# Design and Development of Machine Learning Model for Crop Yield Prediction

Taman Kumar[1], Kiran Jyoti[2], and Sandeep K.Singla[3]

[1,2,3] *Guru Nanak Dev Engineering College, Ludhiana (GNDEC), Panjab, India*

**Abstract**

Agriculture is one of the major sources of employment as well as contributor in the GDP of India. Machine learning is the latest technology which can be used to help the agriculture sector. This paper will focus in using the machine learning technique to predicting the wheat crop yield. The regression algorithms which are used in it are simple regression, gradient booster, polynomial regression and random forest. The results of every algorithm are compared with actual results in the last.

**Keywords**

Crop yield prediction, machine learning, regression.

## 1. Introduction

Agriculture is majorly adopted by population of India as a source of livelihood. Almost all industries depend on raw materials produced by agriculture. That is why agriculture and allied sectors contribute 15.4% in the GDP of India. India is second largest in producer and seventh largest exporter of agricultural goods. The boom in this sector is measured after the green revolution of 1967. The production of crops are depend on different parameters such as rainfall, irrigation, temperature, different climate conditions, quality of seeds, consumption of NPK (Nitrogen, Phosphorus, Potassium) and many more. Many changes are required in agricultural domain to improve the changes in Indian economy (Ramesh et al. 2019). The agricultural information can be extracted by two methods manual and by using computer and IT tools. However, manual methods have some limitations:

1. Biasing: The manual information is always one person’s perspective. Each and every person has their own perspective and the provided information is not fit in every situation.
2. Time delay: Delayed information is not useful.
3. Correctness: To err is human, that is why there is always probability of mistakes.
4. Reliability: All above factors affects the reliability of manual methods.

On the other hand, technology enhancements are well known for precision. Recently the most common used technological enhancements for agriculture domains are:

1. Machine Learning.
2. Deep learning.

Machine learning: Machine learning is used in many domains such as malls to predict the behavior of customer’s shopping, stock market trends, moreover it is used in agriculture fields also. There are many processes that are included in agriculture like irrigation scheduling, crop diseases, by-products, transportation etc. All procedures ultimately lead to crop yield. Despite going for mini procedures we opted for main task i.e. crop yield. Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate,

International Conference on Emerging Technologies: AI, IoT, and CPS for Science & Technology Applications, September 06–07, 2021, NITTTR Chandigarh, India

EMAIL: tamankumar0808@gmail.com (A. 1); kiranjyoti@gmail.com (A. 2); sandeepkumar.singla@gmail.com (A. 3)
ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)

weather, soil, use of fertilizers, and seed variety. This indicates that crop yield prediction is not a trivial task; instead, it consists of several complicated steps. Nowadays, crop yield prediction models can estimate the yield, but a better performance in yield prediction is still desirable (Klompenburg et al. 2020).

## 2. Literature

### A. Agricultural Information Extraction

1) **Raorane and Kulkarni (2015),** used datamining tools in crop management system. They used regression algorithms. The disadvantage is the model is not specified.
2) **Kushwaha and Bhattachrya (2015),** concluded the method which is helpful in finding the suitable crop according to the land. Agro algorithm is used in this paper.
3) **Santra et al. (2016),** used artificial neural network, decision tree algorithm and regression analysis to providing the information of crops and help in increasing the yield rate. The negative is method is not clearly specified.

### B. Crop Yield Estimation

1) **Kumar et al. (2015),** suggested the method which is helpful in improving the yield of crops. Classifications are used and the parameters are compared. The demerit is the accuracy and performance is not proper.
2) **Babu and Babu (2016),**gave method which provide solutions to some farming problems such as water and fertilizers. They have also used the agro algorithm and the accuracy is also the problem in it.
3) **Jain et al. (2017),** in their paper found the better sequence according to which the crops should be sown so that the maximum yield is extracted. Not only sequence they also used machine learning for irrigation and crop diseases.
4) **Djodiltachoumy (2017),** used K means algorithms (Clustering) on previous years data and predict yield according to that database. The demerit is they used fewer amounts of data and it is suitable only for association rule.
5) **Nigam et al. (2019),** have concluded the random forest regression gives the highest yield prediction accuracy. Simple recurrent neural network performs better on rainfall prediction while LSTM is good for temperature prediction.

