A Multiple Linear Regressions Model for Crop Prediction with Adam Optimizer and Neural Network Mlraonn

M. Lavanya[1]
Department of Computer Science
School of Computing Sciences
VISTAS, Chennai
India

Dr. R. Parameswari[2]
Department of Computer Science,
School of Computing Sciences
VISTAS, Chennai
India

***Abstract*—Due to the increase in population, demand for the food is increasing day by day. Crop prediction is necessary or need of the hour to fill the gap between the demand and the supply. Instead of following a traditional system for crop selection method, a successful crop selection for the given soil properties will help the farmers to get the expected crop yield. The objective of the proposed work is to develop one such system. The proposed system is developed using real data with various soil parameters acquired from soil laboratory located in Chennai. This system uses 16 parameters of soil which includes all the micro, macro nutrients along with that pH, EC, OM values and the recommended crop for the soil parameter. The proposed Mlraonn (Multiple Linear Regression with Adam Optimization in Neural Network) model is developed using Keras software mainly used for Deep Learning. A neural network approach is used to construct a regression model. The model is evaluated with Loss Metrics such as RMSE, MSE, and MAE. The proposed algorithm is compared with the existing standardized machine learning algorithms. It is found that the proposed algorithm gave very minimal error as output in all the above three categories of loss metrics than the standardized algorithm such as Random Forest Regression and Multiple Linear Regression.**

Keywords*—Multiple Linear Regression; Adam Optimization; Neural Network; Keras; Machine learning algorithm; Root Mean Square Error (RMSE); Mean Square Error (MSE); Mean Absolute Error (MAE); presence of Hydrogen (pH); Electrical Conductivity (EC); Organic Matter (OM)

I. Introduction

First and foremost method in statistics is linear regression; the mathematical equation representation for the same is Y =m x + c; where y is the predicted output; x is the input variable; m is the slope and c is the bias. The above idea can be extended to multiple linear regression where more than one input features which produces single output feature. The mathematical representation of multiple linear regression is; Y = m1*x1+ m2*x2+m3*x3+.........+mn*xn + c. A neural network model can be created by calculating Weights and bias value at each and every node [23]. The layer consists of various nodes; layers are classified in to input; hidden and output layers. Inputs are multiplied with weights of the node to form a summation of the activation function. The activation is a transformation function that may be a linear or non-linear; applied to every input before it gets transferred to the next layer or to the output layer. Different types of activation function available some of those are Sigmoid; RELU; Leaky RELU and Tanh; all activation function has its own purpose [23]. Linear activation function is very simple than non-linear. RELU and Sigmoid is an example for linear and non-linear activation function respectively. Rectified Linear activation (RELU) requires no transformation and model can be easily trained mainly used for multiple linear regression. The performance of the neural network can be optimized with the optimization function one such is gradient descent. In order to adjust the weights; gradient descent algorithm is used; from which the relation between the error and a single weight can be obtained. This optimization step used to arrive at a conclusion that at which point of weight a very low error is generated. Minimizing the error value is the overall aim of developing any model. In the Feed forward step the weights for all the nodes are calculated with the activation function. Whereas in the back propagation step weights of the network is adjusted based generated error. The model can be trained quickly and its performance will be increased with optimization algorithm. There exists many optimization algorithm; some examples are Sgd; Rmsprop; Nestrov; Adagrad; Adadelta; Adam [26] and so on. Adam optimizer is used to update the node weights. This algorithm is a variation of gradient descent algorithm. It uses two momentum first order momentum is a mean value and second order momentum is variance value. Section II of this paper tells about the related works using various machine learning algorithms. Section III explains about data collection and pre-processing works carried with the dataset. Section IV gives the pseudo code for the proposed algorithm. Section V gives the comparison of the results. Section VI gives the conclusion part.

