University of Barishal

Department of Computer Science & Engineering

Project Report

Crop Yield prediction in Bangladesh Based on Machine Learning

Mst. Masphy Masphy Akter

Student Id: 09-02-17

Batch: 02

Supervisor

Dr. Tania Islam

Assistant Professor

Department of Computer Science & Engineering

University of Barishal

APPROVAL BY THE SUPERVISOR

This project “Crop Yield prediction in Bangladesh Based on Machine Learning” conducted by Mst. Masphy Akter (09-02-17) is comply the requirements of academic integrity.

……..

Supervisor: Dr. Tania Islam
Assistant Professor
Department of Computer Science & Engineering
University of Barishal

Abstract

The prediction of agriculture yield is the one of the challenging problems in smart farming in Bangladesh. Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms have been applied to support crop yield prediction research.

We have predicted the yield of rice in different district of Bangladesh with the help of Machine Learning by considering the area, annual rainfall, fertilizer and pesticide. Here we have used Decision Tree Regression, Random Forest Regression and Linear Regression in order to predict the rice yield. From the experiments we got Random Forest regression to be the best with 98.77% accuracy.

Table of Contents

Chapter 1

1.Introduction

1.1 Background

Agriculture is a vital element that has a significant role in nourishing the world's growing population. To keep pace with the increasing demand for foodstuffs, farmers need to make the best use of them to reap output while minimizing losses. Forecasting and examining reap growth is a serious part of modern agriculture, and machine learning has become a powerful tool to achieve this goal line. Smart farming, or precision agriculture, is a modern farming conduct that utilizes recent technology to optimize reap production and minimize waste. Smart farming aims to increase reap output while minimizing using resources such as water, fertilizer, and energy.

In recent years, machine learning applications have entered our lives in many areas, from health to defense industries and education to urbanization, and have taken an effective way in decision-making situations. At the same time, it started to produce information and technology solutions by forming the basis of the newly emerging search engine infrastructure.

1.2 Objectives

Data Collection and Preprocessing: We collect and compile relevant datasets, including historical crop yield records, weather data, soil properties. We clean and preprocess the data to handle missing values, standardize features, and ensure consistency for machine learning model training.

Exploratory Data Analysis (EDA): We app data analysis to uncover trends, correlations, and patterns between input features (weather, soil, etc.) and crop yields. Visualize data insights and identify key factors that impact crop yields.

Machine Learning Model Development: We select multiple machine learning models (Random Forest, Linear Regression) to predict crop yields based on input data.

Model Evaluation: We use relevant metrics, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared and compared the models to determine the best-performing algorithm for crop yield prediction.

Feature Importance Analysis: To analyze the relative importance of different features (e.g., weather variables, soil quality, satellite data) to identify which factors have the most significant influence on crop yields.

Model Deployment and Testing: We deployed the final crop yield prediction model in a cloud-based or local environment and tested the model with real-world data and assess its performance in practical agricultural settings.

Chapter 2

2. Literature review

In its earliest, many analyses and models were being done for estimating the negative and positive effect of climate on various types of crops in Bangladesh and different nations too. Furthermore, they made a lot of algorithms for getting precise result. The worldwide interest in agriculture yields is expected to generally twofold by 2050, driven by expansions in the populace. Obviously, these new developments in worldwide yield produce lingers behind the ordinary requests, leaving us with a critical inquiry.

Aruvansh Nigam, Saksham Garg, Archit Agrawal, Parul Agrawal investigates Indian government dataset and it's been laid out that Arbitrary Timberland AI the calculation gives the best yield expectation precision. Random Forest Regressor gives the highest yield prediction accuracy. Simple Recurrent Neural Network performs better on rainfall prediction LSTM is good for temperature prediction. It can be improved by extends parameter to predict the more accurate result.

E Setiawati and W A Yusuf is gaining practical experience in the exactness and strength and relationship of irregular woods calculations. Arbitrary timberland calculation makes choice trees on various information tests and afterward predicts the information from every subset and afterward by casting a ballot offers better the response for the framework. Irregular Backwoods utilized the sacking strategy to prepare the information. To support the precision, the arbitrariness infused needs to limit the connection ρ while keeping up with strength.

