Predictive Agriculture Using Data Analysis and Machine Learning

Aastha Doshi
Dept. Computer Science and Technology
Usha Mittal Institute of Technology, SNDT University Santacruz, Mumbai, India aasthadoshi31@gmail.com

Anuradha Chopade
Dept. Computer Science and Technology
Usha Mittal Institute of Technology, SNDT University Santacruz, Mumbai, India anuradhachopade99@gmail.com

Sonali Bodekar
Dept. Computer Science and Technology
Usha Mittal Institute of Technology, SNDT University Santacruz Mumbai, India sonali.bk86@gmail.com

***Abstract*—Agriculture is India's prime occupation. The majority of people in the rural and urban areas depend upon this sector directly or indirectly. Agriculture and its components contribute to approximately 20% of India's GDP in 2020-21. Increasing population, increasing average income, and globalization effects in India will increase demand for quantity, quality and nutritious variety of food. There is immense pressure on farmers to keep up with this demand and supply chain. Even though the technology is upgrading enormously, the rural areas are still using traditional methods of agriculture. Therefore, to accelerate the production process our paper focuses on the prediction of various aspects of agriculture with the help of *Machine Learning* algorithms and *Data Analytics*. In this study, we have considered four aspects i.e. Rainfall, Crop production, soil and weather undergoing an ensemble learning algorithm for predictions. This will lead to better yield with minimum usage of resources.**

***Index Terms*—Machine learning, big data analysis, predictions, ensemble learning algorithm, Random Forest Regressor.**

I. INTRODUCTION

Agriculture is the primary source of livelihood for about 59% of India's population. Gross Value Added (GVA) by agriculture and supporting industries was estimated at 19.48 trillion in FY20. According to Indian Council for Agricultural Research (ICAR), the total food grain production in the country is estimated to be a record 291.95 million tonnes and the demand for food grains would increase to 345 million tonnes by 2030. Although these estimations are fascinating, this industry faces many unpredictable and drastic problems like drought, lack of rain, weather changes and floods. The other problems are overuse of chemical fertilizers which leads to degradation of soil nutrients, not enough subsidies by the governing body and corruption does not let the subsidies reach to the farmers lead to an increase in the debt which ultimately leads to suicides. All the above reasons make it crucial to use statistical methods in this field to provide us with better yield with minimum use of capital and other accessories. Using the advances in computer science and data science we can predict to some extent the crop production weather prediction, rainfall pattern. These predictions can be made with the Machine learning branch of computer science. Machine learning has become the most fascinating stream of computer science. Machine Learning and artificial intelligence are synonyms used from time to time. Machine algorithms analyzed any typical situation based on multiple combinations of conditions fed into it. It can predict the future related to crop yield, nutrition value, etc. up to a significant extent. The prediction accuracy depends upon the data given to the algorithm. The statistical nature of these algorithms can lead to a significant increase in yield. In India, the concept of applying artificial intelligence (AI) and machine learning (ML) is not fully grasped and the traditional methods of farming sometimes can put farmers at great risk, instead of using technology can be at great practice.

II. LITERATURE REVIEW

In this paper [1], they have presented diverse procedures for crop prediction for the states of Uttar Pradesh and Karnataka. These use models like Naive Bayes, Random Forest, KNN, etc. the performance is evaluated on data using techniques such as cross-validation, accuracy, RMSE, precision, recall. The best performing model is selected and subsequently used to classify and recommend crops. Wheat crop production [2], studies the use of machine learning on wheat crop production. The methodology consists of taking digital image processing techniques for feature extraction for the maturity of crop and determining the stages of growth by classification using supervised machine learning technique Support Vector Machines (SVM).Data mining techniques [3] like K-Means Clustering, KNN, SVM, and Bayesian network algorithms where high accuracy was achieved to predict the crop production.