II. Related Works

As a part of crop management for wheat crop its biomass was estimated using machine learning algorithm such as Random Forest Regression; SVC Regression and ANN. It is that Random Forest produced accurate estimation than other two algorithms. Experiment took place in southern China [12]. Data collected from weather department to predict most profitable crop; analyzed parameters such as pH; soil & weather data. Author used multiple linear regressions; a machine learning algorithm for prediction. Detecting crop diseases to help farmers has been discussed by the author can be considered as future enhancement [7]. Sensing soil parameters and atmospheric parameters; author used ANN algorithm to predict the crop yield [21]. Rice yield prediction in Maharashtra state data collected from the year of 1998 to 2002. Neural network algorithm is used for prediction; cross validation technique is used to validate the result; accuracy of 97.5% sensitivity and 96.3% spacificity [17]. Crop yield with respect to climate and biophysical change; a huge data were collected; algorithm such as Random Forests and MLR. It is concluded that Random Forest performance seems to be better than MLR [8]. Random forest is used to calculate the accuracy with the climate data [3]. Yield prediction of crops like wheat; maize and potato is done with huge data divided in to training and testing. Algorithm such as Random Forests is compared with Multiple Linear Regression –MLR. Found that RF performed better than MLR. The RMSE value of RF was 6 and14% whereas MLR ranges from 14 to 49% [18]. Four statistical prediction models such as MLR; SGD; RFR and SVM to find soil information in south western Burkina. High spatial resolution satellite data along with soil sample data from laboratory was used. It is found that RFR performed better than other four algorithms. Internal validation is done through cross validation [9]. Crop yield prediction was carried with the data obtained from soil testing lab at Jabalpur; Madhya Pradesh. Naive Bayes and KNN models were generated. Soil were classified in to three categories low; medium and high based on their nutrients values found in the soil. The outcome of this study helped the farmers for choosing a better sowing land. Since the test carried with small dataset the author wishes to have big dataset as future work to get better accuracy [19]. An optimization algorithm such as Gradient Descent with Momentum was used to train neural network pattern classification algorithm to find the soil moisture in an hour advance for irrigation which helped farmers the follow an irrigation pattern. The MSE and RMSE obtained by Gradient Descent with Momentum based neural network pattern classification are 0.039622 and 0.19905 [20]. Three crops such as rice; maize and wheat were considered for study. Machine learning algorithms like Multiple Linear Regression; Random Forest Regression and Multivariate Adaptive Regress-ion Splines (Earth) were used for predict the yield of chosen crops. Multiple Linear Regressions gave good prediction [6]. State wise prediction of rainfall was carried by MLR algorithm [11]. Random Forest Regression was used by the author to predict sugarcane yield [13]. To predict agricultural yield using various algorithms such as linear; non-linear and MLR; experiment was carried at Andhra Pradesh; Telangana state [14]. Crop yield prediction were carried by ANN and MLR; C-ANN and D-ANN algorithm were compared for their performance [15]. Climate change on mustard yield prediction was carried in Haryana state using MLR [16]. Agriculture data is analyzed to find the optimal parameters for maximizing crop production using algorithm Multiple Linear Regressions; PAM; CLARA; DBSCAN [10].

III. Data Collection and Pre-Processing

A. Data Collection

Crop prediction with this proposed system developed with only by using soil properties such as micro; macro nutrients Ec; Om & pH values as input or independent features and suggested crop as output or dependent features. The above mentioned soil properties was collected from a soil lab. Dataset consists of nearly 1600 samples. The dataset is analyzed before generating a model. The sample dataset is as shown in "Fig. 1".

