ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter58872020.pdf
https://doi.org/10.30534/ijeter/2020/58872020

Towards a Unified Approach for Crop Yield Prediction

Shilpa Mangesh Pande[1], Dr. Prem Kumar Ramesh[2]
[1]Associate Professor, Department of Information Science and Engineering, Research Scholar, Department of Computer Science and Engineering (VTU-RC), CMR Institute of Technology, Bengaluru, India and affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, shilpa.p@cmrit.ac.in[1]
[2]Professor, Department of Computer Science and Engineering, CMR Institute of Technology, Bengaluru, India and affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, premkumar.r@cmrit.ac.in[2]

ABSTRACT

Agriculture is one of the fundamental occupations for majority of the countries in the world. Especially, in developing nations like India, the country is primarily driven by agriculture sector, where agriculture and its associated businesses are the backbone of the Economy making it the integral revenue generator. With technological advancements in the recent years, crop yield prediction has gained wide importance, and has shown to have significant impact on the revenue generated from agriculture in every season. Multiple factors influence crop yield prediction, which in turn makes it a non-trivial and challenging task. Despite many proposed works in the area, crop yield prediction lacks a unified solution. This paper brings out the need for a unified framework through a comparative study of standard algorithms and attributes. The algorithms considered are Linear Regression, Random Forest, K Nearest Neighbors (KNN) and Stochastic Gradient Descent (SGD). Our results show that Random Forest outperforms the other standard algorithms by showing 91.62% accuracy in crop yield prediction. Further evaluation is done where the attributes that affect the crop yield most are ranked according to their impact based on Mean Absolute Error (MAE). With this, we make a case for the need for a unified approach for crop yield prediction.

Key words : Crop Yield, Linear Regression, Random Forest, K Nearest Neighbors (KNN) and Stochastic Gradient Descent (SGD), Machine Learning, Generic Solution.

1. INTRODUCTION

Agriculture is one of the important sources of income in India [3]. Farmers grow crops which suits for the environmental condition of their region. Different types of crops are grown in different region because each crop region is defined based on homogeneity in the soil type, climatic conditions, rainfall, farming practices, availability of labor, irrigation facilities and many more [1]. Crop yield prediction is a fundamental agriculture problem to solve. In traditional farming, most of the farmers generally predict the crop yield based on their experience on the specific crop. This ad-hoc approach of crop yield prediction is unreliable and is prone to errors, as the multiple changing environmental factors drastically affect crop yield. In the recent years, various data analytics techniques have been applied in the field of farming to predict the crop yield accurately, and farmers are provided timely advice for planning the future crops boosting up the crop yield as well as quality of crops [2]. Machine Learning algorithms enables handling of vast data, does intelligent analysis by learning patterns in the data and thus guide to improved decision making without human involvement.

A significant amount of research work has been carried out in field of Agriculture for accurate prediction of crop yield for different type of crops [3]. Despite the availability of many such models, the farmers in developing nations, such as India, lack the knowledge of these available scientific tools which would guide them for getting best possible farming results. India is a land of diversity in physical features, season, crops, climate and culture, which directly or indirectly influence the annual yield of food crops. For this reason, this paper focuses on crop yield prediction in India. However, the authors believe the same model is applicable globally.

This paper provides an analysis of the various solutions available for crop yield prediction technique and attempts to bring out the factors required to develop a generic solution for crop yield prediction. Through a detailed survey of various prediction techniques this work attempts to justify a need for a unified framework for crop yield prediction. Specifically, the major contributions are enlisted below.

1. Study of various solutions proposed/used in crop yield prediction, and the effectiveness of the various parameters that influences their results.
2. Comparative analysis of standard algorithms on crop yield prediction and identifying the most suitable algorithm for a generic set of crops
3. Study of the impact of geographical location on crop yield prediction.
4. Evaluation of various parameters which affects the crop yield and ranking them according to their impact.

