International Journal of Modern Agriculture, Volume 9, No.3, 2020
ISSN: 2305-7246

Crop Yield Prediction based on Indian Agriculture using Machine Learning

N BanuPriya[1], D Tejasvi [2], P Vaishnavi [3]

[1]Computer Science and Engineering, R.M.K. Engineering College
[2,3] Student, Computer Science and Engineering, R.M.K. Engineering College.

Abstract

In our analysis, that we tend to have discovered withinside the previous studies papers is that everybody makes use of environmental condition factors like rain, daylight and agricultural factors like soil sort, nutrient possessed via the soil (Nitrogen, Potassium, etc.) however the matter is we want to gather the data so a third party will do this prediction and later it is explained to the farmer and this takes a variety of attempts for the farmer and he doesn't perceive the technological study behind these factors. To make it straightforward and which can be directly utilized by the farmer this paper uses easy factors like the state and district is the farmer from, the crop and in what season (as in Kharif, Rabi, etc.). In India, there are more than a hundred vegetation plants across the entire country. These vegetation are categorized for better understanding and image. The data for this analysis has been nonheritable from the Indian Government Repository [1]. The statistics includes attributes like– State, District, Crop, Season, Year, Space and Production with around 2.5 Lakh observations. We used advanced regression techniques – Random Forest, Gradient Boost and Decision Tree to predict the yield and used Ensemble algorithms to minimize the error and reap higher predictions.

Keywords: Python, Ensemble Algorithms, Machine Learning.

I. Introduction

Agriculture is regarded as India's primary and most significant community. Since ancient people grew crops on their own land, they were able to adapt to their needs. As a result, natural crops are grown and used by a number of species, including humans, livestock, and birds. The creature has taken the greenish goods created in the soil, which has resulted in a healthy and welfare life. The agricultural sector has been steadily degrading since the advent of new advanced technologies and techniques. People are concentrating on cultivating artificial goods, which are hybrid products, because of these plentiful inventions, which contributes to an unhealthy existence. Nowadays, most people are unaware of the importance of cultivating crops at the appropriate time and place. Seasonal climatic conditions are also being altered because of these cultivation methods, posing a challenge to fundamental assets such as soil, water, and air, resulting in food insecurity. By evaluating all these issues and challenges, such as weather, temperature, and several other variables, we have discovered that there is no suitable solution or technology to help us solve the situation we are in. In India, there are many options for increasing agricultural economic development. There are numerous methods for increasing and improving crop yield and quality. Data mining can also be used to forecast crop yield production. Data mining, in general, is the methodology of analyzing information from different angles and synthesizing it into helpful information.

Data mining software is an analytical tool that allows users to look at data from a variety of perspectives, categorize, and summarize the relationships found. Data mining is the method of detecting trends or associations between hundreds of fields in broad relational databases. All of this data's patterns, associations, and relationships can provide information. Knowledge about historical patterns and future trends can be derived from data. Summary information about crop production, for example, may assist farmers in identifying crop losses and preventing them in the future. Crop yield forecasting is a significant agricultural problem. Every farmer is always curious as to how much yield he can expect based on his expectations. Yield prediction used to be determined by considering a farmer's previous experience with a specific crop. Weather conditions, livestock, and harvest process preparation all play a role in agricultural yield. Accurate knowledge about crop yield history is crucial when making decisions about agricultural risk management. Therefore, this paper proposes a method for estimating the cop's yield. Before planting the field, the farmer will check the yield of the crop per acre.

II. LITERATURE SURVEY

[2] M. G. Ananthara et al. suggested the CRY algorithm for crop yield using beehive clustering techniques as a prediction model for agricultural datasets. Agricultural research around the world emphasizes the need for a reliable system to forecast and boost crop growth. There is a clear need for an prophetic modelling system with efficient predictive yield management methodology die to multi- dimensional variable metrics and also the lack of a prophetic modelling approach, which leads to crop yield loss. Using a dynamically updated historical crop information set, this paper proposes a crop yield prediction model (CRY) that employs an adaptational cluster approach to predict crop yield and enhance exactness agriculture decision- making employing a dynamically changed historical crop knowledge assortment. CRY analyses and classifies crops using a beehive simulation approach based on crop growth trend and yield. Clementine was used to measure the CRY classified dataset against established crop domain information.