B. Data Pre-Processing

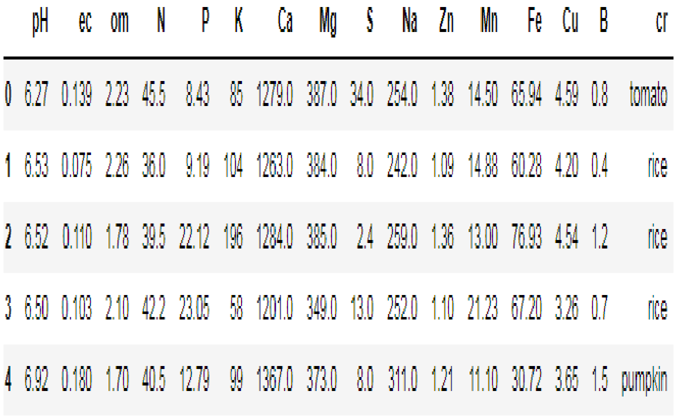
1) Finding correlation among features: The study of data reveals the nature of data as numerical data. In order to find a relationship between features a correlation map called heat map is generated which is shown in "Fig. 2" shows the correlation value and the generated heatmap [25] for the dataset. Heat map is used to find the correlation between each and every feature in the dataset. The correlation values ranges from -1 to +1; the correlation value of a feature which is near to -.01 to +.01 can be dropped since it denotes the value is equal to zero which mean there is no correlation. In this dataset the features such as N; P; Na; Zn and B were removed for crop prediction; which is shown in the "Fig. 3".

2) Dataset scaling: The dataset is examined to find out the range of the feature it is found that the values differ their exits no uniformity; with this; it is not possible to generate a correct model. A solution is to scale all the values in a predefined range which is nothing but -1 to +1. The above step called scaling and it is implemented with the help of Standard Scalar a pre processing function in sklearn. The code is as follows;

```
from sklearn.preprocessing import StandardScaler

scaledX1 = StandardScaler().fit(X)

Xsca = scaledX1.transform(X)
```



	pH	ec	om	N	P	K	Ca	Mg	S	Na	Zn	Mn	Fe	Cu	B	cr
0	6.27	0.139	2.23	45.5	8.43	85	1279.0	387.0	34.0	254.0	1.38	14.50	65.94	4.59	0.8	tomato
1	6.53	0.075	2.26	36.0	9.19	104	1263.0	384.0	8.0	242.0	1.09	14.88	60.28	4.20	0.4	rice
2	6.52	0.110	1.78	39.5	22.12	196	1284.0	385.0	2.4	259.0	1.36	13.00	76.93	4.54	1.2	rice
3	6.50	0.103	2.10	42.2	23.05	58	1201.0	349.0	13.0	252.0	1.10	21.23	67.20	3.26	0.7	rice
4	6.92	0.180	1.70	40.5	12.79	99	1367.0	373.0	8.0	311.0	1.21	11.10	30.72	3.65	1.5	pumpkin

Fig. 1. Shows the Sample Rows in the Crop Dataset.

