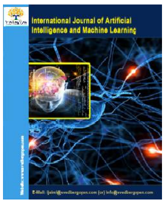
SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

An optimized machine learning approach for predicting various crop yields

Mahender Reddy Sheri[1*], Sriman Naini[2] and Sai Kiran Thatipamula[3]

[1] Otto-Friedrich University of Bamberg, Germany. E-mail: mahendersheri@gmail.com
[2] Rosenheim University of Applied Sciences, Germany. E-mail: srimannaini141@gmail.com
[3] National Institute of Technology, Trichy, Tiruchirappalli, Tamil Nadu, India. E-mail: saikithati83@gmail.com

Article Info

Volume 1, Issue 1, July 2021
Received : 24 February 2021
Accepted : 18 June 2021
Published : 05 July 2021
doi: 10.51483/IJAIML.1.1.2021.18-23

Abstract

Agriculture being the most essential and crucial thing for the mankind as well as for the economy for the countries like India, various crop patterns and their yearly production statistics derives many conclusions for many places where the actual prediction for the crop yield plays a major role with respect to certain factors concerned, in our work we optimize the real time data and use machine learning approaches such as random forest, multilinear regression, normalization and pearsons correlation coefficient for the prediction of yield concerned to the state of Telangana considering the factors such as temperature, humidity, underground water, canals, soil type, season etc. Our model is helpful for more accurate prediction of the yield for different crops for a farmer friendly and profitable cultivation. As the algorithms used are of more supervised and powerful it gives the best results for the user.

Keywords: *Machine learning, Crop yield, prediction, Regression, Random forest, Normalization*

1. Introduction

Agriculture being played as a major role in the economy and livelihood of a huge country as india, it makes more interesting for the data scientists to model and predict the crop yields at various locations. As there are always more optimistic and reliable patterns followed from the generations in india, they were more helpful previously but the fact of increasing drastic changes in climatic conditions as well as the weather changes, population expansion, usage of pesticides, availability of water, requirement of the crop leads to more helpful ways for the farmers for a proper productive yield of the crops. There are many prediction models already being researched and on use in various fields as in retail business, stock market, banking sector, medical field, pharmaceutical field with considerable factors as per the required output.

In the field of agriculture the analysis and evaluation of the historical data plays a key role in maintaining proper standards of crops and their yield however the methods to be considered should be in a way that relates all of the

* *Corresponding author: Mahender Reddy Sheri, Otto-Friedrich University of Bamberg, Germany.*
E-mail: mahendersheri@gmail.com

existing conditions for any certain type of crop with available resources. There are many machine learning techniques used for classifying and predicting a defined dataset by exploring the data in terms of preprocessing, i.e, data cleaning and data modification, analyzing, data reduction, applying algorithms and at last summarizing and obtaining the graphical representation of the given data and the results. There are certain standards to be incorporated for and every algorithm in order to differentiate its performance with others such as *R*-square value for regression models.

In the state of Telangana, India, there is a rapid changes in the resources available for cultivation, such as, the worlds largest lift irrigation project (Kaleshwaram project) providing water to the fields throughout the year and making the farmers to move from water independent crops, such as, cotton to the water dependent crops, such as, paddy but it lacks with the production factors and competency between various crops leading to over production, water availability from different sources, i.e, rain, canals, underground, rivers etc.

The study of such patterns and providing the prediction system makes the farmer to prevent loses. The current work provides the information of the yield of particular crop based on the factors such as area of production, available resources for the crop, i.e., water from various sources by applying regression techniques as well as random forest and also gives the summarized details of other factors to be considered for the future enhancement. It is necessary for a farmer to have an insight of crop yield for an efficient control and management of the cultivated crops which leads to provide some useful techniques for the usage of prediction systems.

2. Literature Review

Machine learning makes most of the prediction and recommendation systems up to date with more efficient methods to run smoothly with variable competing accuracies between different algorithms and approaches used. An approach for different crops based on various factors are discussed by Kodimalar *et al.* (2019) and given an insight of various machine learning algorithms, such as, support vector machines, artificial neural networks, random forests based on the metrics, such as, root mean squared error, mean squared error, and mean absolute error.

