors in crop yield. Thus the crop yields are vulnerable on three key factors like Yearly Rainfall, Cultivation Area, and Food Price Index. The line fit to plot shows the predicted values against the residuals are shown in fig 2.

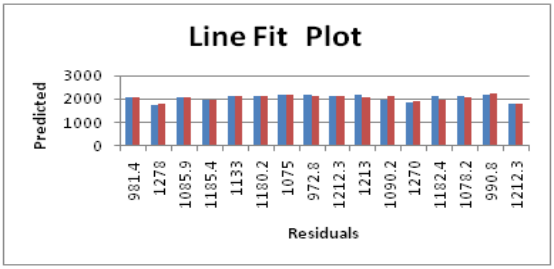

Fig 2: Line Fit to Plot

Here, Experiments based on ANN were conducted by with the Fitting Parameters are listed as in table III.

Table III: Fitting Parameters

Random seed for initial weights	12345
Maximum # Hidden Layers	2
Maximum # Neurons in Hidden Layer 1	3
Maximum # Neurons in Hidden Layer 2	2
Learning rate	0.25
Weight change momentum	0.6

A. Predictions using MLR and ANN

The statistic value R^2 for the MLR and ANN was measured to find its accuracy in prediction. The closeness the data to the fitted regression line was calculated by using the statistical measure R^2 is given in equation (4),

$$R^2 = 1-((n-1/n-p)*(SSE/SST)) \quad \text{------- (4)}$$

Where, sum of squared error (SSE), sum of squared total (SST), n, p are the number of observations and regression coefficients respectively. The prediction accuracy of the given model is obtained by calculating the value of R^2 and the value of R^2 is higher, then the prediction by the given model is better. The accuracies or the MLR and ANN models are shown in Table IV.

Table IV: Accuracies based on R^2 value

Prediction	Accuracy in (%) during		
models	Training	Validation	Testing
MLR	100	100	90.12
ANN	95	89	94

During training and validation phase of MLR model, the obtained R^2 value was 100%. Due to the non-linear relationship among the parameters, the R^2 value obtained during test phase of the MLR model was reduced to 90.12% but when using ANN model with the learning rate of 0.25 had a higher accuracy of 94%. This pinpoints that the ANN model forecasts better during test phase when compared to the MLR with an improvement of nearly 4% over MLR models.

When making deeper insights in fig. 4, MLR performs well during training and validation performs better but the ANN performs well in testing.

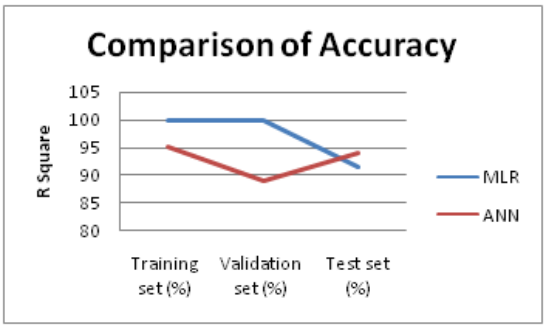

Fig 4: MLR Vs ANN

V. CONCLUSION

In the present technological era, it is mandatory for any agriculturalist to update them from the traditional approach into the modern scientific approach for doing the agriculture. In this study different parameters were considered for crop yield prediction because in agriculture crop yield prediction is substantial. After comparing the results of MLR with ANN model during test phase shows that ANN has maximum accuracy that is higher R^2 value with minimum prediction error than MLR model. Thus the ANN models predicts crop yield better than MLR model. Present research work is continued by combining more factors like crop diseases, different irrigation patterns, climatic conditions etc. that influences the yield of a crop and these influencing factors were analyzed by using different machine learning techniques. Successful integration of machine learning with agriculture helps the farmers to maximize their yield and optimize the use of available resources.

REFERENCE

1. Subhadra Mishra, Debahuti Mishra and Gour Hari Santra, "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper", *Indian Journal of Science and Technology (INDJST)*, 2016, Vol. 9(38), pp. 1-14.
2. https://en.wikipedia.org/wiki/Data_analysis
3. Veenadhari, S., Misra, B., & Singh, C, "Machine learning approach for forecasting crop yield based on climatic parameters", *International Conference on Computer Communication and Informatics (ICCCI - 2014)*, 2014 IEEE.
4. D.S.Zingade1 et.al, "Crop Prediction System using Machine Learning", *International Journal of Advance Engineering and Research Development*, 2017, Vol. 4, Special Issue 5.
5. Chusnul Arif et. al, "Estimation of soil moisture in paddy field using Artificial Neural Networks", *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, 2012, Vol.1, No. 1, pp. 17-21.
6. Simpson G, "Crop yield prediction using a CMAC neural network", *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, 1994, pp. 160–171.
7. Louis, E. K., and Yan, X.-H., "A neural network model for estimating sea surface chlorophyll and sediments from Thematic Mapper imagery", *Remote Sensing of Environment*, 1998 – Elsevier, Vol. 66, Issue 2, pp. 153–165.
8. Ji, B. Sun, Y. Yang. S. & Wan, J. "Artificial neural networks for rice yield prediction in mountainous regions" *Journal of Agricultural science*, 2007, Vol. 145, Issue 3, pp. 249-261.
9. Snehal S.Dahikar and Sandeep V.Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach". *International journal of innovative research in electrical, electronics, instrumentation and control engineering*, 2014, Vol. 2, Issue 1 pp. 683-686.
10. Shivnath Ghosh. & Santanu Koley. "Machine Learning for Soil Fertility and Plant Nutrient Management", *International Journal on Recent and Innovation Trends in Computing and Communication*, 2014, Vol. 2 Issue 2, pp. 292-297.
11. Hemageetha, N., "A survey on application of data mining techniques to analyze the soil for agricultural purpose," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 3112-3117.
12. Ankalaki, S., Chandra, N., Majumdar, J., "Applying Data Mining Approach and Regression Model to Forecast Annual Yield of Major Crops in Different District of Karnataka", *International Journal of Advanced Research in Computer and Communication Engineering*, 2016, Vol. 5, Special Issue 2, pp.25- 29.

13. Raorane, A.A., Kulkarni R.V., "Data Mining: An effective tool for yield estimation in the agricultural sector", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2012. Vol. 1, Issue 2, pp.75-79.
14. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, Ojeda-Bustamante Waldo, "Predictive ability of machine learning methods for massive crop yield prediction", *Spanish Journal of Agricultural Research*, 2014, Vol. 12, No 2, 313–328.

AUTHORS PROFILE

D.Maghesh Kumar, received the M.Sc., degree in Computer Science from the Bharathidasan University, Tiruchirappalli and M.E., in Computer Science from Anna University, Tiruchirapalli. Currently pursuing Ph.D., as a part time research scholar in the PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Tamil Nadu, India. He qualified in TNSET and he is having 25 years of rich teaching experience. His focuses his research work mainly on Data analytics and Machine Learning. He published more than 20 research articles in the peer reviewed International journals.

Dr. K. Mohan Kumar, pursed M.Sc., and Ph.D in Computer Science from Bharathidasan University, Tiruchirappalli, He received M.Phil in computer science from Manonmaniam Sundaranar University, Thirunelveli, India and currently heading the PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India. His research work focuses mainly on Network Security, Data analytics, Machine Learning and IoT. He published more than 50 research papers in reputed International journals. He has more than 25 years of teaching experience blended with 20 years of Research Experience.