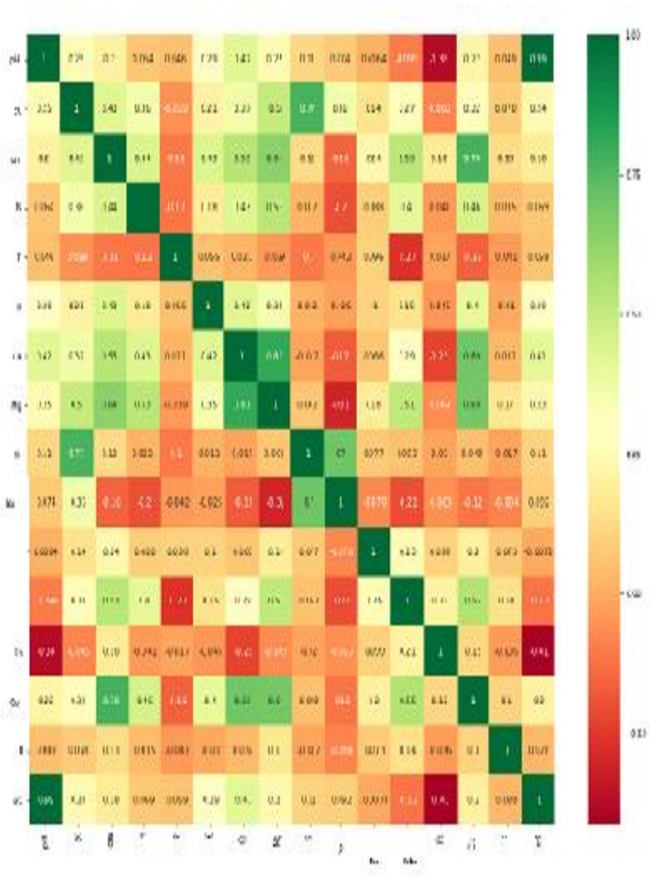
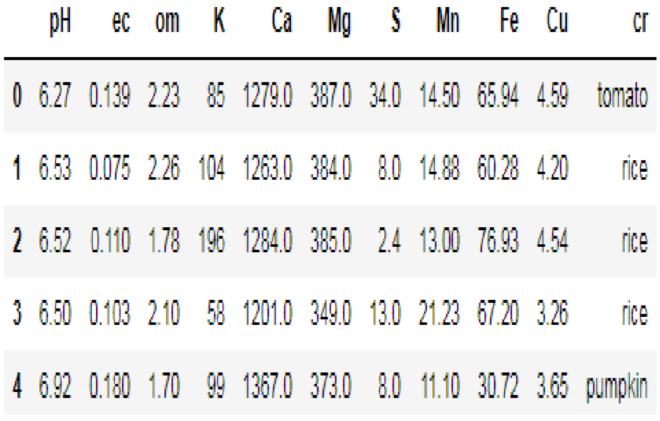


Fig. 2. Showing Correlation among All the Features the Values Ranges from – 0.0 to + 1.0 [25].



	pH	ec	om	K	Ca	Mg	S	Mn	Fe	Cu	cr
0	6.27	0.139	2.23	85	1279.0	387.0	34.0	14.50	65.94	4.59	tomato
1	6.53	0.075	2.26	104	1263.0	384.0	8.0	14.88	60.28	4.20	rice
2	6.52	0.110	1.78	196	1284.0	385.0	2.4	13.00	76.93	4.54	rice
3	6.50	0.103	2.10	58	1201.0	349.0	13.0	21.23	67.20	3.26	rice
4	6.92	0.180	1.70	99	1367.0	373.0	8.0	11.10	30.72	3.65	pumpkin

Fig. 3. Shows the Dataset after the Removal of the Less Correlated Field.

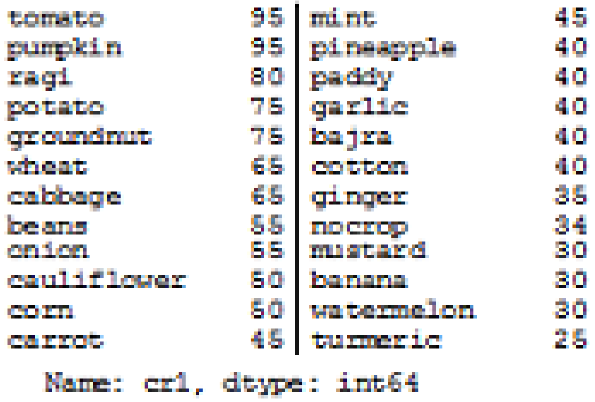
3) Label encoding: The target feature such as suggested crops seems to be of string data type; it is wise if it get converted in to numeric. The dataset has 27 different crop names; which comes under multiclass. Label encoding is a method which automatically assigns numerical value when it is called for a particular feature and the data type is converted in to integer array. The code is shown below;

from sklearn import preprocessing

le1=preprocessing.LabelEncoder()

YSca=le1.fit_transform(y)