[3] Awan, A. M. et al. suggested creating a software program framework for prognostication crop yield primarily based totally on weather and plantation information. The core of this scheme is an approach for unattended data partitioning for locating spatio-temporal trends in climate knowledge using kernel strategies, which provides strength in dealing with complicated data. For this purpose, a strong weighted kernel k-means formula with spatial constraints is given. The algorithm can efficiently manage noise, outliers, and auto-correlation in spatial data for successful and efficient data analysis, and so are often accustomed to predicting oil-palm yield with the aid of using different studying factors affecting yield.

[4] Chawla, I. et al. Fuzzy logic-based Crop Yield prediction using Temperature and Rainfall parameters predicting through ARMA, SARIMA and ARMAX models: Agriculture is extremely important in India's economy. As a result, crop yield forecasting is an essential task for India's development. Crops are affected by a variety of weather conditions, including temperature and rainfall. As a result, when forecasting a crop's yield, it's important to take these factors into account. Forecasting the weather is a difficult task. Three forecasting methods are used in this study: ARMA, SARIMA, and ARMAX (Auto Regressive Integrated Moving Average) (ARMA with exogenous variables). The three models' output is compared, and the best model is used to forecast rainfall and temperature, which are then used to forecast crop yield using a fuzzy logic model.

[5] Crop Yield prediction using data analytics and hybrid approach: Agricultural data is constantly and massively generated. As a result, the age of bid data has arrived for agricultural data. Data collection via electronic devices is aided by smart technologies. We will analyze and mine this agricultural data in our project to obtain useful results using technologies such as data analytics and machine learning, and this information will be provided to farmers in order to improve crop yield in terms of production and productivity.

[6] Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. proposed work on Data mining techniques for crop yield prediction: Predicting crop yields well ahead of harvest will help farmers and government agencies make more informed decisions about storage, sale, setting minimal assist rates, importing/exporting, and other activities. A detailed analysis of vast quantities of data derived from varied variables including soil quality, pH, EC, N, P, K, and so on. Since crop prediction requires many databases, this prediction method is an ideal candidate for data mining. We derive information from massive amounts of data using data mining. This research discusses the various data mining techniques that have been used to predict yield of the crop. The precision with which features are extracted and how well they are named defines the efficiency of any crop yield prediction system.

[7] Gandhi, N., Petkar, O., & Armstrong, L. J. Artificial Neural Network Rice crop yield prediction. The intention of this study was to use neural networks to predict rice manufacturing yield and to research about the factors that impact rice crop yield in Maharashtra, India. Data for 27 districts in Maharashtra state, India, were collected from publicly available Indian government records. For the Kharif season from 1998 to 2002, the parameters considered were common temperature, highest temperature, reference crop evapotranspiration, location, production, and yield. The data was analyzed using the WEKA program. It was created a Multilayer Perceptron Neural Network. The data was validated using the cross-validation process. The precision was 97.5 percent, with a sensitivity of 96.3 and specificity of 98.1 percent. Additionally, for this analysis, mean absolute

error, root mean squared error, relative absolute error, and root relative squared error have been measured. The WEKA tool's Information Flow was also used to run the study dataset. The WEKA tool's Information Flow was also used to run the study dataset. The ROC curve is used to visually summarize the classifier's results.

III. PROPOSED SOLUTION

To design an application where we compare the different machine learning to predict the crop yield. We build a new decision system using ensemble regression system. The user would provide input of season type, year of production, area of production, crop type, cloudburst, climate condition, located yield within side the remaining and the system would predict the yield and relying at the value set, the crop may be classified and attain the results. In the first step it allows the admin to login and load the data. Second, it allows the admin to perform analysis by considering all the input conditions. Finally, a report is generated for the crop yield and the accuracy of the models are also generated. The accuracy which is near to 1 is considered as an ideal model and the model which has accuracy near to 0 are considered as unideal model. The input for the system will be a season, rainfall, area of production, crop type, district name, state name and output will be the production of crop yield and accuracy of each model.

This project provides:

- Allows Admin to login.
- Allows admin to load data.
- Allows admin to predict the death cause.
- Allows admin to perform analysis.
- Allows admin to generate reports.

