Crop Yield Prediction Using Machine Learning Algorithms

Aruvansh Nigam
Dept. Of Computer Science and Information Technology
Jaypee Institute of Information Technolology, India
aruvanshn@gmail.com

Saksham Garg
Dept. Of Computer Science and Information Technology
Jaypee Institute of Information Technolology, India
sakshamgarg500@gmail.com

Archit Agrawal
Dept. Of Computer Science and Information Technology
Jaypee Institute of Information Technolology, India
archit.agrawal321@gmail.com

*Parul Agrawal
Dept. Of Computer Science and Information Technology
Jaypee Institute of Information Technolology, India
*parul.agrawal@jiit.ac.in *Corresponding Author*

***Abstract*—Agriculture is one of the major and the least paid occupation in India. Machine learning can bring a boom in the agriculture field by changing the income scenario through growing the optimum crop. This paper focuses on predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.**

Keywords: crop yield prediction;long short-term memory(LSTM);simpleRNN; random forest;xgboost;machine learning classifiers;ensemble learning

I. INTRODUCTION

The history of agriculture in India[1] dates back to the Indus Valley Civilization Era. India ranks second in this sector. Agriculture and allied sectors like forestry and fisheries account for 15.4 percent of the GDP (gross domestic product) with about 31 percent of the workforce. India ranks first globally with the highest net cropped area followed by US and China. Agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic fabric of India. Due to the revolution in industrialization, the economic contribution of agriculture to India's GDP is steadily declining with the country's broad-based economic growth. The problem that the Indian Agriculture sector is facing is the integration of technology to bring the desired outputs. With the advent of new technologies and overuse of non-renewable energy resources patterns of rainfall and temperature are disturbed. The inconsistent trends developed from the side effects of global warming make it cumbersome for the farmers to clearly predict the temperature and rainfall patterns thus affecting their crop yield productivity. In order to perform accurate prediction and handle inconsistent trends in temperature and rainfall various machine learning algorithms like RNN, LSTM, etc can be applied to get a pattern. It will complement the agricultural growth in India and all together augment the ease of living for farmers. In past, many researchers have applied machine learning techniques to enhance agricultural growth of the country.

Balamurugan et al. [2] have implemented crop yield prediction using only random forest classifier. Various features like rainfall, temperature and season were taken into account to predict the crop yield. Other machine learning algorithms were not applied to the datasets. With the absence of other algorithms, comparison and quantification were missing thus unable to provide the apt algorithm. Mishra et al. [3], has theoretically described various machine learning techniques that can be applied in various forecasting areas. However, their work fails to implement any algorithms and thus cannot provide a clear insight into the practicality of the proposed work. Manjulas et al.[4] research aimed to propose and implement a rule-based system to predict the crop yield production from the collection of past data by applying association rule mining on agriculture data from 2000 to 2012. The dataset used in this research is limited to the southern district of India thus limiting its scope for pan India implementation. Dahikar and Rode[5] in their research provided the datasets of different features and applied Artificial Neural Networks for crop yield prediction. However, Dahikar fails to give a practical implementation of his proposed work. The dataset in the research done by Sanchez et al.[6] is limited to only 6000 records.

This paper focuses on the practical application of machine learning algorithms and its quantification. The work presented here also takes into account the inconsistent data from rainfall and temperature datasets to get a consistent trend. Crop yield prediction is determined by considering all the features in contrast with the usual trend of determining the prediction considering one feature at a time.

Rest of the paper is organized as follows. Section 2 gives the methodology applied for predicting crop yield by using different features. Section 3 focuses on Experimental Results and Analysis. Section 4 concludes the paper.

II. METHODOLOGY

A. Factors affecting Crop Yield

There are number of factors that influence the yield of any crop prediction. These are basically the features that help in predicting the production of any crop over the year. Some of the major factors are as follows:

1. Temperature 2. Rainfall 3. Area 4. Season

978-1-7281-0899-5/19/$31.00 ©2019 IEEE

B. Machine Learning Algorithm

Among the various factors affecting crop yield prediction temperature and rainfall are the most impactful. Rainfall and temperature data are sequential data thus time series machine learning algorithms are applied to them. The algorithms are given below:

1) Simple RNN:: A recurrent neural network (RNN) [7] is a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior.

2) LSTM:: Long short-term memory is an artificial recurrent neural network (RNN) architecture [8] used in the field of deep learning. Standard feed forward neural networks do not have feedback connection like LSTM. This makes LSTM helpful for a "general purpose computer" [9].