4) Handling imbalanced dataset: The value counts of the target variable in the dataset are found to be imbalanced. For example: **y.value_counts()** .



```
tomato          95 | mint          45
pumpkin         95 | pineapple     40
ragi            80 | paddy         40
potato          75 | garlic        40
groundnut       75 | bajra         40
wheat           65 | cotton        40
cabbage         65 | ginger        35
beans           55 | nocrop        34
onion           55 | mustard       30
cauliflower     50 | banana        30
corn            50 | watermelon    30
carrot          45 | turmeric      25
     Name: cr1, dtype: int64
```

The figure shows the feature value in the decreasing order. Resultant prediction will be incorrect if the above dataset is used for the same. It is necessary to follow a technique which balances the dataset in order to get the correct prediction. SMOTE- Synthetic Minority Over Sampling Technique [24]; in order to increase the sample size; synthetic data need to be generated; Smote uses KNN for oversampling the data [24]. Below line shows the training dataset shape of the independent features x and the dependent feature y. After the implementation of the code the size of the x and y features have changed; showing that the dataset is a balanced one.

x_train.shape;y_train.shape

((1199; 9); (1199;))

from imblearn.over_sampling import SMOTE

smo1=SMOTE('minority')

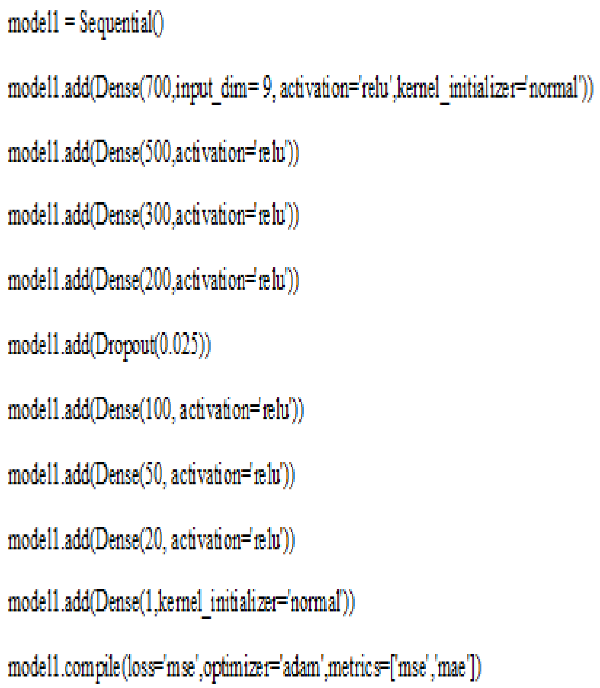
x_sm1;y_sm1=smo1.fit_sample(x_train;y_train)

print(x_sm1.shape;y_sm1.shape)

(1254; 9) (1254;)

IV. Proposed Algorithm Explaination

A neural network model with Multiple Linear Regression [2] [5] is generated for training the dataset. The generated model does the functionality of regression using Keras Regressor [22]. Keras is loaded with more built-in libraries through which neural network model can be built efficiently and easily. In order to develop a fully connected network keras layer is imported with Dense [22]. To avoid over fitting this model uses Dropout. Sequentially the layer can be added till the expected result reaches. The input dimension is assigned with the required value. The activation function relu is used here; this is a linear activation function; it denotes that the weights will be taken as it is only for positive output otherwise negative values will be assigned to zero. In output layer is declared with the value 1; which is nothing but the output dimension. Since this is regression problem the model can be evaluated with loss metrics. Since this is a regression model loss is included in model compile. Two metrics MSE and MAE were calculated for this model. Below code shows the neural network model for crop prediction using keras.



```
model1 = Sequential()

model1.add(Dense(700,input_dim=9, activation='relu',kernel_initializer='normal'))

model1.add(Dense(500,activation='relu'))

model1.add(Dense(300,activation='relu'))

model1.add(Dense(200,activation='relu'))

model1.add(Dropout(0.025))

model1.add(Dense(100, activation='relu'))

model1.add(Dense(50, activation='relu'))

model1.add(Dense(20, activation='relu'))

model1.add(Dense(1,kernel_initializer='normal'))

model1.compile(loss='mse',optimizer='adam',metrics=['mse','mae'])
```

The dataset is split in to training and the validation set. The 1004 samples is considered as training and 495 samples is taken as validation set. The metrics of training loss and the validation loss are calculated in parallel by the model for 200 epochs. Below Table I shows all the loss values for training and the validation set for every 50 epoch.