This is achieved through these concepts. (i)Random Forest regressor. (ii)Gradient Boost regressor. (iii) Decision Tree regressor.

System architecture consists of 6 modules namely,

(i) Requirement gathering

(ii) Analysis

(iii) Design

(iv) Coding

(v) Testing

(vi) Maintenance

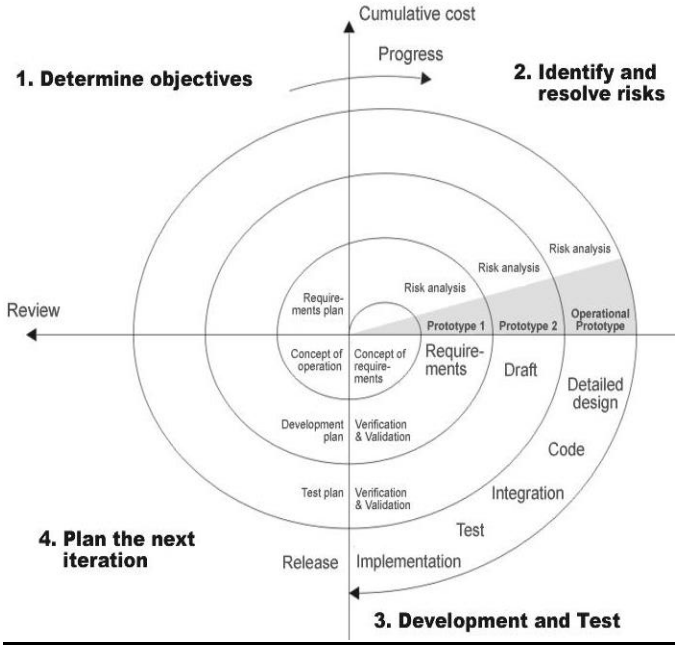

Fig. 1: Software Development Lifecycle Methodology

Requirement gathering stage: This stage takes as its input the objectives defined within the high-level portion of the project set up. Each target will be broken down into one or extra specifications. These specifications describe the intended application's key functions, operational information areas, and reference in areas, as well as the first data entities. The key roles include managing sensitive processes as well as mission crucial inputs, output, and reports. These core functions, information regions, and data entities are organized according to a user class hierarchy. Each of these definitions is referred to as a Prerequisite. Requirements are diagnosed via means of precise requirement identifies and, at minimum, include a requirement identity and textual description. In this stage, all the necessities are well specified withinside the primary deliverable: The Requirements Document and the Requirements Traceability Matrix (RTM). The specifications document provide comprehensive explanations of and condition, including diagrams and references to outside documents, as required. The title of each condition, as well as the title of each target from the project plan, is included in the initial version of the RTM. The RTM's aim is to demonstrate that the merchandise elements developed throughout every stage of the computer code development lifecycle area unit formally related to the components developed in previous stages. The RTM in the requirement stage consists of a list of high- level necessities, or targets, organized by title, along with a list of related requirements for each goal, organized by requirement title. The RTM demonstrates that every criterion identified throughout this stage is joined formally to a selected product goal during this hierarchical listing. The output of the requirement stage includes the necessities document, the RTM, and an updated project set up.

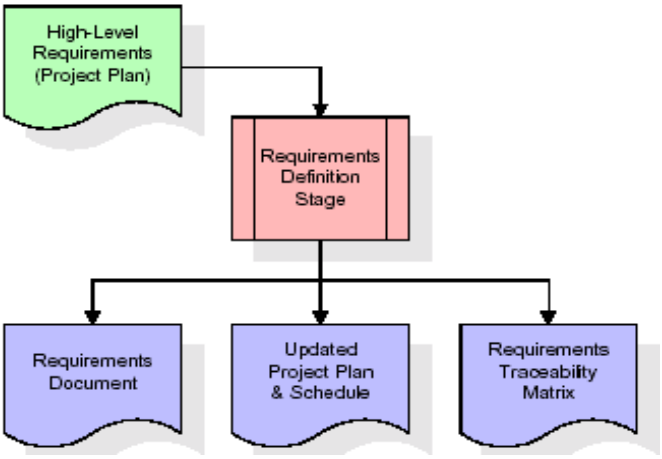

Fig. 2: Requirement gathering stage

Analysis stage: This planning stage establishes a bird's eye view of the intended software product and makes use of this to set up the fundamental challenge structure, examine feasibility and dangers related to the project, and describe suitable management and technical approaches. The maximum critical phase of the project pan is a list of high- degree product necessities, additionally known as goals. During the requirements specification stage, all the software program product necessities so as to be created a glide from one or more of these objectives. The minimal records and references to outside files can be included. The configuration control plan, the fine guarantee plan, and the project plan and schedule are the outputs of the project planning stage, with a complete list of deliberate responsibilities for the approaching necessities degree and high- degree estimates of attempt for the out stages.