The crop production dataset that is used to predict the name and yield of the crop is fed into classification and regression algorithms. The algorithms applied are ensemble learning algorithms like Random Forest Classifier[10] and XGBoost[11] , KNN Classifier[12], Logistic Regression[13], Linear Regression[14] and Artificial Neural Networks[15].

III. Experimental Results and Analysis

This research focuses on enhancing yield production through various machine learning techniques. Experiments are conducted on agricultural dataset. Machine learning classifiers i.e. Random Forest, XGBoost, Logistic Regression, Linear Regression, Artificial Neural Networks are implemented inorder to find the best classifier that gives accurate predictions. These machine learning algorithms are implemented on Python 3.7 (Jupyter Notebook) using following inbuilt libraries : Numpy, Scikit-learn, Pandas and Keras.The hardware configuration for impelmentation are 8GB RAM with Intel i-5 processor.

A. Datasets used

The first three datasets used were found on the official website of the Indian government. The datasets can be categorized as:

1) Temperature:: 100 years of the dataset is sorted according to the month and the districts of India. [16]

2) Rainfall:: 100 years of the dataset is sorted according to the month and the districts of India. [16]

3) Production:: 17 years of dataset that is sorted according to the districts, crops, season and area. [17]

4) Final Dataset:: The temperature and the rainfall dataset was recategorized on the basis of seasons. This data was then appended according to the districts into the production dataset.

B. Parameter Settings:

1) Rainfall: In order to predict crop yield on the basis of rainfall trends following parameters are deployed in the algorithms implemented as shown in Table I.

Table I. Parameter Value of Rainfall Model

	LSTM	SimpleRNN
Layers Used	5	5
Neurons used	2024	2024
Optimizer	RMSProp	RMSProp
Loss	Mean Absolute Error	Mean Absolute Error
Epochs	40	100

2) Temperature: In order to predict crop yield on the basis of temperature trends, following parameters are deployed in the algorithms implemented as shown in Table II.

Table II. Parameter Value of Temperature Model

	LSTM	SimpleRNN
Layers Used	2	2
Neurons used	362	362
Optimizer	RMSProp	RMSProp
Loss	Mean Absolute Error	Mean Absolute Error
Epochs	50	40

3) Crop Name Prediction: In order the predict the name of crops, following classifiers along with the corresponding parameter values are implemented.

a) Random Forest Classifier:

- min_samples_leaf = 1
- min_samples_split=2
- n_estimators=10
- criterion = mse

b) Artificial Neuron Network:

- Optimizer : sgd
- Learning_rate : 0.1
- Layers : 3087
- Loss : mean_absolute error

c) K-Nearest Neighbours Classifier:

- n_neighbors : 24
- leaf_size : 30
- metrics : minkowski

d) XGBoost Classifier:

- max_depth : 4
- eta : 0.4
- n_round : 25

e) Stochastic Gradient Descent Classifier:

- learning_rate : optimal
- validation_factor : 0.1
- n_iter_no_change : 5

4) Crop Yield Production: In order the predict the yield of crops,following classifiers along with the corresponding parameter values are implemented.

a) Random Forest Regressor:

- fit_intercept : True
- n_jobs : None
- n_estimators : 70
- max_feature : sqrt

b) K-Nearest Neighbour Regressor:

- n_neighbors : 500
- algorithm : ball_tree
- leaf_size : 30

c) Artificial Neuron Network:

- Layers :385
- Optimizer : Stochastic Gradient Descent
- Loss : mean_absolute error

C. Experimental Result and Analysis

1) Temperature and Rainfall: Models are created district wise and time series algorithms are trained by taking a look back of 10 years.

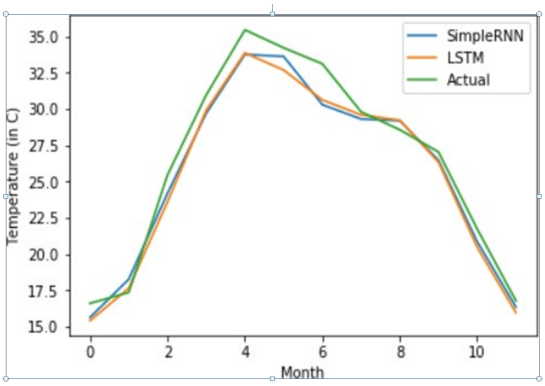

Fig 1. Deviation of predicted temperature with actual values.

Table III. Results of Models applied on Temperature Dataset

MODEL	MEAN ABSOLUTE ERROR
Simple RNN	0.9202
LSTM	0.8137

Fig 1 gives deviation of temperature predicted by Simple RNN and LSTM from the actual one. It can be observed from Table III and Fig 1 that prediction given by LSTM is more close to the actual data as it has a low mean absolute error.