A part of the training is kept separately as a validation set for checking purpose. Validation set consists of new set of data which is not trained so for. A graph is plotted [27] between the training and validation loss in order to check how the model behaves with an unseen data which is one shown in “Fig. 4”. From the figure model behaviour is good since there is a minimum deviation between the two loss lines; as it is understood from the theory if the deviation is high the behaviour of the model is not good towards validation set ; further tuning is required in order to improve the same.

TABLE I. SHOWS THE LOSS VALUES FOR TRAINING AND THE VALIDATION

Metrics	**50th iterations**	**100th iterations**	**150th iterations**	**200th iterations**
Training loss	14.3294	5.7992	0.2365	0.1673
Training_mse	14.3294	5.7992	0.2365	0.1673
Training_mae	2.8082	1.5482	0.3556	0.2944
Validation loss	13.9797	3.6981	0.1327	**0.1185**
Validation_mse	13.9797	3.6981	0.1327	**0.1185**
Validation_mae	2.7521	1.3439	0.2577	**0.2476**

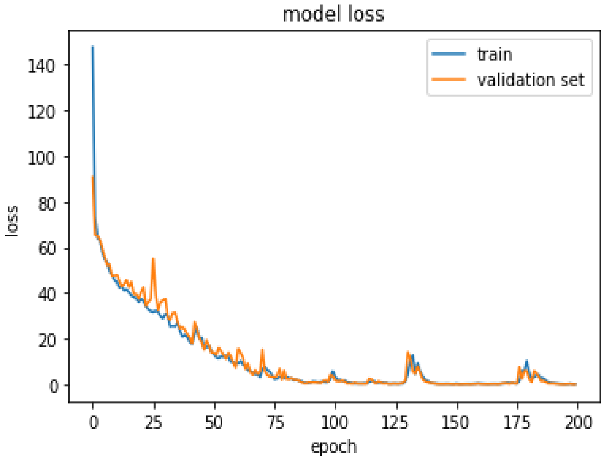


Fig. 4. Shows Model Loss in Validation and Training Data [27].

V. RESULTS AND DISCUSSION

Random Forest Regression algorithm is a machine learning algorithm which can be used for both regression and classification problem. It is an ensemble technique [1] with multiple decision trees; it also uses a Boosting technique called Bagging. The result obtained here by combining multiple trained decision trees; which seems to more effective; than taking decision with single decision tree. Since the structure of the dataset has multiple independent variables and a dependent variable. It is better to choose Multiple Linear Regression algorithm [4] for the above circumstance. Loss Metrics such as Root Mean Square generally calculate the average squared difference between the actual table value and the predicted value. Mean Absolute Error is the average; absolute value of the all the residual data points. Mean Square Error is similar to MAE; it take square of all the residual points and sum those values.

The study considered proposed Mlraonn model and the other two standard algorithms such as Random Forest Regression and Multiple Regression Algorithm. The Loss metrics such as Mean Squared Error; Mean Absolute Error and Root Mean Squared Error of all the standardized algorithms; along with the proposed Mlraonn model were compared. It is found that the proposed Mlraonn model performs better than the standardized algorithms which is shown in the below Table II.