To pursue this, various standard algorithms and factors/attributes are considered. First, the geography of different states which may have different climatic conditions affects the crop yield. To get more reliable results diverse states of India which are geographically apart, such as Punjab

(north), Mizoram (east) and Karnataka (south/west), are chosen. Standard ML models, such as Linear Regression, Random Forest, and SGD KNN algorithms are considered for crop yield prediction. Our results indicate Random Forest Regression yields the best performance with a dataset containing multiple parameters and different crops, across independent of geographical locality. This is a significant find, considering the multitude of models each claiming to work for specific crop at specific regions. The paper also provides a weightage to various factors that influence the crop yield, allowing researchers to discard insignificant attributes. In summary, the paper provides solid evidence towards the development of a unified framework for crop yield prediction. The rest of the paper is organized as follows: Section 2 captures a background study of scholars emphasizing their contributions to the field of farming, crop yield prediction, and analysis. In section 3, a standard model is proposed outlining the details of crop yield prediction model, identification of the right data, and its application to generate results. The evaluation and results are in Section 4. The paper concludes in Section 5 with glimpse of future work

2. BACKGROUND

This section captures survey on previous published research work. A tabulated summary of the same capturing the typical parameters, and the algorithms used to predict the crop yield for the selected crops are presented.

D. Ramesh and team proposed a crop yield prediction for a specific region, the East Godavari district of Andhra Pradesh, India for rice, as it is the prominent crop in that state [2]. MLR (Multi Linear Regression) algorithm was selected as it models the linear relationship between a dependent variable with one or more independent variables. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm was proposed to make approximately six clusters. The results achieved using MLR were verified using DBSCAN. The predicted results using MLR algorithm varied between -14% to +13%, whereas for DBSCAN the range was between -13% and +8%. Rainfall was used are one of the primary input parameters.

In spite of achieving good results, the model is not applicable throughout the country, as there are many regions where the primary mode of irrigation does not depend on rainfall. Thus, there were models proposed to suit this particular scenario. For instance, E. Manjula et al. considered tanks, bore-wells and open-wells for the Tamilnadu region (a southern state in India), for rice yield prediction as this region gets lesser rainfall compared [4]. A Data mining technique based on association rules has been used to predict rice yield prediction for selected regions of Tamilnadu. Using K-means algorithms clusters were formed on which association rule mining was applied. “Apriori” algorithm played important role for finding frequent pattern mining. The experiment was successful with 6 frequent patterns with highest support value of 0.993. However, like other former works, there was still a need for selecting optimal parameters to maximize the crop yield.

To address this, Majumdar et al. opted to find optimal parameters to achieve higher production with various data mining techniques [5]. The optimal parameters proposed here are optimal temperature (25.4^{o}C to 29.9^{o}C**),** worst temperature (30.2^{o}C to 31.15^{o}C**),**) and best rainfall (548-580mm). The research work was carried out using Karnataka region, a State in India, for various crops like cotton, groundnut, jowar, rice and wheat crop. PAM (Partitioning Around Medoids**)** and CLARA (Clustering LARge Application**)** algorithms were applied to form clusters based on districts which produces maximum crop yield. DBSCAN algorithm was shown to yield better clustering than PAM and CLARA. However, CLARA showed significant results compare to PAM.

Most of the related works in crop yield prediction were developed on standalone systems without any standard application service designed [6]. Agricultural cloud framework was very much needed in order to provide selected services. Shastry et al. proposes two services, viz., Soil Classification-as-a-Service and Crop Yield Prediction-as-a-service. The main advantage of deploying services to the cloud is availability of a software interface to the users anytime, and inherent security with proper backup associated with the cloud infrastructure. In this work, wheat yield prediction was considered for Karnataka region using a customized ANN (Artificial Neural Network). Amazon S3 was used to store the data while for deployment is done on Heoku cloud. The soil classifier (M-SVM-Support Vector Machine) and crop yield predictor (M-ANN) provided accurate and reliable services. The parameters used for prediction model are soil data, crop data, weather data and fertilizer data.

In similar lines, Tanuja and B.V. Pawar developed an application portal through which farmers can fill required information for crop yield prediction for Maharashtra region [8]. The model not only predicted the crop yield but also suggested the farmers which seasonal crops they plan with maximum yield and profit. Prediction model is built using ANN and SVM. ANN model performed far better that SVM model for the given dataset with 86.8% accuracy. The results of this research is helpful for the farmers for proper selection of crops.