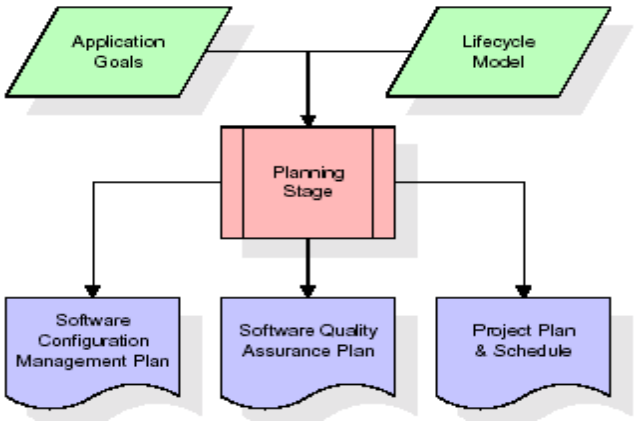

Fig. 3: Analysis stage

Design stage: It takes as its preliminary input the necessities recognized withinside the accredited necessities report. For every requirement, a set of one or more layout factors can be produced due to interviews, workshops, and/ or prototype efforts. Functional hierarchy diagrams, display format diagrams, tables of enterprise rules, enterprise process diagrams, pseudo code, and a entire entity-dating diagram with a complete records dictionary are all examples of layout factors that designate the preferred software program functions in detail. These design elements are intended to provide enough information about the software so that professional programmers can create it with minimal help. The RTM is revised after the layout report is completed and accredited to signify that every layout characteristic is formally aligned with a precise requirement. The outputs of the design degree are the layout report, an up-to-date RTM, and an up to date venture plan.

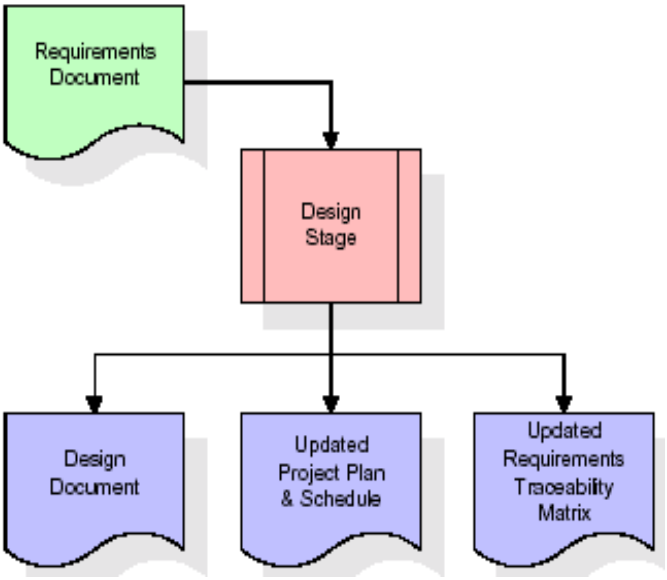

Fig. 4: Design stage

Coding Stage: The Software package artifacts, on-line facilitate and test records migrated from the event surroundings to a separate test environment. All test cases are run at this point to confirm that the program is true and complete. The test suite's prosperous completion demonstrates a stable and full migration capability. Production users' square measures are outlined and connected to their acceptable roles throughout this time, and reference information is finalized for production use. The Production Initiation Plan contains the ultimate reference knowledge and production user list. An associate integrated assortment of tools, an online support system, an implementation map, a development plan that identifies reference knowledge and production users, an approval arrangement that has the ultimate suite of test cases, and an updated project plan are all products of this level.