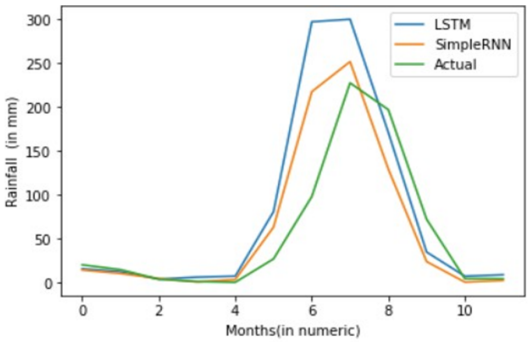

Fig 2. Deviation of predicted rainfall with actual values.

Table IV. Results of Models applied on Rainfall Dataset

MODEL	MEAN ABSOLUTE ERROR
Simple RNN	22.17
LSTM	34.14

Fig 2 gives deviation of rainfall predicted by Simple RNN and LSTM from the actual one. It can be seen from Table IV and Fig 2 that prediction given by SimpleRNN is better more close to the actual data as it has low mean absolute error than LSTM.

2) Crop Name Prediction: Fifteen widely grown crops in India were selected and their name was predicted on the basis of area, production, rainfall, season and temperature. Different classifiers were trained on the dataset and confusion matrix was plotted to showcase the performance of a model for which the true values were known.

Rainfall and Temperature predicted by the above algorithms are then used in crop name prediction and crop yield prediction. Crop Yield data instances are grouped season wise (Kharif, Rabi) and used as an input feature for crop name prediction. This will provide us an insight regarding the crop to be sown in next season.