A host of other studies [13][16][18][19][20][21][22] have been proposed applying different features, and feature selection algorithms. Maya Gopal and R. Bhargavi applied different feature selection algorithms to select most prominent five features out of 16 features for better prediction accuracy [7]. The chosen features include area, number of open wells, tanks, canal length and maximum temperature. The proposed model predicts paddy yield for Tamilnadu state, India. The intrinsic relationship between MLR and ANN is established. The hybrid MLR-ANN model was designed for the paddy yield prediction. The standard performance matrix was used to calculate the prediction accuracy. The results were compared with various conventional methods like KNN,

random forest, SVR (Support Vector Regression). The results proved that the hybrid MLR-ANN model gives better accuracy than the conventional models.

Current literature suggests that rainfall, temperature, precipitation, fertilizers and soil type play a very important role in crop yield prediction model [14][15][17]. It has been observed that most of the research work done in India has been done on typical crops such as Rice, Wheat, and Maize [2][4][7][11].

Even though many works have contributed towards this area of research, there is still a gap due to the lack of a unified framework to predict crop yield. Every crop is having different requirement for soil, climate, temperature, rainfall, and many others. This makes the problem even more challenging. As a first step to bridge this gap, this paper presents a comparative study of various standard parameters and highlights the attributes that should constitute in a unified model.

Table 1 depicts various parameters and algorithms used in crop prediction model.

Table 1: Parameters and Algorithms used in Published Papers to Predict the Crop Yield

References	[2]	[10]	[11]	[9]	[4]	[8]	[4]	[7]
Crops / Parameter	Rice	Wheat Maize	Wheat Rice Soyabean	Barley	Paddy Ragi Cumbu	Kharib Rabi	Rice	Paddy
Precipitation	✓	✓		✓			✓	
Rainfall	✓	✓				✓	✓	✓
Humidity		✓						
Temperature		✓		✓		✓	✓	✓
PH						✓	✓	
Soil Type		✓	✓			✓	✓	
Yield	✓	✓			✓		✓	✓
Area of Sowing	✓				✓			✓
Fertilizer	✓		✓				✓	✓
Ground Water Level		✓						
Year	✓				✓	✓	✓	
Seed Quality			✓					
Proposed Model	MLR DBScan	RMS Error	CSM	MLR	K-means Association Mining Rules	ANN SVM	M-SVM M-ANN	Hybrid MLR ANN

3. PROPOSED SYSTEM

Standard Machine Learning techniques are available to solve problems where the relationship between input variables and output variables is hard to obtain or is not known. In this section, a simple framework using machine learning algorithms is proposed and implemented in a generic sense to predict the crop yield as show in Figure 1. Geographically apart regions, namely, Karnataka, Punjab and Mizoram states of India are selected. Data is collected from various reliable and available standard sources[2]. Collected data consists of 10 years of data for all the regions and districts of India. Collected data is prepared for data set. Data sets used for demonstrating the proposed model are crop, rainfall and temperature. Considering the data associated with all the above parameters, a final data set is obtained. Collected data is preprocessed. The data preprocessing is very important and critical step for Machine Learning Algorithms. In preprocessing step raw data is filtered out for Karnataka region, Punjab region and Mizoram regions. In raw data, every field is not important for a selected algorithm. Data cleaning takes care of this. In the proposed model depicted in Figure 1, the functionalities implemented in the preprocessing step are:

- null values are removed, and missing values are replaced with mean values
- used “labelencoder” function for changing categorical data to numerical data
- used “standard_scalar” function for various distributed ranged data values to specific ranged values

The preprocessed data set is split into 2 distinct sets namely train set and test set with division ratio of 80% to 20%. Random forest regression, linear regression, K Nearest Neighbor (KNN) regression, and Stochastic Gradient Descent (SGD) algorithms are implemented. For each of these algorithms predicted value is calculated and compared against the actual value reported in the data set. All selected algorithms are used to predict the values in Karnataka state, Punjab state as well as Mizoram state. All these states are geographically diverse states of India hence selected to verify the accuracy of the algorithms. A superset of all the parameters which influences the crop yield was identified for the given dataset. The working of the proposed model includes historical data that has been collected from various reliable resources[†]. On the acquired data, the data cleaning is done, and the raw data is preprocessed. Preprocessed data is spilt into training and test data and the decision model is built.