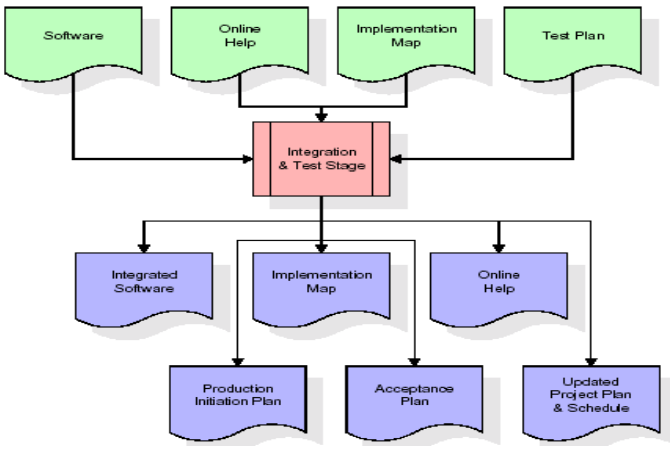

Fig. 5: Coding stage

Testing stage: The software program artifacts, on-line help, and preliminary production statistics are loaded onto the production server. All test cases are run at this point to ensure that the program is right and complete. The test suite must be completed successfully before the program can be accepted by the customer. The customer formally approves the delivery of this system after customer workers have checked that the preliminary production statistics load is correct and that the test suite has been achieved with perfect results. A production application, a accomplished acceptance test suite, and a memorandum of customer acceptance of the system are the primary outputs of this level. Finally, the PDR enters the very last piece of real labor statistics into the project agenda and saves it as a everlasting undertaking record. The PDR "locks" the project at this

factor via means of archiving all application objects, the implementation map, the supply code, and the documentation for future use.

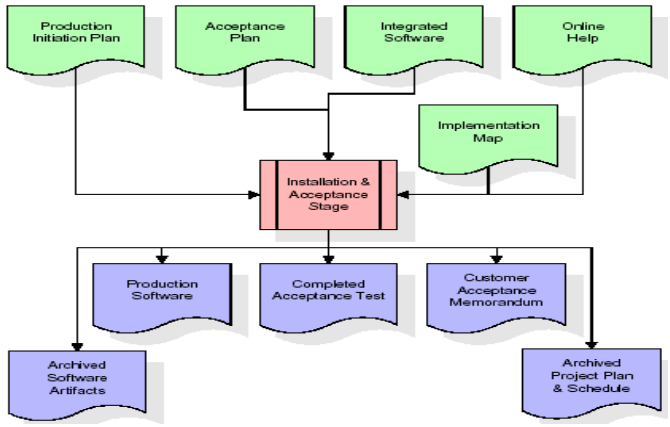

Fig. 6: Testing stage

Maintenance:

The outer rectangle represents project maintenance; the maintenance team will begin by studying specifications and understanding documentation; after that, employees will be allotted work and receive coaching within the class to which they have been allotted.

Training our model:

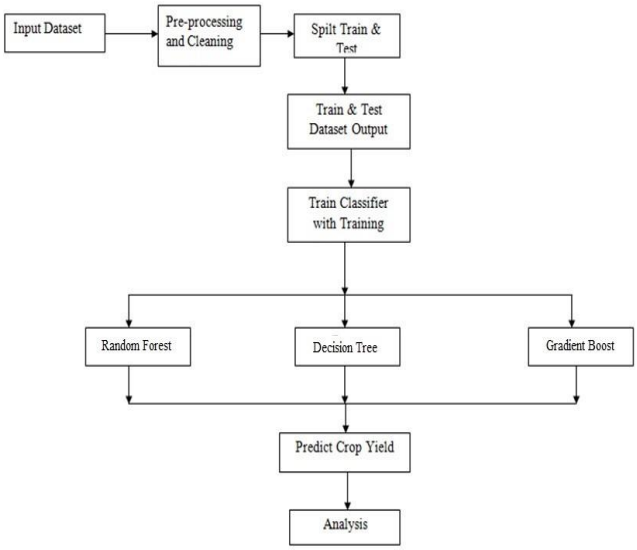

Fig. 7: Architecture Diagram

Methodology

To train our model we need data.