Md. Toukir Ahmed1 *, Md. Niaz Imtiaz1 , and Nurun Sakiba Mitu1 have carried out crop yield expectation by utilizing simply the arbitrary backwoods classifier. Different elements like

precipitation, temperature, and season were considered to anticipate the harvest yield. Other AI calculations were not applied to the datasets. With the shortfall of other calculations, correlation and evaluation were missing and subsequently incapable to give the able calculation.

Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Prof. Vinu Williams has hypothetically portrayed different machine learning strategies that can be applied in different gauging regions. In any case, their work neglects to execute any calculations what's more, consequently can't give a reasonable understanding of the common sense of the proposed work.

Chapter 3

3. Proposed Method

This section outlines the methodology for developing a machine learning-based crop yield prediction model. The process includes data collection, preprocessing, model selection, feature engineering, training and validation, integration of satellite imagery, and the deployment of the final model.

3.1 Data Collection

Our intention was prediction of crop yield in different area of Bangladesh. To achieve our goal, we collected some raw data of rice. We collect data from different agricultural website, OpenWeatherMap and Kaggle.

To collecting data , we mostly focused on latest data with different variables , specially area, growing seasons, annual crop production , amount of pesticides used, total yield. For weather data, we collect data based on humidity, annual rainfall, average temperature etc.

3.2 Dataset Description

The dataset we collected has 10000 rows and 9 columns. There was no null value in the columns. The dataset contains data for different crops cultivated in various states across different seasons. Yield shows the highest percentage (26.15%), suggesting that the average yield is relatively closer to its maximum potential compared to other features. The average annual rainfall was 1218.39 mm .

For data analysis , we can conform that there yield is highly related to annual rain fall, fertilizer, pesticides and area. We can also notice that in summer season high yield produced. So there are must be some correlation between season, rainfall and fertilizer with crop yield.

3.3 Data Processing

3.3.1 Data Cleaning

Data cleaning is an importan part of data preprocessing, which is applied to enhance the quality of the dataset. In this step, we handle missing values in our dataset; and also drop irrelevant columns.

Table 3.1: Before Data Cleaning

	Crop	Season	State	Area	Production	Annual_Rainfall	Fertilizer	Pesticide	Yield
0	Rice	Autumn	Dhaka	607358.0	398311	2051.4	5.780226e+07	188280.98	0.780870
1	Rice	Summer	Dhaka	174974.0	209623	2051.4	1.665228e+07	54241.94	1.060435
2	Rice	Winter	Dhaka	1743321.0	1647296	2051.4	1.659119e+08	540429.51	0.941304
3	Rice	Monsoon	Chittagong	1031530.0	2340493	1266.7	9.817071e+07	319774.30	2.233500
4	Rice	Winter	Chittagong	53889.0	109350	1266.7	5.128616e+06	16705.59	2.073846

We checked the missing values but we didn't find any missing value. We drop some irrelevant columns like 'Crop', 'Season' and 'State' as they contain objective value.

Table 3.2: After Data Cleaning

	Area	Production	Annual_Rainfall	Fertilizer	Pesticide	Yield
0	607358.0	398311	2051.4	5.780226e+07	188280.98	0.780870
1	174974.0	209623	2051.4	1.665228e+07	54241.94	1.060435
2	1743321.0	1647296	2051.4	1.659119e+08	540429.51	0.941304
3	1031530.0	2340493	1266.7	9.817071e+07	319774.30	2.233500
4	53889.0	109350	1266.7	5.128616e+06	16705.59	2.073846

3.3.2 Data Encoding

The dataset contains three non-numerical variables such as crop, season, state .In order to apply the machine learning algorithm we need to convert these non-numerical feature into numerical form. we replace the categorical value with a numeric value between 0 and the number of classes minus 1. In python the LabelEncoder encodes the labels with a value between 0 and n_classes-1 where n is the number of distinct labels. When the labels getting repeated it will assign the same value to the labels as assigned earlier. Here the categorical data is converted into numerical form. Here LabeEncoder() package to encode the labels which have text data.

3.3.3 Dataset Splitting

In order to apply the machine learning algorithms to predict the rice yield we need to split the data into train and test sets. Here for this particular research work the data splits into 80% for training and 20% for testing. After the successful splitting we can apply the different supervised algorithms to predict the rice yield.

3.4 Apply Machine Learning Model

3.4.1 Linear Regression

Linear Regression

Linear regression is used to represent the relationship between two variables by applying a linear equation to the observed data. Here one variable is supposed to be an independent variable, and the other is to be a dependent variable. We can say that the measure of the extent of the relationship between two variables is shown by the correlation coefficient. The range of this coefficient lies between -1 to +1. The coefficient shows strength of the association of observed data for two variables.