In [4], selection of crops, they have included two different Methods-Firstly Naive Bayes method and secondly K Nearest neighbour method. To predict the crop yield rate a java application was created. This application includes three parts. The first is managing datasets, second is testing datasets and third is analyzing the datasets. This technique gave 91.11% accuracy. In this paper [5], high precision agriculture is achieved with the performance of machine learning algorithms on aerial image object detection. This methodology uses the Decision Tree Ensemble with an accuracy of 94%.The research focuses on the prediction of different crops yield using neural network regression modelling [6]. This research paper describes the development of a different crop yield prediction model with ANN, with a 3 Layer Neural Network. In this paper [7], Crop yield prediction incorporates forecasting the yield of the crop from past historical data. These predictions are done by a machine learning algorithm called Random Forest with the best accurate value. In [8], the author has proposed a model that can predict soil series with land type and according to prediction it

can suggest suitable crops. Several machine learning algorithms such as weighted k-Nearest Neighbor (KNN), Bagged Trees, and SVM are used for soil classification. The results show that the proposed SVM-based method performs better than many existing methods. A review [9] on different machine learning methods such as SVM, Naive Bayes, Random forest, ANFIS, and Gradient Boosting Methods(GBM) used in predicting the soil type, soil moisture and soil nutrient content are presented. SVM and Multi-layer Perceptron performed better by achieved the highest accuracy. This paper [10], has performed the heuristic prediction of rainfall using machine learning techniques. The paper also measures the different categories of data by linear regression method in metrics for effective understanding of agriculture in India.

In This paper [11], the author focuses on optimizing the effect of weather on agriculture using various techniques like Correlation Analysis, multidimensional modeling, k- means, ANN, SVM, KG-classification, PAM, CLARA, DBSCAN, etc. They have also given a comprehensive review of agricultural vulnerabilities and research done to analyze the impact of climate change in climate on agriculture.

In This paper [12], Real-time farm monitoring, Cloud data analytics, and a mobile application is made which perform predictive analytics on sensed data, soil type, landscape, climate, day-to-day market price, and farmer's Economy.

In this study [13], the authors have collected data from different government organizations, after Pre-processing of data applied the Predictive Apriori algorithm using the Data Mining tool - WEKA for analyzing of daily temperature, daily rainfall and paddy yield to predict the yield and to analyze the effect of temperature and rainfall on the paddy yield.

The table below shows the distinguishing study of the literature review.

Sr no.	Reference	Methodology	Inference
1	[1]	Naive Bayes, Ran-dom Forest, KNN	Random forest gave Highest accuracy 79%
2	[2]	Support Vector Ma-chine	To determine the stages of growth by classification using SVM
3	[3]	K-Means Clustering, KNN, SVM, and Bayesian network algorithm	To predict the crop production.
4	[4]	Naive Bayes, K Nearest neighbour	To predict the crop Yield with 91.11% accuracy.
5	[5]	The Decision Tree Ensemble	High precision agriculture is achieved with an accuracy of 94%.
6	[7]	Random Forest	Crop yield prediction incorporates forecasting the yield of the crop from past historical data.
7	[8]	k-Nearest Neighbor (KNN), Bagged Trees, SVM	Soil classification using SVM-based method performs better.
8	[10]	linear regression	For understanding Rainfall Pattern in India
9	[13]	Predictive Apriori algorithm	to analyze the effect of temperature and rainfall on the paddy yield

III. METHODOLOGY

After studying the multiple research papers, it's been inferred that we have to test the different machine learning algorithms on data and as it gives different results depending on that data fed to the algorithm. In this section we display our approach with a flowchart Fig.1 below. We are implementing our work into five modules. Each module consists of examining the factors that affect agriculture directly. The final module will be deployed using the Streamlit library in Python.

Module 1: This consists of the production data for agriculture from 2001-2014 for different crops and regions through- out India. We have divided this production data state-wise and then applied a Regression algorithm to predict production for each state. We are only considering those states for final implementation who have R2 score above 50%.