The collected data can have little 'NA' values filtered out in Python. Since the data set is composed of digital data, the robust scaling we use is very similar to normalization, but it uses interquartile range, and normalization compresses the data in units of 0 and 1.

Ensemble Algorithms

- On average, this is an improvement. Here, we add a meta model and use creases in addition to creases. The predictions of other models is used to train the basic meta-model.
- The entire training set is divided into two different sets (train and test/holdout datasets).
- We train the chosen base models with initial part (train dataset).
- Then we check them with the second half (holdout set).
- Now, the predictions obtained from the check half are inputs to the train higher level learner referred to as meta-model.

Algorithms:

- Random Forest Regression: It generates multi decision trees from which each decision tress uses a part of data sample and predicts the result. Then, the result which was achieved by maximum number of trees is considered as the final prediction. This is a supervised learning algorithm which uses ensemble learning method for classification and regression. It is a bagging technique and the trees in random forests run in parallel without any interactions.
- Decision Tree Regression: Trees are constructed through an algorithmic approach that identifies ways to split the data set based on different conditions. It is one of the most widely used practical methods for supervised learning. These are non-parametric method used for both classification and regression.
- Gradient Boost Regression: This method converts the weak learners into strong learners by boosting their capability. It is a sequential process of learning from the previous trees and increases the model accuracy.

Dataset results:

The scene in this dataset is straightforward. It contains the accuracy of the three algorithms. The R^2 score is used to predict the accuracy of the algorithms. The model closer to 1 is an ideal model whereas the model closer to 0 is an unideal model. The accuracy of Random Forest Regression is 0.88, Gradient Boost Regression is 0.84 and Decision Tree Regression is 0.71.

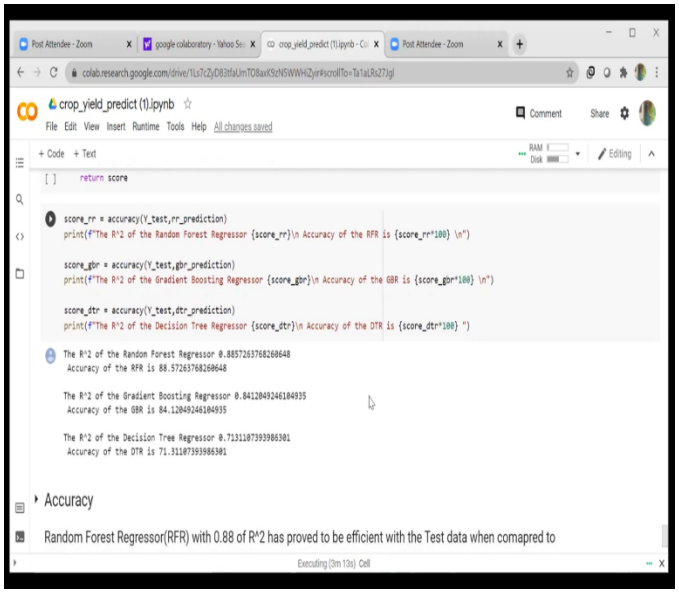

Fig. 8: Accuracy of the algorithms

IV. CONCLUSION:

The outcomes have extremely improvised after applying Ensemble algorithms. The Random Forest, Decision Tree and Gradient Boost Regressors were successfully implemented.

Random Forest Regressor (RFR) with 0.88 of R^2 score has proved to be efficient with the Test data when compared to DTR and GBR.

V. FUTURE SCOPE:

The production shown in the figure is presently an online application, however our future work will include developing an application that farmers can use and translating the entire system into their native language.

REFERENCES

1. "data.gov.in." [Online]. Available: https://data.gov.in/
2. Ananthara, M. G., Arunkumar, T., & Hemayathy, R. (2013, February). CRY- an improved crop yield prediction model using beehive clustering approach for agricultural data sets. In 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering.
3. Awan, A. M., & Sap, M. N. M. (2006, April). An intelligent system based on kernel methods for crop yield prediction. In Pacific- Asia Conference on Knowledge Discovery and Data Mining (pp. 841-846). Springer, Berlin, Heidelberg.
4. Bang, S., Bishnoi, R., Chauhan, A. S., Dixit, A. K., & Chawla, I. (2019, August). Fuzzy logic-based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA and ARMAX models. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.
5. Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. (2018, August). Crop Yield Prediction using data analytics and hybrid approach. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

6. Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) (pp. 420-423). IEEE

7. Gandhi, N., Petkar, O., & Armstrong, L. J. (2016, July). Rice crop yield prediction using artificial neural networks. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 105-110). IEEE.