Unlike other techniques machine learning deals with the training of data with the given features and as it is a part of artificial intelligence it forms a foundation in building a strong prediction system which may be automated by realtime attributes. Machine learning consists of clustering, classification and regression techniques with variety of factors such as *r*-value, *t*-value, precision etc., One such technique based on support vector machine for rice is demonstrated by Su *et al.* (2017) at its development stages and yield prediction by classifying the rice at different stages of production, i.e., early, middle season, late season rice where they have used various kernel functions for training and later cross fold validation to test and model the data. These insights provide to approach various other machine learning approaches, such as, neural networks, random forests for the betterment or availability of the required output prediction models. There are some softwares developed such as cropadvisor (Veenadhari *et al.*, 2014) but it is limited to a specific region of Madhya Pradesh state where it provides the influence of climatic factors on each other for the final crop yield using *k*-means, ANN, SVM techniques by considering additional factors, such as, atmospheric pollution, geographical location where the author adopted his own featured decision tree algorithm named as C4.5.

Regression techniques are the most common and efficient approaches used for the prediction systems considering the predictor or independent variables which are responsible for predicting the expected output. A normal regression equation is in the form of:

$$Y = mX + c$$

where, *X* is the independent variable, and *Y* is the variable which is dependent. This method has *R*-squared value which is responsible for finding the accuracy of the given model for the provided dataset. Other factors such as beta coefficients and residual terms stating the change or variation between the actual and predicted values.

Decision trees consisting of nodes, branches are used for classification of the data as well as regression. For the state of Telangana for the year 2020 the state government has recommended to grow paddy of certain kind and few other crops as well but the fact of over production of rice as well as excess rainfall lead to loss of final yield due to the crop damage which is a considerable factor for the farmer where the yield was lower than the expected and so the recommended system failed. This leads to consider various approaches based on the historical data and other influencable factors, such as, population, demand, usage, dependent products and also expected growth in areas based on the yearly data and available crop utilities such as seeds, pesticides etc.

3. Approach

A proper historical data is collected from the state of Telangana and the central board of agriculture properly and the collected data is explored in terms of relativity such as year of production, season of production by merging all the related data. There are fields such as higher, lower, average temperature which are not so important factors and so pearsons correlation coefficient is used in order to remove the less important features and also leastly corelated features. Later the data is explored redistributed properly by normalizing the given data within a specific range for the fields of production, area, rainfall, humidity, underground water level, temperature.

The final dataset obtained after all the preprocessing and data modification is qualitatively analyzed for predicting the crop yield by using machine learning techniques such as multi linear regression and random forest. The output field for the prediction is the production in tons for certain inputs, such as,, area, rainfall which are predictor or independent variables. For random forest is based on n number of decision trees where decision tree represents the classification of data in terms of different features represented as nodes for test from root to leaf level in a tree format and each branch used to provide the output of a given test. The selection of root is important for a good model it is selected on the basis of reduced standard deviation reduction, i.e., higher the value more likely to be considered as a root node which are calculated for the given labels. Here after the trees selected for the random forest are taken randomly for classification and prediction.

4. Implementation and results

For the implementation of the algorithm training and testing data are divided in the ratio of 70 and 30 where it is further reconsidered by changing the ratios of train vs. test for better results. The overall data is summarized by calculating the