Module 2: Consists of rainfall data from year 1980 - 2017 for different subdivisions in India.

Module 3: This consists of soil data; what kind of soil it is, fertility of land etc.

Module 4: This module has weather data such as daily temperature, humidity, wind, description and precipitation.

Module 5: The final implementation is integrating all four modules together and deploying the end product.

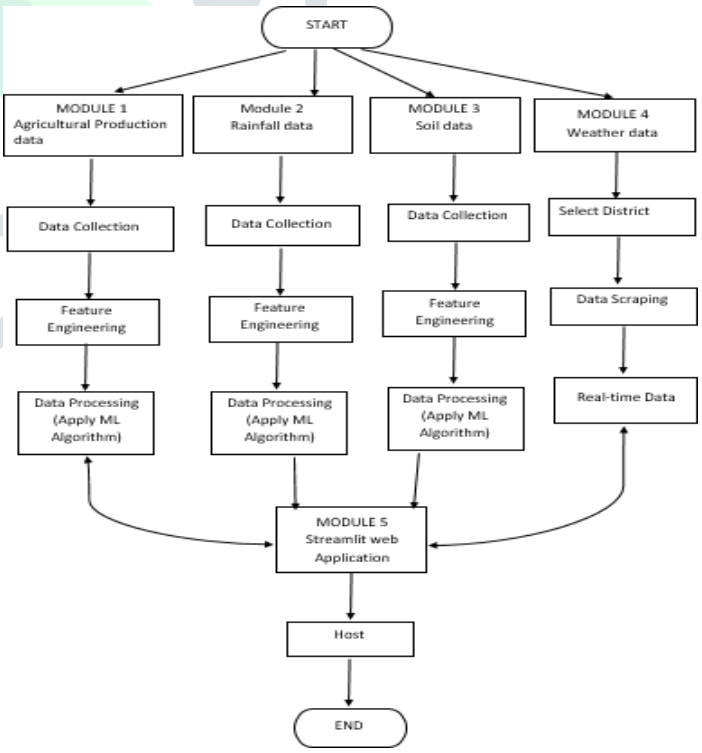

Fig. 1. Methodology Flowchart

The first step is collecting data or data scraping. Data col- lection will be carried out through web scraping from few of the below

mentioned websites and datasets available on the website [14]. Data undergoes a cleaning step where missing data is imputed and outliers are handled. Then we perform Exploratory Data Analysis where we understand patterns in data. It is where we start analyzing trends and begin to find some insights. The next step is building models, we implement different Machine Learning algorithms to see which performs the best to produce more accurate results. The two main algorithms we apply are Random Forest regressor (crop production and rainfall prediction) and Random Forest Classifier (soil data).

V IMPLEMENTATION AND RESULT

Implementation is executed with Python programming language. The programming is done using a Jupyter notebook. Jupyter is a free, open-source, interactive web tool known as a computational notebook. The deployment is done using the Streamlit web application.

A. Datasets

The crop production datasets are gathered from the web-site [14]. It has data of crop production from 2001-2014 .The rainfall datasets are gathered from [14] and later merged to single dataset for whole India, it is from 1980-2017. The weather data is extracted using web scraping from [15], it is real-time weather data. The soil data is gathered from [16] which was in topographic map form and later we converted it into a tabular dataset.

B. Random forest regression

A Random Forest is a sophisticated ensemble learning based algorithm efficient for performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. Bagging involves training each decision tree on a different data sample where sampling is done with replacement. The idea behind this is to merge multiple decision trees in determining the final output rather than relying on individual decision trees.
To perform prediction using the trained Random Forest algorithm uses the below steps:

1) It takes the test features and uses the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome.
2) Calculate the votes for each predicted outcome.
3) Consider the highly voted predicted outcome as the final prediction from the random forest algorithm.

After the EDA process Random forest was applied on crop production dataset which has 242,361 rows and consists of the production of every state in India. The model is trained state wise. The input to the model is district, season, area (hectares) and crop. The output is production.