TABLE II. SHOWS THE COMPARISON BETWEEN GENERALISED AND THE PROPOSED ALGORITHM

SI. No.	**Algorithms**	**Mean Squared Error (MSE)**	**Mean Absolute Error (MAE)**	**Root Mean Squared Error (RMSE)**
1.	Random Forest Regression	0.470	0.27	0.685
2.	Multiple Linear Regression	55.06	6.29	7.420
3.	**Mlraonn model**	**0.1185**	**0.2476**	**0.3442**

VI. CONCLUSION

It is found from the table which compares the obtained result of the propose model with standardized algorithm; shows that the result of the proposed Mlraonn model performed better with less epochs or iterations than the two standard machine learning algorithm. From the graph it is very clear that the model is performing uniformly with both training data and unseen validation data. This can be understood from loss value mentioned for every 50th epochs up to 200th epochs. The graph also explains the same. The evaluated loss metrics such as RMSE; MSE and MAE shows very less value for the proposed algorithm. From this it is concluded that Regression model built with neural network suits well for this soil dataset.

VII. LIMITATIONS AND FUTURE ENHANCEMENT

Every developed system has its own limitations; the proposed system also has some limitations which can also be considered as future enhancement; which are as follows; here the crop is predicted only by using the soil parameter. Other parameters like weather condition; wind speed, etc. is not included for the study. As a future enhancement more parameters other than the soil parameters can be included.

From the heatmap; it is found that the correlation of pH is high which is equal to 0.99. A future study for prediction of crop only using pH value can also be tried; using pH sensor. Various optimization algorithms can also be compared with the same dataset to justify the effectiveness and strength of Adam optimizer algorithm.

REFERENCES

[1] Han Chena; Jinhui Jeanne Huanga; Edward McBean; Partitioning of daily evapotranspiration using a modified shuttleworth wallace model; rand-om Forest and support vector regression; for cabbage farmland; Elsevier; 2019.

[2] Khaoula Abrouguia; Karim Gabsib; Benoit Mercatorisc;Chiheb;khemisa; Roua Amamia; Sayed; Chehaibia Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR); Elsevier; 2019.

[3] Abhijeet Pandhe; Praful Nikam; Vijay Pagare; Pavan Palle; Dilip Dalgade; Crop Yield Prediction based on Climatic Parameters; e-ISSN: 2395- 0056; p-ISSN: 2395-0072; IRJET; Volume:06 Issue:03 | Mar 2019.

[4] Noel Dougba Dago; Nafan Diarrassoub; Martial Didier Yao Saraka;jean-Luc Aboya Moroh; Inza Jesus Fofana; Lamine Baba-Moussa and Adam-a Coulibaly; Predicting maize and soybean crops dry biomass through rhizobacteria microorganisms activity on foliar bio-fertilizer in an aridagro-climate: A multiple linear regression analysis Vol. 12(34); pp. 835-848; 14 September; 2018.

[5] Moslem Abdipoura; Mehdi Younessi-Hmazekhanlub; Seyyed Hamid;Re Za;Ramazanic; Amir hassan midid;Artificial neural networks and multipl linear regression as potential methods for modeling seed yield of safflower (Carthamus tinctorius L.); 0926-6690; Elsevier;2018.

[6] Suvidha Jambekar; Shikha Nema; Zia Saquib; Prediction of Crop Production in India Using Data Mining Techniques; IEEE;2018.

[7] D.S. Zingade; Omkar Buchade; Nilesh Mehta; ShubhamGhdekar;Chanda Mehta; Machine Learning based Crop Prediction System Using Multi-Linear Regression; Volume: 3; Issue: 2; IJETCS;ISSN:2455-9954; April 2018.

[8] R.Karthikeyan; M.Gowthami; A.Abhishhek; P.Karthikeyan; Implementation of Effective Crop Selection by Using the Random Forest Algorithm; International Journal of Engineering & Technology;7 (3.34) ; 287- 290; 2018.