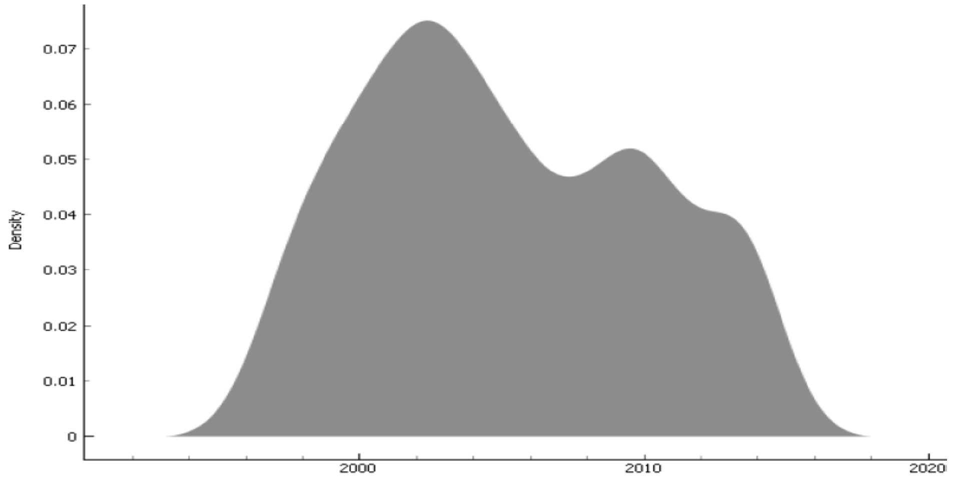


Figure 1: Production

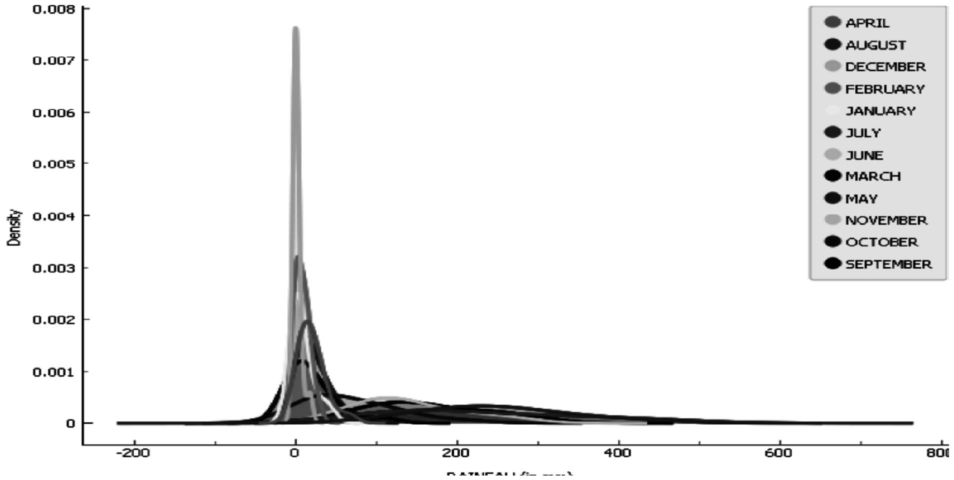


Figure 2: Rainfall

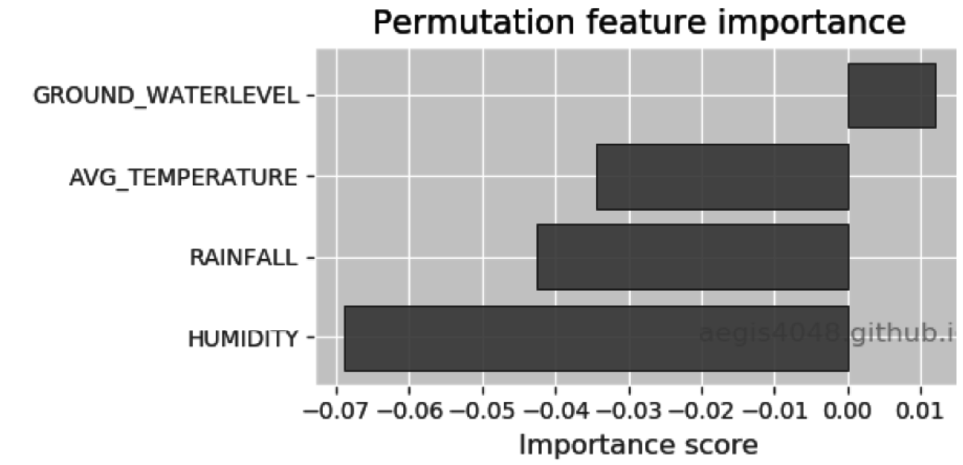


Figure 3: Feature importance

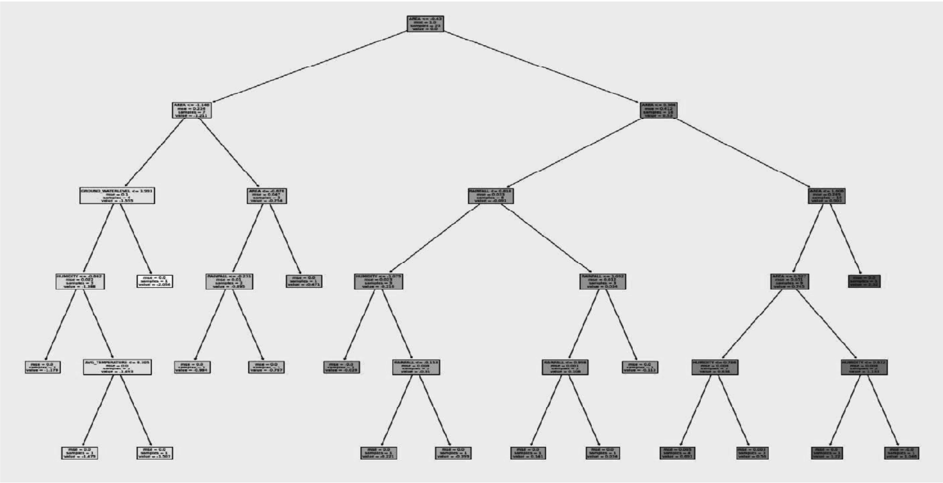


Figure 4: Sample decision tree

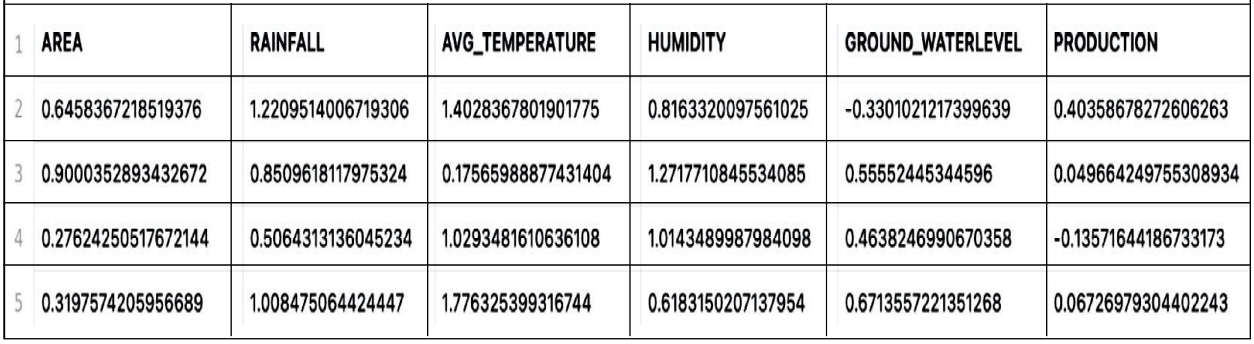
Table 1: Correlation matrix

	Area	Rainfall	Avg_Temperature	Humidity	Ground_waterlevel	Production
Area	1.000000	0.190990	0.147645	0.170493	–0.143155	0.953746
Rainfall	0.190990	1.000000	0.828912	0.875114	–0.153803	0.132751
Avg_Temperature	0.147645	–	1.000000	0.723408	–0.146155	0.082748
Humidity	0.170493	0.828912	0.723408	1.000000	–0.093422	0.096442
Ground waterlevel	–0.143155	–0.875114	–0.146155	–0.093422	1.000000	–0.169040
Production	0.953746	0.132751	0.082748	0.096442	–0.169040	1.000000

Table 2: Sample data

	Area	Rainfall	Avg_Temperature	Humidity	Ground_waterlevel	Production
0	108608	8.18	24.1	64.0	4.71	376544
1	141154	166.00	26.4	74.3	5.86	429108
2	103207	10.98	24.1	60.1	6.07	332223
3	133365	137.04	28.0	73.0	5.67	345015
4	48545	7.30	24.6	64.0	6.47	152965