### C. Machine Learning Algorithms

1) **Khairunniza-Bejo et al. (2014),** defined a method using Artificial Neural Network to help the farmers solving some of their problems. The disadvantage is the proposed method is very time consuming.
2) **Ramesh and Vardhan (2015),**used multiple linear regression method to analyze and verify the database. The demerit is this method is of less accuracy.
3) **Savla et al. (2015),** suggested the framework using Normalization, Clustering and Classification to understand the crop yield rate zones based on attributes.
4) **Sindhura et al. (2016),** also used multiple linear regression methods to predict and support the decision making in many sectors.

The comprehensive study of literature review revealed that the crop yield estimation and agricultural information extraction from the ancillary data as well as historical data is an open problem. Various machine learning models and other algorithms have been used in past for the yield estimation.

## 3. Methodology

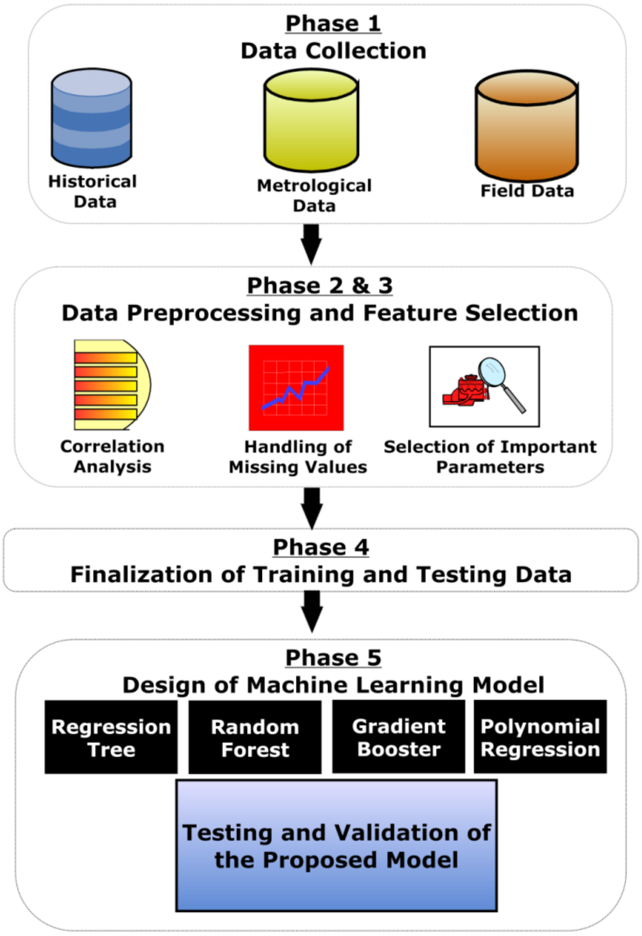
Phase 1
Data Collection
Historical Data
Metrological Data
Field Data
Phase 2 & 3
Data Preprocessing and Feature Selection
Correlation Analysis
Handling of Missing Values
Selection of Important Parameters
Phase 4
Finalization of Training and Testing Data
Phase 5
Design of Machine Learning Model
Regression Tree
Random Forest
Gradient Booster
Polynomial Regression
Testing and Validation of the Proposed Model

**Figure 1**: Flowchart of Proposed Methodology

Elaboration of methodology:

Step 1: The datasets are collected and processed.

Step 2: If there are any impurities in dataset, these are removed.

Step 3: The data is normalized if needed and can be converted into smaller volume of data.

Step 4: The data is converted into supporting format.

Step 5: Processed data is stored in the databases.

Step 6: The required method is applied.

Step 7: Final results are collected.