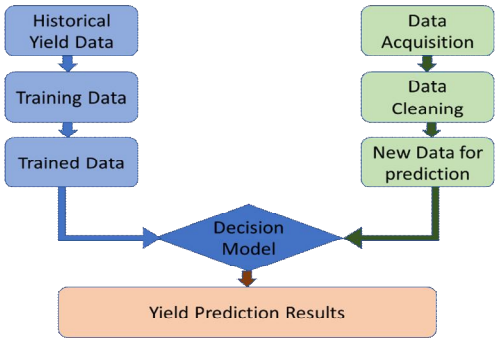

Figure 1: Proposed Model

4. RESULTS AND DISCUSSIONS

This section discusses results obtained by the selected algorithms on various crop sets of Karnataka Region, Punjab region and Mizoram regions of India. Parameters used for crop yield prediction are sowing area, type of crop, season, rainfall and temperature. For all the identified machine learning algorithms, Mean Absolute Error (MAE) values are calculated and compared where MAE and Accuracy are calculated with the following formula:

MAE= $((\sum^{n}_{i=1} abs(y_i - x_i))/n)$, where y is actual value and x is predicted value

[2] data.gov.in, indianwaterportal.org, kaggle.com

Accuracy = 1 – MAE

Figure 2 illustrates the MAE results for Karnataka, Punjab, and Mizoram regions and Table 2 Captures Comparison

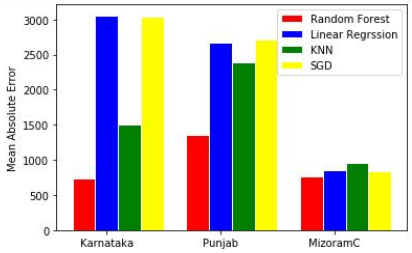

Figure 2: Comparison of Results

Table 2: Comparison with MAE values

State / Algorithm	Karnataka	Punjab	Mizoram
Random Forest	732.59	1350.76	758.22
Linear Regression	3058.48	2673.52	851.19
KNN	1500.68	2387.87	950.87
SGD	3034.49	2709.06	837.03

Figure 3 demonstrates the graphical view of actual values and predicated values of crop for Random Forest Regressor. Figure 4 confirms that the temperature is the most impactful factor in crop yield prediction on the given data set where MAE is calculated with all the selected attributes, a single attribute is taken into consideration for crop yield prediction. Towards building unified framework in addition to the temperature being a primary factor, other selected impactful factors will result in better yield prediction. This justifies there is a need for unified framework.

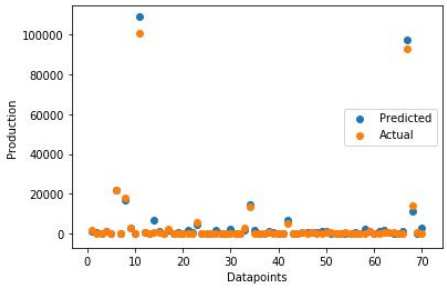

Figure 3: Predicted Vs Actual Production values for Random Forest Regressor

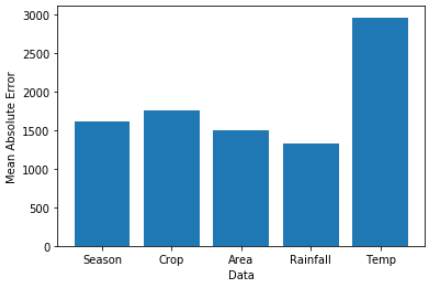

Figure 4: Most impactful factor in Crop Yield Prediction

5. CONCLUSION

This paper highlighted the various factors affecting the crop yield along with multiple solutions to predict the crop yield more accurately. Striding through the proposed model, various machine learning algorithms were implemented on available agricultural data for generic set of crops from different diverse geographical states of India to predict the crop yield. The results obtained suggests the Random Forest Regression as the best among the set of standard algorithms, which included Linear Regression, KNN and SGD. The statistical measure technique used for comparison was accuracy and MAE. It has also been shown that out of all the evaluated parameters, "temperature" is the most impactful factor that affects the crop yield. This is a first step towards unified framework, future work will involve bringing in additional factors like demand and supply and IoT for real time monitoring to establish a unified framework.