Table 3: Normalized data



1	AREA	RAINFALL	AVG_TEMPERATURE	HUMIDITY	GROUND_WATERLEVEL	PRODUCTION
2	0.6458367218519376	1.2209514006719306	1.4028367801901775	0.8163320097561025	-0.3301021217399639	0.40358678272606263
3	0.9000352893432672	0.8509618117975324	0.17565988877431404	1.2717710845534085	0.55552445344596	0.049664249755308934
4	0.27624250517672144	0.5064313136045234	1.0293481610636108	1.0143489987984098	0.4638246990670358	-0.13571644186733173
5	0.3197574205956689	1.008475064424447	1.776325399316744	0.6183150207137954	0.6713557221351268	0.06726979304402243

mean, variance, standard deviation in order to obtain an overview of the data. Later multilinear regression is applied for the prediction of production for different crops based on the independent variables, i.e., underground water level in meters, rainfall in centimeters, temperature in Celsius, area in acres. The *R*-square value obtained for multilinear regression is 0.9293 and 0.713 for random forest. Below are some plots of rainfall in year and production in years, i.e, stating the density of values from low to high where rainfall represents high in July and low in December as well as overall production is higher in initial 2000's than to later 2020. The correlation matrix below represents the interdependency of the features. There are many crops considered for the prediction where each crop is grouped before predicting the final yield. Figures 1 to 4 and Tables 1 to 3 are the sample for rice. The predicted production for area = 15000, rainfall =10, average temperature = 30, humidity = 50, Underground water level = 2 is 14,839 tons with an accuracy of 92% for regression and for random forest.

5. Conclusion

Production of various crops in the state of Telangana for the various available resources is properly predicted with the expected crop yield for the given conditions and helps to choose the crop with the given resources with more accuracy and hence reducing the losses incurring due to the shortage of resources for the crop at that particular area.

6. Future scope

More prediction techniques and optimization techniques can be used by considering the other factors such as new irrigation projects proportion of expected income for the available production, i.e, over production or under production as well as the demand of the crop in the market.

References

Arun, Kumar., Naveen, Kumar., and Vishal, Vats. (2018). Efficient crop yield prediction using machine learning algorithms. *International Research Journal of Engineering and Technology*, 5(6), 3151-3159.

Jayaram, M.A., and Marad, Netra. (2012). Fuzzy inference systems for crop yield prediction. *Journal of Intelligent Systems*, 21(4), 363-372.

Medar, R., Rajpurohit, V.S., and Shweta, S. (2019). Crop yield prediction using machine learning techniques. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). Bombay, India, 1-5. doi: 10.1109/I2CT45611.2019.9033611.

Palanivel, Kodimalar., and Surianarayanan, Chellammal. (2019). An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, 10(3), 110-118.

Rale, N., Solanki, R., Bein, D., Andro-Vasko, J., and Bein, W. (2019). Prediction of crop cultivation, 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). 0227-0232. Las Vegas. NV. USA. doi: 10.1109/CCWC.2019.8666445.

Sellam, V., and Poovammal, E. (2016). Prediction of crop yield using regression analysis. *Indian Journal of Science and Technology*, 9(38), 1-5.

Shahhosseini, M., Hu, G., Archontoulis S.V., and Huber, I. (2021).Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci Rep.*, 11, 1606.

Su, Y.H., Xu, H., and Yan, L.J. (2017). Support vector machine-based open crop model (SBOCM): Case of rice production in China. *Saudi J Biol. Sci.*, 24(3), 537-547.

Thomas, van, Klompenburg., Ayalew, Kassahun., and Cagatay, Catal. (2020). Crop yield prediction using machine learning: A systematic literature review, *Computers and Electronics in Agriculture*, 177, 105709.

Veenadhari, S., Misra, B., and Singh, C. (2014). Machine learning approach for forecasting crop yield based on climatic parameters, 2014 International Conference on Computer Communication and Informatics, 1-5. Coimbatore. India. doi: 10.1109/ICCCI.2014.6921718.

Cite this article as: Mahender Reddy Sheri, Sriman Naini, and Sai Kiran Thatipamula. (2020). An optimized machine learning approach for predicting various crop yields. *International Journal of Artificial Intelligence and Machine Learning*, 1(1), 18-23. doi: 10.51483/IJAIML.1.1.2021.18-23.