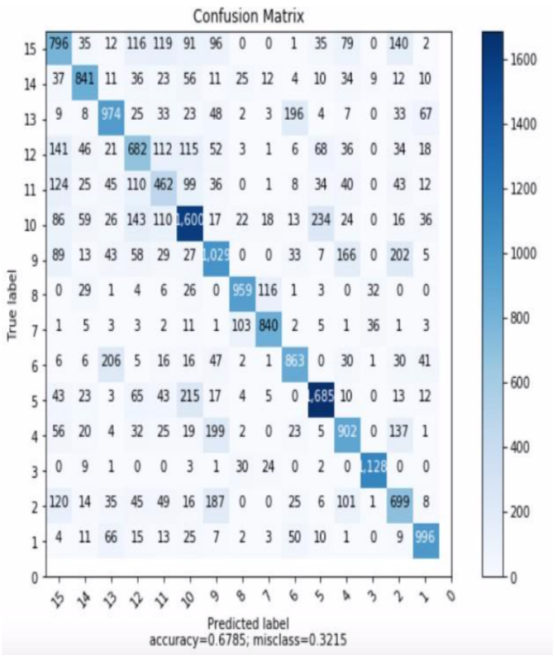

Fig 3. Confusion Matrix of Random Forest Classifier

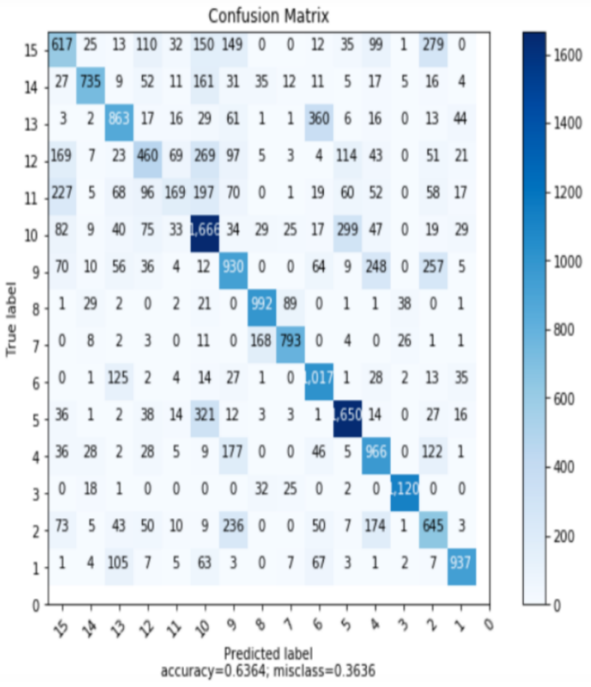

Fig 4. Confusion Matrix of XGBOOST Classifier

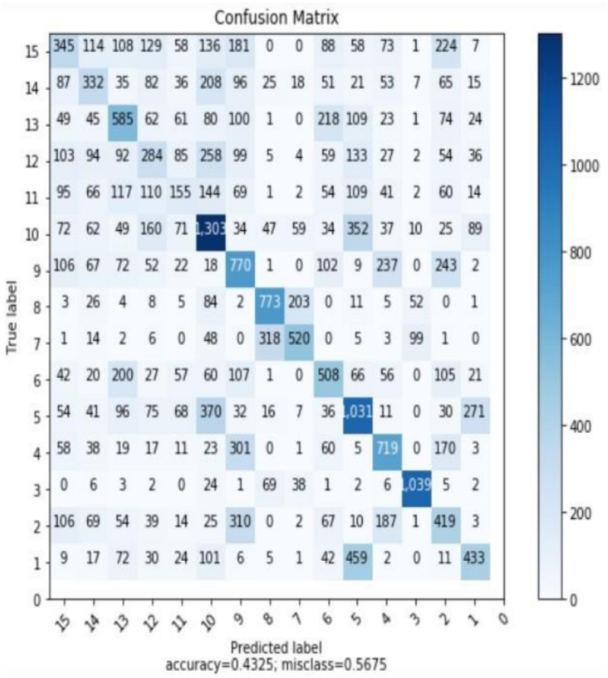

Fig 5. Confusion Matrix of KNN Classifier

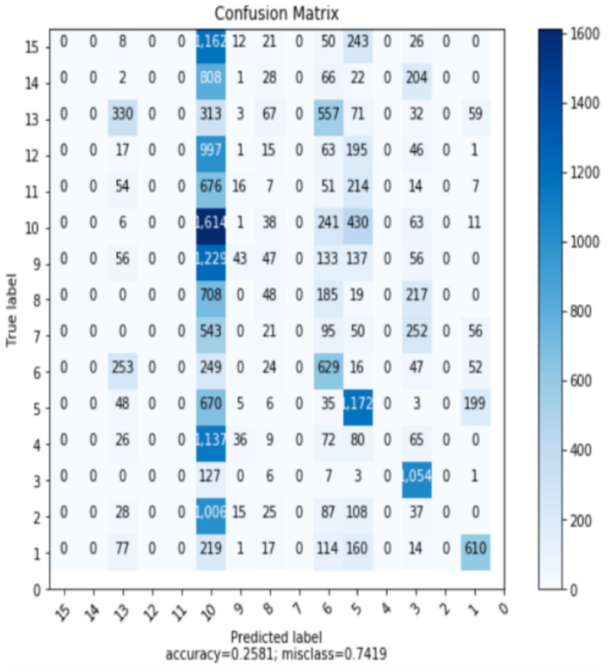

Fig 6. Confusion Matrix of Logistic Regression

Fig 3, Fig 4, Fig 5 and Fig 6 shows confusion matrix of Random Forest Classifier, XGBoost Classifier, KNN Classifier and Logistic Regression respectively. Each confusion matrix depicts the relationship between actual and predicted values. The vertical axis denotes actual values and the horizontal axis denotes predicted values. The range of the axes is from 1 to 15 indicating the 15 widely grown crops. The darker the value of the diagonal cells in the matrix, the

more the accuracy.

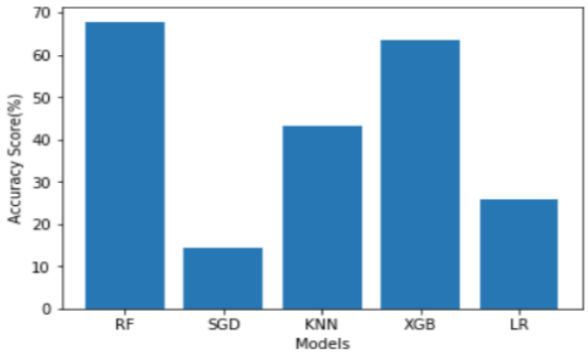

Fig 7. Comparison of Various Models

Table V. Accuracy of Models Applied

MODEL	Accuracy (in percentage)
Random Forest Classifier	67.80
XGBoost Classifier	63.63
KNN Classifier	43.25
Logistic Regression	25.81

It can be noticed from Table V and Fig 7 , Random forest Classifier gives better accuracy and thus outperformed all the other techniques.

3) Yield Prediction: This focuses on district wise yield prediction according to the crop sown in the district. Yield is being predicted for given crops district wise and crops with best yield is suggested to grow in future so that farmers can get maximum benefit from the analysis.

$$Yield = Production/Area \tag{1}$$

Various regression models were used to predict production using the dataset[17].