3.4.2 Random Forest Regression

Random Forest is one of the best algorithms that are used for Machine Learning studies. Here we used random forest algorithm for regression studies. One of the great advantages of this algorithm is it predicts outcomes with higher accuracy most of the times even when data sets that does not have proper parameter tuning. Therefore, we can say that it has a simplicity compared to the other algorithms and it is very much popular. In the Random Forest it creates a forest in random manner. Here Multiple decision trees area created with this algorithm and are merged together in order to produce even more accurate predictions. From this algorithm we can say that, as the number of decision tree in this algorithm increases the stability of the predictions also increasing. Random forest itself is working as an ensemble algorithm hence it will give good accuracy in most of problems.

3.4.3 Decision Tree Regression

Here in the decision tree regression multiple regression trees are created from the dataset and then from these trees it will predict the future yield. With the help of a set of questions the decision tree is created. Based on the questions and the answers the decision nodes are created. By using MSE (mean squared error) value the nodes splits into sub nodes. Here we used decision regressor to predict the rice and pepper yield.

Chapter 4

4.Result and Discussion

4.1 Result

4.1.1Pre-processing Result

As there was no null values or missing values, we dropped down irreverent columns, Then we directly distlplot the dataset. It is represented that pesticides have a high effect on yield.

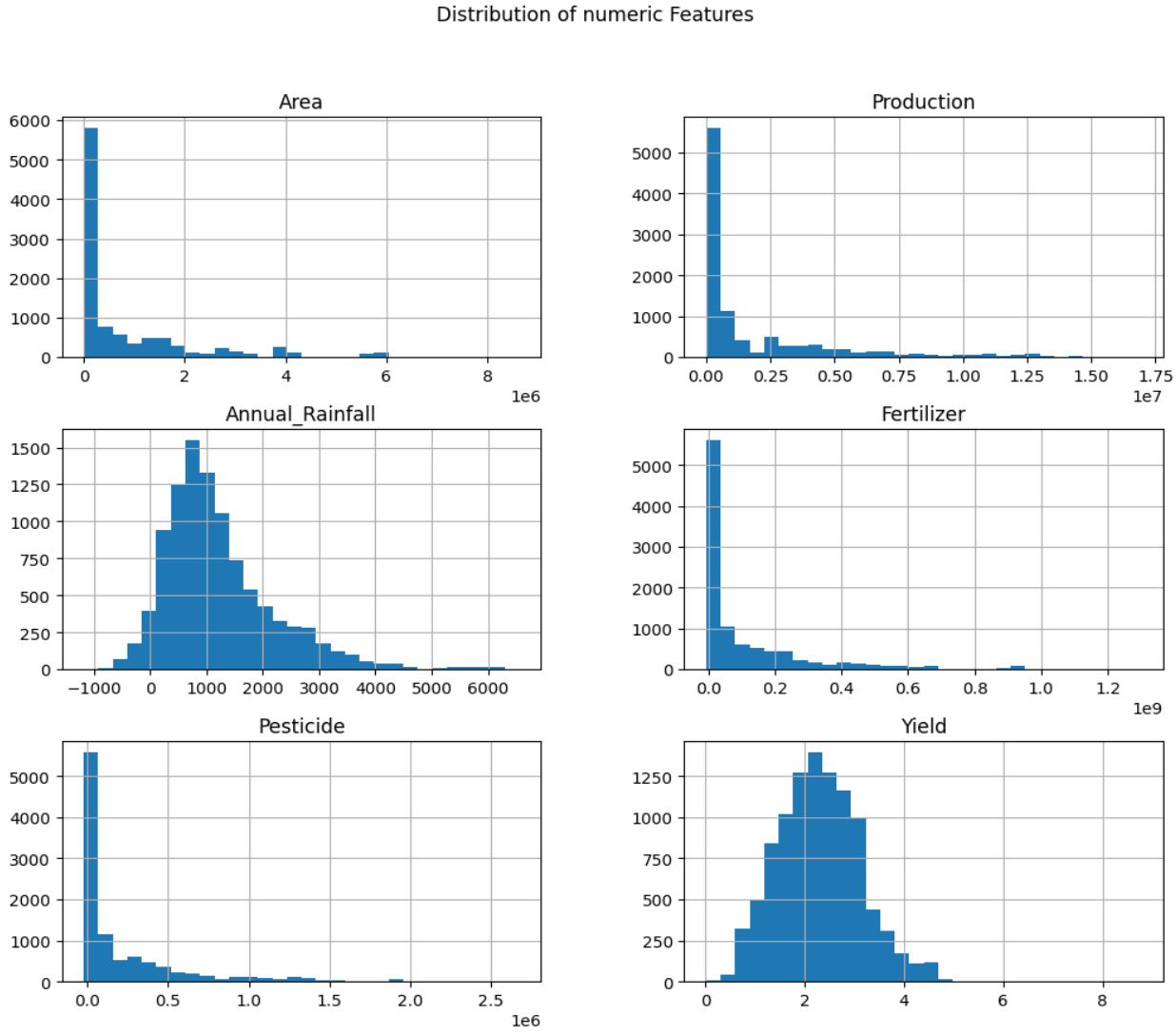


Fig 1: Distribution of numerical features

We visualize the correlation relation among features are given below. The diagram shows negative relation between high annual rainfall and yield, that means heavy rainfall decreases amount of yield. Conversely fertilizer and pesticide value are positively increasing the yield production.

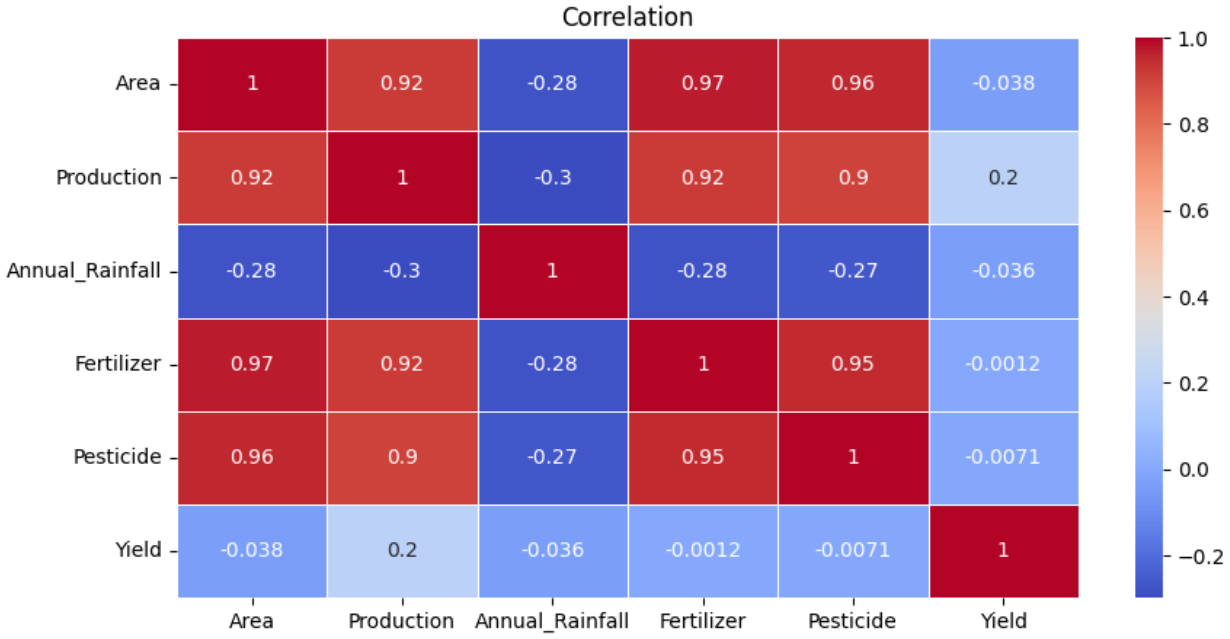


Fig 2: Correlation

4.1.2 Result Analysis for Machine Learning Model

We split the data into training and testing dataset. We applied different machine learning models to count mean squared error and R-squared error.