### A. Working of model:

i. Real time datasets of different parameters such as Precipitation, Wheat Crop Yield, NPK Consumption, Mean Temperature, Relative Humidity, Surface Pressure, Annual Rainfall is collected and downloaded from authentic sites such as data.gov.in and power.larc.nasa.gov/data-access-viewer. The area chosen is widely from Punjab, India. The variables and their respective units of measures are given below in table 1:

**Table 1**

Units of Measures

| Variable | Units of Measures |
|---|---|
| Precipitation | mm/day |
| Relative Humidity | % |
| Surface Pressure | kPa |
| Mean Temperature | C |
| Mean Wind Speed | m/s |

| | |
|---|---|
| Earth Skin Temperature | C |
| NPK Consumption | TNT |

ii. Collected data is preprocessed. There were some 'NA' values which are filled by taking average value of the above and below column.

iii. Feature selection is applied to extract important parameters for modeling framework. A process to find correlation between all the parameters is applied and the parameters which were not affecting the crop yield are eliminated. Image of correlation is given below:

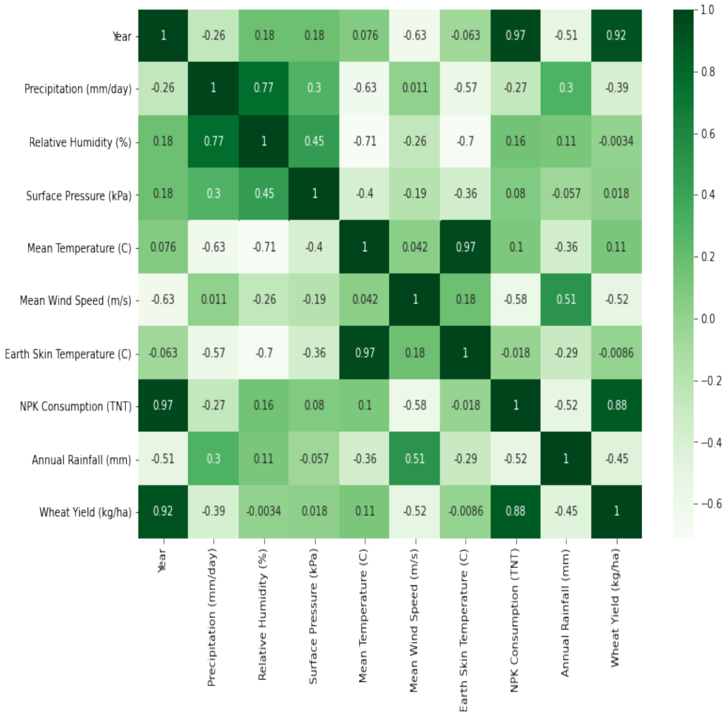

**Figure 2**: Correlation Between Variables

iv. Dataset is partitioned into training and testing set. 80% of data is used for training purpose of the model and 20% is used for testing of the model.

v. Various machine learning algorithms named as Random Forest trees, Polynomial Regression, GBM, Multiple Linear Regression and Linear Regression are implemented on the dataset to predict the output.

### B. Output

1. Results of applied machine learning algorithms are compared to evaluate the model. The table of results are given below in table 2:

**Table 2**

Comparison of Predictions with Actual Results

| Actual Results (kg/ha) | Random Forest Trees (kg/ha) | Gradient Booster (kg/ha) | Simple Regression (kg/ha) | Polynomial Regression (kg/ha) |
|---|---|---|---|---|
| 4693 | 4152.03 | 4474.749 | 4507 | 3994.286 |
| 5097 | 3943.56 | 4352.341 | 4507 | 3772.596 |
| 4724 | 4149.2 | 4184.22 | 4179 | 3449.539 |
| 5017 | 3945.25 | 3933.049 | 3853 | 3843.369 |
| 4304 | 4001.91 | 4208.408 | 4179 | 4038.906 |
| 4583 | 4277.22 | 4369.696 | 4221 | 3825.896 |
| 5046 | 4160.33 | 4224.234 | 4221 | 4004.914 |
| 5077 | 4233.55 | 4226.763 | 4207 | 3716.017 |

2. The representation of all the predicted values and actual values from year 2011 to 2018 is also given below in line and bar graph:

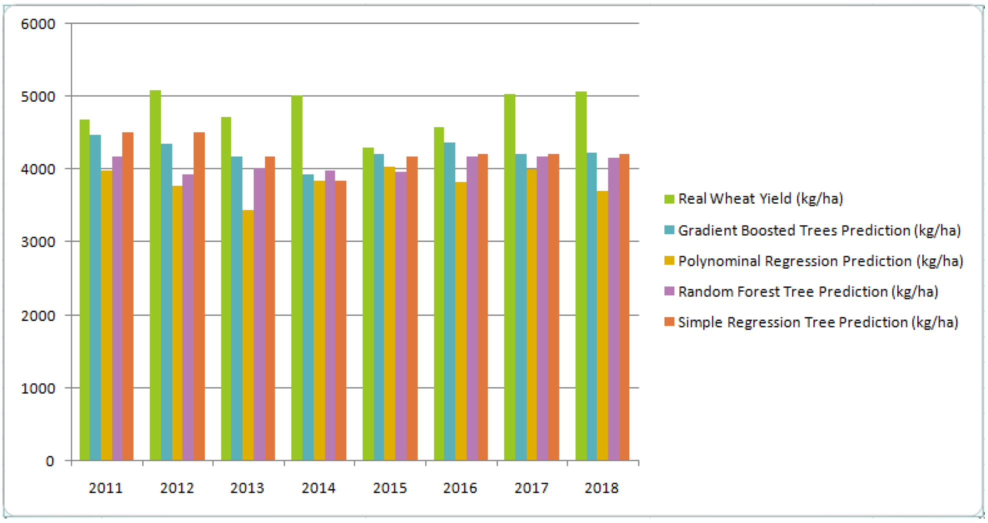

**Fig. 3.** Bar Graphical Representation of Predictions with Actual Results.

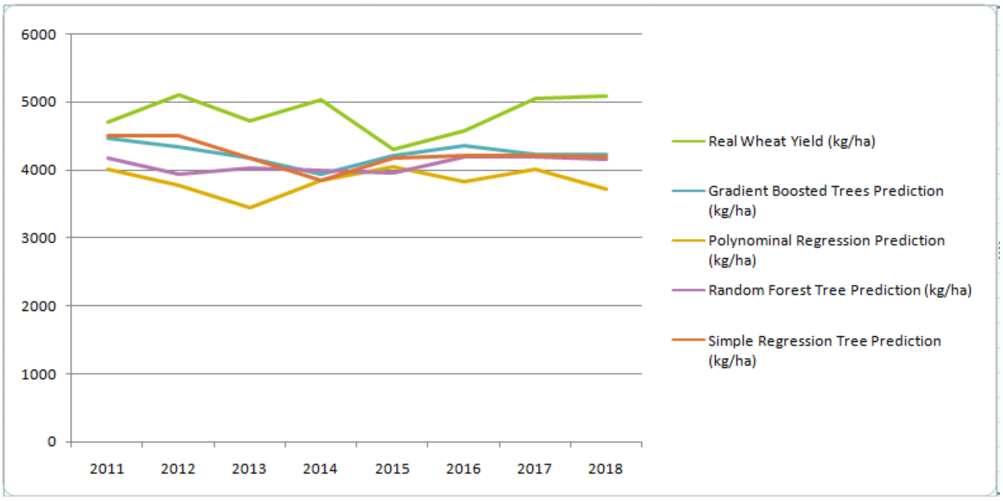

**Fig. 4.** Line Graphical Representation of Predictions with Actual Results.

3. The table of performance evolution measures such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and Mean Absolute Percentage Error of applied algorithm is given below in table 3:

**Table 3.** Table of Performance Evolution Measures

| Type of Errors | Random Forest Trees | Gradient Booster | Simple Regression | Polynomial Regression |
|---|---|---|---|---|
| **Mean Absolute Error** | 709.744 | 570.942 | 583.375 | 986.935 |
| **Mean Squared Error** | 597,836.813 | 440,162.565 | 452,351.375 | 1,102,940.261 |
| **Root Mean Squared Error** | 773.199 | 663.447 | 672.571 | 1050.21 |
| **Mean Absolute Percentage Error** | 0.144 | 0.115 | 0.118 | 0.202 |

4. Accuracy of applied models is given below in table 4:

**Table 4.** Accuracy of Applied Models

| **Random Forest Trees** | **Gradient Booster** | **Simple Regression** | **Polynomial Regression** |
|---|---|---|---|
| **85.6%** | 88.5% | 88.2% | 79.8% |

## 4. Conclusion and Future work

From the results it is clearly shown that Gradient booster gives the maximum accurate results. The results are obtained currently using the Knime software but our future work is to develop an application so that the farmers can operate it easily.