REFERENCES

1. Sagar, B. M., and N. K. Cauvery. "**Agriculture Data Analytics in Crop Yield Estimation: A Critical Review.**" Indonesian Journal of Electrical Engineering and Computer Science 12.3 (2018): 1087-1093. https://doi.org/10.11591/ijeecs.v12.i3.pp1087-1093
2. Ramesh, D., and B. Vishnu Vardhan. "**Analysis of crop yield prediction using data mining techniques.**" International Journal of research in engineering and technology 4.1 (2015): 47-473. https://doi.org/10.15623/ijret.2015.0401071
3. Shah, Ayush, et al. "**Smart Farming System: Crop Yield Prediction Using Regression Techniques.**" Proceedings of International Conference on Wireless Communication. Springer, Singapore, 2018.
4. E. Manjula,S. Djodiltachoumy, "**A Model for Prediction of Crop Yield**", International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4, March 2017
5. Majumdar, Jharna, Sneha Naraseeyappa, and Shilpa Ankalaki. "**Analysis of agriculture data using data mining techniques: application of big data.**" Journal of Big Data 4.1 (2017): 20.
6. Shastry, K. Aditya, and H. A. Sanjay. "**Cloud-Based Agricultural Framework for Soil Classification and Crop Yield Prediction as a Service**" Emerging

Research in Computing, Information, Communication and Applications. Springer, Singapore, 2019. 685-696 https://doi.org/10.1007/978-981-13-5953-8_56

7. Gopal, P. M., & Bhargavi, R. (2019). "**A novel approach for efficient crop yield prediction**" Computers and Electronics in Agriculture ma, 165, 104968
8. Tanuja K. Fegade,B. V. Pawar. "**Crop Prediction Using Artificial Neural Network and Support Vector Machine**", Springer Nature Singapore Pvt Ltd. 2020 Data Management, Analytics and Innovation, Advances in Intelligent Systems and Computing,1016, Page 311-324
9. Guan, Leran, et al. "**Yield modeling for prediction of regional whole crop barley productivity.**" Grassland Science 65.3 (2019): 179-188
10. Dubey, Swatantra Kumar, and Devesh Sharma. "**Assessment of climate change impact on yield of major crops in the Banas River Basin, India.**" Science of The Total Environment 635 (2018): 10-19 https://doi.org/10.1016/j.scitotenv.2018.03.343
11. Kumar, Rakesh, et al. "**Crop Selection Method to maximize crop yield rate using machine learning technique.**" 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM). IEEE, 201
12. Manjunatha, Manasa, and A. Parkavi. "**Estimation of Arecanut Yield in Various Climatic Zones of Karnataka using Data Mining Technique: A Survey.**" 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018
13. Dholu, Manishkumar, and K. A. Ghodinde. "**Internet of things for precision agriculture application.**" 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018
14. Soureshjani, Hedayatollah Karimzadeh, Ayoub Ghorbani Dehkordi, and Mahmoud Bahador. "**Temperature effect on yield of winter and spring irrigated crops.**" Agricultural and Forest Meteorology 279 (2019): 107664. https://doi.org/10.1016/j.agrformet.2019.107664
15. Dubey, Swatantra Kumar, and Devesh Sharma. "**Assessment of climate change impact on yield of major crops in the Banas River Basin, India.**" Science of The Total Environment 635 (2018): 10-19.
16. Fernandez-Ordoñez, Yolanda M., and Jesus Soria-Ruiz. "**Maize crop yield estimation with remote sensing and empirical models.**" 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, 2017
17. Iizumi, Toshichika, et al. "**Global crop yield forecasting using seasonal climate information from a multi-model ensemble.**" Climate Services 11 (2018): 13-23.
18. Chlingaryan, Anna, Salah Sukkarieh, and Brett Whelan. "**Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review.**" Computers and electronics in agriculture 151 (2018): 61-69
19. Tiwari, Preeti, and Piyush Shukla. "**Artificial Neural Network-Based Crop Yield Prediction Using NDVI, SPI, VCI Feature Vectors.**" Information and Communication Technology for Sustainable Development. Springer, Singapore, 2020. 585-594
20. Sahu, Shriya, Meenu Chawla, and Nilay Khare. "**An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach.**" 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017 https://doi.org/10.1109/CCAA.2017.8229770
21. Mahule, Ankit Arun, and A. J. Agrawal. "**Hybrid Method for Improving Accuracy of Crop-Type Detection using Machine Learning.**" *International Journal* 9.2 (2020). https://doi.org/10.30534/ijatcse/2020/209922020
22. V.Sudha, S.Mohan, R. Madhan Mohan **"A High Performance on Extemporize Yield of Horticulture Crops with Predictions based Water and Soil properties using Multivariate Analytics and Machine Learning Algorithms"** , *International Journal* 8.4 (2019)