After splitting train and test data, random forest regressor was applied. For accuracy we consider the R2 score.

If $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value for total n samples, the estimated R^2 is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The R2 score function computes the coefficient of determination The R2 score in the case of crop production came out to be 72.7%.

The table below summarizes the R2 score accuracy for productions of different states.

State	Random Forest Regressor R2Score
Andaman and Nicobar Is-lands	0.9611
Andhra Pradesh	0.8471
Arunachal Pradesh	0.9463
Assam	0.6388
Bihar	0.7886
Chandigarh	0.9164
Chhattisgarh	0.5472
Dadra and Nagar Haveli	0.9209
Gujarat	0.9520
Haryana	0.8836
Himachal Pradesh	0.8973
Jammu and Kashmir	0.8346
Jharkhand	0.6125
Karnataka	0.9562
Kerala	0.9653
Madhya Pradesh	0.8539
Maharashtra	0.9510
Manipur	0.9532
Meghalaya	0.9814
Mizoram	0.8462
Nagaland	0.8828
Odisha	0.8077
Puducherry	0.8362
Punjab	0.9787
Rajasthan	0.8871
Telangana	0.9212
Tripura	0.9877
Uttar Pradesh	0.9937
Uttarakhand	0.5844

For the rainfall model the dataset was from 1980-2017, the training data taken as 70% and test data 30%.the accuracy is predicted by R2 score as 68.02%.

C. Random Forest Classifier

Random forest classifier purpose is to create a set of decision trees from randomly selected subsets of the training set. Later the final class of test object is decided based on the aggregate of the votes from multiple decision trees. Finally, it predicts based on the majority of votes from each of the decision trees made. Finally, predictions are made on the basis of majority votes from individual decision trees. The principle of random forest classifier is "Number of weak estimators when combined forms a strong estimator."

For the soil classification we used took the N, P, K (Nitrogen, phosphorus, potassium) values as HIGH, LOW, MEDIUM, VERY HIGH and VERY LOW of the soil as input and encoded it and classified organic carbon content as HIGH, LOW, MEDIUM, VERY HIGH and VERY LOW .The soil classification module was taken state wise Soil organic carbon contributes to nutrient retention and turnover, soil structure, moisture retention and availability, degradation of pollutants, and carbon sequestration.

The Accuracy score for the random forest classifier is given by formula -

Accuracy = TP+TN/TP+FP+FN+TN

Where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

The table below summarizes the classification score for soil's organic carbon of different states.

State	Random Forest Classification score
Andhra Pradesh	0.375
Arunachal Pradesh	0.857
Assam	0.467
Bihar	0.333
Chandigarh	0.467
Gujarat	0.2
Haryana	0.8
Himachal Pradesh	0.714
Jammu and Kashmir	0.428
Jharkhand	0.272
Karnataka	0.444
Kerala	0.75
Madhya Pradesh	0.733
Maharashtra	0.762
Manipur	1.0
Meghalaya	1.0
Mizoram	1.0
Nagaland	1.0
Odisha	0.6875
Punjab	1.0
Telangana	0.5
Tripura	0.6
Uttarakhand	0.375
Uttar Pradesh	0.907

D. Web Application

The web application is made in Streamlit using Python language. The input of the web application is to select the state, district, season, and area in hectares, crop to be grown and N, P, K values. The output will be production, yield, annual rainfall and Organic Carbon (OC). The real time weather will also be displayed along with the above output.

VI. Conclusion

We conclude that we move one step close for accurate yield, rainfall and soil nutrient prediction systems. With the use of ensemble learning methods, we can predict with great accuracy. For large scale predictions, we can use big data analysis and data mining techniques for better performance. The main factor here is the data, we need to deploy own sensors to take know the soil nutrients and weather data. Also there is a need to make farmers aware of such techniques. For future work we need to consider the other aspects like the fertilizers consumption in the farm, the topography of the area. A mobile application which will alert farmers through messaging service the right time to sow and harvest. There is a need to make technology available to the farmers.