## 5. References

[1] Babu, T. Giri, and Dr G. Anjan Babu. "Big Data Analytics to Produce Big Results in the Agricultural Sector." (2016).

[2] Djodiltachoumy, S. "A Model for Prediction of Crop Yield." International Journal of Computational Intelligence and Informatics 6, no. 4 (2017).

[3] Ghadge, Rushika, Juilee Kulkarni, Pooja More, Sachee Nene, and R. L. Priya. "Prediction of crop yield using machine learning." Int. Res. J. Eng. Technol.(IRJET) 5 (2018).

[4] Huang, Jui-Chan, Kuo-Min Ko, Ming-Hung Shu, and Bi-Min Hsu. "Application and comparison of several machine learning algorithms and their integration models in regression problems." Neural Computing and Applications 32, no. 10 (2020): 5461-5469.

[5] Jain, Nishit, Amit Kumar, Sahil Garud, Vishal Pradhan, and Prajakta Kulkarni. "Crop selection method based on various environmental factors using machine learning." International Research Journal of Engineering and Technology (IRJET) 4, no. 2 (2017): 1530-1533.

[6] Kale, Shivani S., and Preeti S. Patil. "A Machine Learning Approach to Predict Crop Yield and Success Rate." In 2019 IEEE Pune Section International Conference (PuneCon), pp. 1-5. IEEE, 2019.

[7] Khairunniza-Bejo, Siti, Samihah Mustaffha, and Wan Ishak Wan Ismail. "Application of artificial neural network in predicting crop yield: A review." Journal of Food Science and Engineering 4, no. 1 (2014): 1.

[8] Kumar, Rakesh, M. P. Singh, Prabhat Kumar, and J. P. Singh. "Crop Selection Method to maximize crop yield rate using machine learning technique." In 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM), pp. 138-145. IEEE, 2015.

[9] Kushwaha, Ashwani Kumar, and Sweta Bhattachrya. "Crop yield prediction using Agro Algorithm in Hadoop." International Journal of Computer Science and Information Technology & Security (IJCSITS) 5, no. 2 (2015): 271-274.

[10] Medar, Ramesh, Vijay S. Rajpurohit, and Shweta Shweta. "Crop yield prediction using machine learning techniques." In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1-5. IEEE, 2019.

[11] Mishra, Subhadra, Debahuti Mishra, and Gour Hari Santra. "Applications of machine learning techniques in agricultural crop production: a review paper." Indian Journal of Science and Technology 9, no. 38 (2016): 1-14.

[12] Nigam, Aruvansh, Saksham Garg, Archit Agrawal, and Parul Agrawal. "Crop yield prediction using machine learning algorithms." In 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 125-130. IEEE, 2019.

[13] Ramesh, D., and B. Vishnu Vardhan. "Analysis of crop yield prediction using data mining techniques." International Journal of research in engineering and technology 4, no. 1 (2015): 47-

473.

[14] Raorane, A. A., and R. V. Kulkarni. "Application of DataMining tool to crop management system." Russian Journal of Agricultural and Socio-Economic Sciences 37, no. 1 (2015).

[15] Rajak, Rohit Kumar, AnkitPawar, MitaleePendke, PoojaShinde, Suresh Rathod, and AvinashDevare. "Crop recommendation system to maximize crop yield using machine learning technique." *International Research Journal of Engineering and Technology* 4, no. 12 (2017): 950-953.

[16] Savla, Anshal, Himtanaya Bhadada, Parul Dhawan, and Vatsa Joshi. "Application of machine learning techniques for yield prediction on delineated zones in precision agriculture." IJNCAA (2015): 48

[17] Son, Nguyen-Thanh, Chi-Farn Chen, Cheng-Ru Chen, Horng-Yuh Guo, Youg-Sing Cheng, Shu-Ling Chen, Huan-Sheng Lin, and Shih-Hsiang Chen. "Machine learning approaches for rice crop yield predictions using time-series satellite data in Taiwan." International Journal of Remote Sensing 41, no. 20 (2020): 7868-7888.

[18] D. Sindhura, B. Navya Krishna, K. Sai Prasanna Lakshmi, B. Mallikarjun Rao, Dr. J Rajendra Prasad, Effects of Climate Changes on Agriculture International Journal of Advanced Research in Computer Science and Software Engineering,2016.

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