[9] Gerald Forkuor; Ozias K. L. Hounkpatin; Gerhard Welp; Michael Thiel; High Resolution Mapping Of Soil Properties Using Remote Sensing Va riables In South-Western Burkina Faso: A Comparison Of Machine Learning And Multiple Linear Regression Models; PLOSONE | journal . pone; January 23; 2017.

[10] Jharna Majumdar; Sneha Naraseeyappa and Shilpa Ankalaki; Analysis of agriculture data using data mining techniques: application of big data Majumdar et al. J Big Data ; 4:20;2017.

[11] Jesleena Rodrigues ;Arti Deshpande; Prediction of Rainfall for all the states of India using Auto-Regressive Integrated Moving Average; Model and Multiple Linear Regression; IEEE;2017.

[12] Li'ai Wang; Xudong Zhou; Xinkai Zhu; Zhaodi Dong; Wenshan Guo Estimation of biomass in wheat using random forest regression algorithm and remote sensing data; 2214-5141; CAAS; 2016.

[13] Yvette Everingham; Justin Sexton; Danielle Skocaj & Geoff Inman-Bamber; Accurate prediction of sugarcane yield using a random forest algorithm; Agron. Sustain. Dev.; 36: 27; 2016.

[14] S.Nagini; T. V. Rajini Kanth; B.V.Kiranmayee; Agriculture Yield Prediction Using Predictive Analytic Techniques; IEEE; 2016.

[15] K. Aditya Shastry; H.A.Sanjay ; Abhijeeth Deshmukh; A Parameter Based Customized Artificial Neural Network Model for Crop Yield Prediction. Journal of Artificial Intelligence; 9: 23-32; 2016.

[16] U. Verma; H. P. Piepho; A. Goyal; J.O. Ogutu1 and M.H. Kalubarme; Role Of Climatic Variables And Crop Condition Term For Mustard Yield Prediction In Haryana ; ISSN : 0973-1903; Int. J. Agricult. Stat Sci. Vol.12; No. 1; pp. 45-51; 2016.

[17] Niketa Gandhi; Owaiz petkar ; Leisa J. Amstrong ; Rice Crop Yield Prediction Using Artificial Neural Networks; International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR 2016); IEEE; 2016.

[18] Jig Han Jeong Jonathan ;P. Resop; Nathaniel D. Mueller; David H.Fleisher;Kyungdahm Yu; Ethan E. Butler; Dennis J. Timlin; Kyo-Moon Shim; James S. Gerber; Vangimalla R. Reddy; Soo-Hyung Kim ;Niketa Gandhi Owaiz Petkar; Random Forests for Global and Regional Crop Yield Predictions; journal.pone; 2016.

[19] Monali Paul; Santosh K. Vishwakarma; Ashok Verma ;Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach; IEEE; 2015.

[20] Saroj Kumar Lenka; Ambarish; G. Mohapatra ; Gradient descent with momentum based neural network pattern classification for the prediction of soil moisture content in precision agriculture; IEEE;2015.

[21] Snehal S.Dahikar; Sandeep V.Rode; Agricultural crop yield prediction using Artificial Neural network approach; Vol. 2; Issue 1; IJIREEICE; January 2014.

[22] Jason Brownlee; https://machinelearningmastery.com/tutorial-first Neural-network-python-keras; july 24; 2019.

[23] Chris Nicholson; https://pathmind.com/wiki/neural-network;2019

[24] Javaid Nabi; http://towardsdatascience.com/machine-learning Multiclass-classification with Imbalanced-data-set-29f6a177c1a; Dec 23; 2018.

[25] Milind paradar; https://blog.quantinsti.com/creating-heatmap-using-Python-seaborn; Dec 19; 2016.

[26] An Overview of gradient Descent Optimization Algorithms; Sebastian Ruder; 19 Jan 2016.

[27] Jason Brownlee; https://machinelearningmastery.com/tensorflow-tutorial-deep-learning-with-tf-keras/; December 19; 2019.