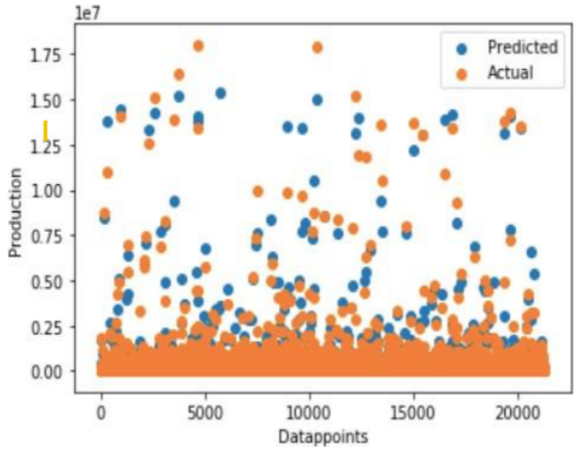

Fig 8. Scatter Plot of Predicted Vs Actual Yield Prediction

Fig 8 portrays the mapping of predicted and actual values of crop production. The range of the X axis was from 0 to 20000(approximately) and Y axis ranged from 0 to 1.75*(1e7).

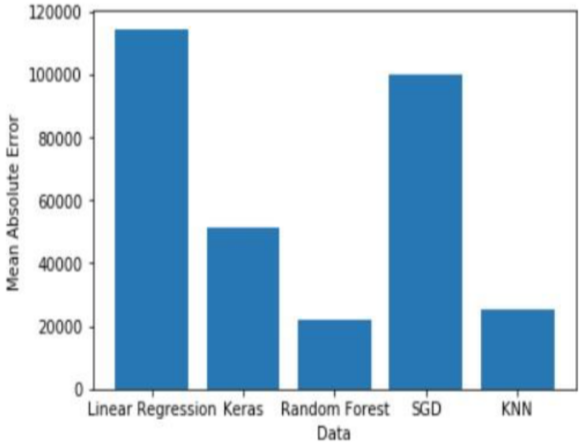

Fig 9. Mean Absolute Error of various Algorithms

Fig 9. depicts that Random Forest Regressor shows the least mean absolute error among all other machine leaning algorithms. Thus, most suitable for crop yield prediction.

D. Conclusion

The paper presented the various machine learning algorithms for predicting the yield of the crop on the basis of temperature, rainfall, season and area. Experiments were conducted on Indian government dataset and it has been established that Random Forest Regressor gives the highest yield prediction accuracy. Sequential model that is Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction. By combining rainfall, temperature along with other parameters like season and area, yield prediction for a certain district can be made. Results reveals that Random Forest is the best classifier when all parameters are combined. This will not only help farmers in choosing the right crop to grow in the next season but also bridge the gap between technology and the agriculture sector.

REFERENCES

[1] Agriculture Role on Indian Economy Madhusudhan L - https://www.omicsonline.org/open-access/agriculture-role-on-indian-economy-2151-6219-1000176.php?aid=62176
[2] Priya, P., Muthaiah, U., Balamurugan, M. International Journal of Engineering Sciences Research Technology Predicting Yield of the Crop Using Machine Learning Algorithm.
[3] Mishra, S., Mishra, D., Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper.Indian J. Sci. Technol,9(38), 1-14.
[4] Manjula, E., Djodiltachoumy, S. (2017). A Model for Prediction of Crop Yield.International Journal of Computational Intelligence and Informatics,6(4), 2349-6363.
[5] Dahikar, S. S., Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach.International journal of innovative research in electrical, electronics, instrumentation and control engineering,2(1), 683-686.

[6] Gonzlez Snchez, A., Frausto Sols, J., Ojeda Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction.
[7] Mandic, D. P., Chambers, J. (2001). Recurrent neural networks for prediction: learning algorithms, architectures and stability. John Wiley Sons, Inc..
[8] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[9] Sak, H., Senior, A., Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.
[10] Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
[11] Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.
[12] Cover, T. M., Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.
[13] Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., Klein, M. (2002). Logistic regression. New York: Springer-Verlag.
[14] Seber, G. A., Lee, A. J. (2012). Linear regression analysis (Vol. 329). John Wiley Sons.
[15] urada, J. M. (1992). Introduction to artificial neural systems (Vol. 8). St. Paul: West publishing company.
[16] http://www.indiawaterportal.org/met_data/
[17] District-wise, season-wise crop production statistics Vkhullar - https://data.gov.in/catalog/district-wise-season-wise-crop-production-statistics