8. Islam, T., Chisty, T. A., & Chakrabarty, A. (2018, December). A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh. In 2018 IEEE Region 10 Humanitarian Technology Conference (R10- HTC) (pp. 1-6). IEEE.

9. Jaikla, R., Auephanwiriyakul, S., & Jintrawet, A. (2008, May). Rice yield prediction using a support vector regression method. In 2008 5th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications, and Information Technology (Vol.1, pp. 29-32). IEEE.

10. Kadir, M. K. A., Ayob, M. Z., & Miniappam, N. (2014, August). Wheat yield prediction: Artificial Neural Network based approach. In 2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T) (pp. 161-165). IEEE.

11. Manjula, A., & Narisimha, G. (2015, January). XCYPF: A flexible and extensible framework for agricultural Crop Yield Prediction. In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-5). IEEE.

12. Mariappan, A. K., & Das, J. A. B. (2017, April). A paradigm for rice yield prediction in Tamilnadu. In 2017 IEEE Technological Innovations in ICT for Agricultural and Rural Development (TIAR) (pp. 18-21). IEEE.

13. Paul, M., Vishwakarma, S. K., & Verma, A. (2015, December). Analysis of soil behavior and prediction of crop yield data mining approach. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 766-771). IEEE.

14. Shah, A., Dubey, A., Hemnani, V., Gala, D., & Kalbande, D. R. (2018). Smart Farming System: Crop Yield Prediction Using Regression Techniques. In Proceedings of International Conference on Wireless Communication (pp. 49-56). Springer, Singapore.

15. Ahamed, A. M. S., Mahmood, N. T., Hossain, N., Kabir, M. T., Das, K., Rahman, F., & Rahman, R. M. (2015, June). Applying data mining Techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing (SNPD) (pp. 1-6). IEEE.

16. Shastry, A., Sanjay, H. A., & Hegde, M. (2015, June). A parameter based ANFIS model for crop yield prediction. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 253-257). IEEE.

17. Sujatha, R., & Isakki, P. (2016, January). A study on crop yield forecasting using classification techniques. In 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16) (pp. 1-4). IEEE.

18. Suresh, A., Kumar, P. G., & Ramalatha, M. (2018, October). Prediction of major crop yields of Tamilnadu using K-means and Modified KNN. In 2018 3rd International Conference on Communication and Electronics Systems (ICCES) (pp. 88-93). IEEE.

19. Veenadhari, S., Misra, B., & Singh, C. C. (2014, January). Machine learning approach for forecasting crop yield based on climatic parameters. In 2014 International Conference on Computer Communication and Informatics (pp. 1-5). IEEE.

20. Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016, July). Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE.

21. Gandhi, N., Armstrong, L. J., & Petkar, O. (2016, July). Proposed decision support system (DSS) for Indian rice crop yield prediction. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 13-18). IEEE.

22. Manjula, A., & Narisimha, G. (2015, January). XCYPF: A flexible and extensible framework for agricultural Crop Yield Prediction. In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-5). IEEE.

23. Paul, M., Vishwakarma, S. K., & Verma, A. (2015, December). Analysis of soil behavior and prediction of crop yield data mining approach. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 766-771). IEEE.

24. Shastry, A., Sanjay, H. A., & Hegde, M. (2015, June). A parameter based ANFIS model for crop yield prediction. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 253-257). IEEE.

25. Mariappan, A. K., & Das, J. A. B. (2017, April). A paradigm for rice yield prediction in Tamilnadu. In 2017 IEEE Technological Innovations in ICT for Agricultural and Rural Development (TIAR) (pp. 18-21). IEEE.

26. Aruvanash, N., & Saksham Garg, (2019, Nov). Crop Yield Prediction using Machine Learning Algorithm. In 2019 Fifth International Conference on Image Information Processing.

27. Kavitha, M., & Pratistha . M. Crop Yield Estimation in India Using Machine Learning. In 2020 IEEE 5th International Conference on Computing Communication and Automation. (ICCCA).