Table 4.1: MSE and r2 Score of different models

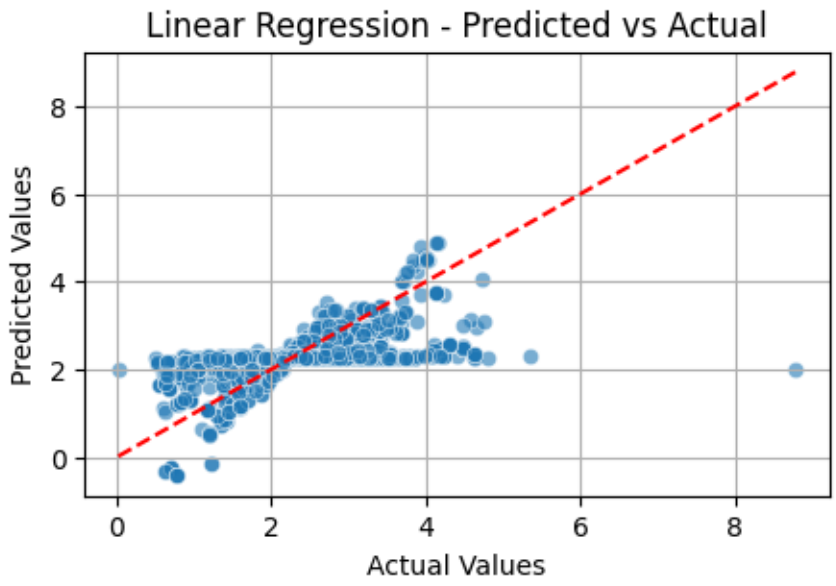
Mean Squared Error of Linear Regression	0.4619998281981817
R-sequared Score of Linear Regression	34.44396381863817
Mean Squared Error of Random Forest	0.039132642904837325
R-sequared Score of Random Forest	94.44722530710257
Mean Squared Error of Decision tree	0.045749528640184216
R-sequared Score of Decision tree	93.50831413398355

If MSE value is less and R-squared Score is high, the prediction is more accurate. The accuracy of prediction with actual data on different models are shown in the following diagram

For Linear Regression:

Mean Squared Error of Linear Regression: 0.4619998281981817

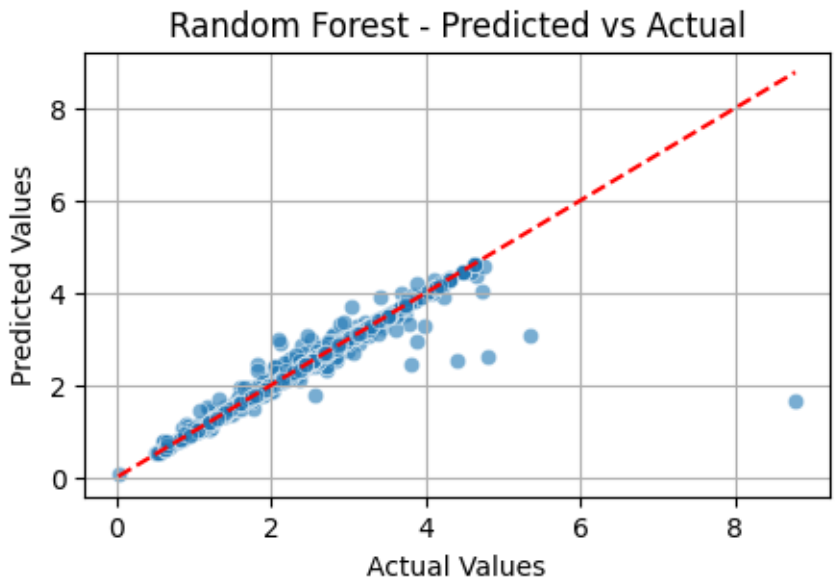
R-sequared Score of Linear Regression: 34.44396381863817



For Random Forest:

Mean Squared Error of Random Forest: 0.039132642904837325

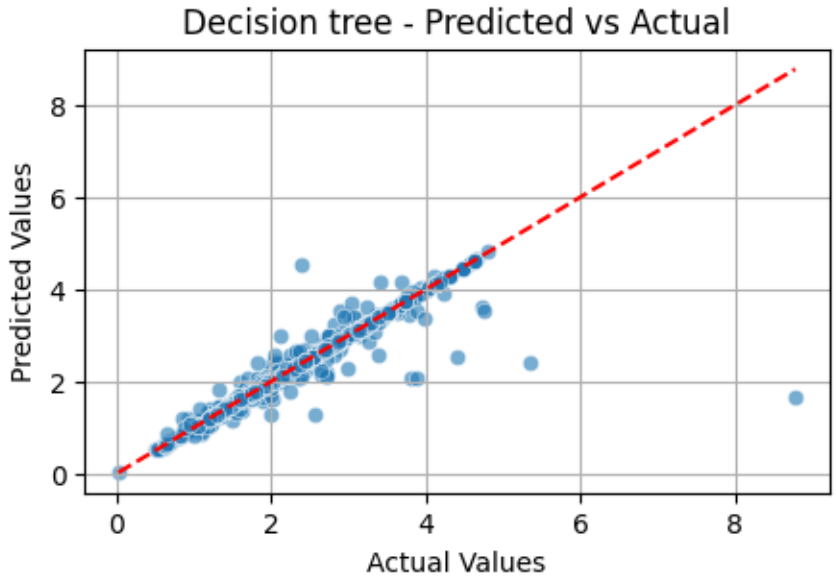
R-sequared Score of Random Forest : 94.44722530710257



For Decision Tree:

Mean Squared Error of Decision tree: 0.04556063726213749

R-sequared Score of Decision tree : 93.535117109348



From those diagrams, we can clearly declare that Random Forest model is suitable for crop prediction.

4.1.3 Features Importance

Different features have different importance to crop prediction. The ration of different models is following below the table.

Table 4.2: Features importance

Feature	**Importance (%)**
Production	46.75
Area	46.66
Fertilizer	4.64
Pesticide	0.99
Annual Rainfall	0.95

This table 4.2 we can summarize that the production and area are the most significant predictors, contributing nearly 93% of the model's predictive power where the fertilizer has a moderate impact with about 4.64% contribution. Pesticide and annual rainfall have very minimal influence, contributing less than 1% each.

4.1.4 Algorithm Result

Table 4.3: Algorithm Result

Model	R2 Score
Linear Regression	34.44
Random Forest	94.45
Decision tree	93.69

These are the different Machine learning regression algorithms used in this particular research. We can say that using these algorithms we can predict the future yield of a crop. Here we can predict the future yield of the crops. The different models predicted the future yield of crops with a good accuracy. The rice dataset giving the highest accuracy for Random Forest regression. Out of those 3 models we can fix a best model for the deployment of the model.

Chapter 5

5. Conclusion

The demand and supply for food have grown more difficult to manage as the population grows. To assist farmers, experts have worked hard over the past few years to anticipate agricultural yield production. In order to forecast crop yield, this study uses various machine learning and deep learning approaches. The study underlines the advantages of cutting-edge procedures. It is beneficial for small-scale ranchers, as they may use the predictions to estimate crop production for upcoming years and plant it appropriately. We apply three machine learning and deep learning algorithms, Decision Tree, Random Forest to the dataset taken into consideration. When data is analyzed at the country level, Random Forest Regression and Decision Tree Regression perform better. Experimental findings demonstrate that the approach has a great potential for precise crop productivity prediction and its effectiveness has been validated using real-time data and interactions with people. More data for each crop year having more historically precise information about the climate and environment is needed. To increase the model's accuracy in crop production prediction, remote sensing data could be amalgamated with statistical data of districts.

Chapter 6

6. Reference

[1] Nigam, A., Garg, S., Agrawal, A. and Agrawal, P., 2019, November. Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE.

[2] Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment agriculture. Environmental Research Letters, 13(11), p.114003. in

[3] Van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, p.105709.

[4] Sellam, V., and E. Poovammal. "Prediction of crop yield using regression analysis." *Indian Journal of Science and Technology* 9.38 (2016): 1-5.

[5] Hegde, Niranjan G., et al. "Survey paper on agriculture yield prediction tool using machine learning." Int. J 5 (2017): 36-39.

[6] Priya, P., U. Muthaiah, and M. Balamurugan "Predicting yield of the crop using machine learning algorithm." International Journal of Engineering Sciences & Research Technology 7.1 (2018): 1-7.

[7] Hassan, Fazal Mahmud, et al. Agricultural yield and profit prediction using data analysis techniques. Diss.2018.

[8] Sivanand, K., M. Sai Amrutha, and J. Chaitra Nayak. "Web Application Development for Site Specific Crop Prediction using Machine Learning." International Journal of Modern Agriculture 10.2 (2021): 3538-3549.

[9] Kumar, Arun, Naveen Kumar, and Vishal Vats. "Efficient crop yield prediction using machine learning algorithms." International Research Journal of Engineering and Technology 5.06 (2018): 3151-3159.