Acknowledgment

Many thanks to our guide and mentor Prof.Sonali Bodekar for helping us throughout and guiding us in the right direction. Also, we would be glad to show appreciation to our examiner Prof.Kumud Wasnik for seeing through our work and motivat- ing us do better.We are grateful that we are able to complete this project on time with the persistent efforts of team members Aastha Doshi and Anuradha Chopade. In the end, I would like to thank our friends and family who believed in us.

References

[1] Rahul Katarya, Ashutosh Raturi, Abhinav Mehndiratta, Abhinav Thap- per, "Impact of Machine Learning Techniques in Precision Agricul- ture",3rd International Conference on Emerging Technologies in Com- puter Engineering: Machine Learning and Internet of Things (ICETCE- 2020), 07-08 February 2020.

[2] Bhawana Sharma,Jay Kant Pratap Singh Yadav,Sunita Yadav, "Predict Crop Production in India Using Machine Learning Technique: A Sur- vey",2020 8th International Conference on Reliability, Infocom Tech- nologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020

[3] Vanitha CN1, Archana N2, Sowmiya R3,"Agriculture Analysis Using Data Mining And Machine Learning Techniques",2019 5th Interna- tional Conference on Advanced Computing Communication Systems (ICACCS)

[4] Ramesh Medar,Vijay S. Rajpurohit,Shweta, "Crop Yield Prediction using Machine Learning Techniques",2019 5th International Conference for Convergence in Technology (I2CT) Pune, India.

[5] Jerome Treboux ,Dominique Genoud,"Precision Agriculture: An Ap- plication Of Improved Machine-Learning Algorithms",2019 6th Swiss Conference on Data Science (SDS)

[6] Shivani S. Kale,Preeti S. Patil,"Machine Learning Approach to Predict Crop Yield and Success Rate",2019 IEEE Pune Section International Conference (PuneCon) MIT World Peace University, Pune, India. Dec 18-20, 2019

[7] Dr. Y. Jeevan Nagendra Kumar1,V. Spandana2, V.S. Vaishnavi3, K. Neha4,V.G.R.R.Devi5,"Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector",Proceedings of the Fifth Interna- tional Conference on Communication and Electronics Systems (ICCES 2020) IEEE Conference Record 48766; IEEE Xplore ISBN: 978-1- 7281-5371-1

[8] Sk Al Zaminur Rahman, Kaushik Chandra Mitra, S.M. Mohidul Islam, "Soil Classification using Machine Learning Methods and Crop Sug- gestion Based on Soil Series",2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018

[9] Juhi Reashma S R K, Anitha S.Pillai, "Edaphic factors and crop growth using Machine learning – A Review",Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386- 1959-9

[10] Chandrasegar Thirumalai,M Lakshmi Deepak,K Sri Harsha,K Chaitanya Krishna "Heuristic Prediction of Rainfall Using Machine Learning Techniques",International Conference on Trends in Electronics and Informatics ICEI 2017

[11] Suman Avdhesh Yadav,Biswa Mohan Sahoo,Smita Sharma,Lipsa Das,"An Analysis of Data Mining Techniques to Analyze the Effect of Weather on Agriculture",2020 International Conference on Intelligent Engineering and Management (ICIEM)

[12] Jai Mahaprabhu A,Praveen Kumar V,Dr B.Latha,Gangadharan P S,"Cloud Analytics based Farming with Predictive Analytics using Artificial In- telligence",2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)

[13] Kuljit Kaur,Kanwalpreet Singh Attwal,"Effect of Temperature and Rain- fall on Paddy Yield Using Data Mining",2017 7th International Confer- ence on Cloud Computing, Data Science Engineering – Confluence

[14] https://data.gov.in/

[15] http://api.openweathermap.org

[16] https://soilhealth7.gov.in/