IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

Crop Yield Prediction using Machine Learning

Mr. V. Shanmugam[1], I. Sriteja[2], K. Sai Dathu[3], K. Raju[4], S. Sai Kumar[5], G. Karun[6]
Assistant Professor, Department of Computer Science & Engineering[1]
UG Students, Department of Computer Science & Engineering[2,3,4,5,6]
Christu Jyothi Institute of Technology & Science, Jangoan, Telangana, India

Abstract: Weather profoundly impacts agricultural outcomes, making accurate crop prediction vital for farmers' decision-making. This abstract presents a comprehensive overview of weather-based crop prediction, emphasizing its significance, keycomponents, and methodologies. The process begins with the collection and analysis of historical weatherdataencompassing variables such as temperature, precipitation, humidity, and sunlight. Utilizing Python programming and data visualization libraries like Pandas and Matplotlib facilitates the exploration and visualization of this data, revealing trends and patterns. Machine learning algorithms, including regression and ensemble methods, are employedto develop predictive models. These models leverage historical weather data to forecast future crop yields accurately. Python's extensive libraries, such as Scikit-learn and TensorFlow, offer robust tools for model development and evaluation.Incorporating advanced technologies like remote sensing and satellite imagery further refines the prediction process. These tools provide real-time insights into crop health and growth, enhancing the precision of forecasts. Ultimately, weather-based crop prediction serves as a valuable decision support tool for farmers, enabling informed choices regarding planting, irrigation, and harvesting practices. By harnessing historical weather data, machine learning algorithms, and innovative technologies, stakeholders can optimize agricultural productivity, mitigate risks, and contributeto global food security.

Keywords: Agriculture, KMeans Clustering, Logistic regression Algorithm, Crop yield prediction, Machine learning method

I. INTRODUCTION

India's economy is built on agriculture, which employs a sizable percentage of the country's workers and makes up a sizable chunk of its GDP. However, the agricultural industry faces many difficulties, such as erratic weather patterns, variable soil, and shifting consumer needs. The effect of climate change on agricultural productivity has been more noticeable in recent years, which emphasizes the need for creative solutions toreduce risks and increase crop yields. This research uses machine learning algorithms and sophisticated data analysis methods to tackle the important problem of crop prediction in India. It makes use of an extensive dataset that includes data on numerous crops as well as important climatic parameters including temperature, precipitation, humidity, and soilcharacteristics. The goal is to create a prediction model that can precisely estimate crop yields based on weather conditions. The project's first stage entails exploratory data analysis (EDA) to learn more about the features and organization of the dataset. Histograms and heatmaps are two examples of data visualization techniques that are usedto find patterns and relationships between various variables. Understanding the distribution of important nutrients such as potassium, phosphorus, and nitrogen among various crop varieties as well as their interactions with environmental conditions is madeeasier with the use of EDA.

Moreover, the dataset is divided into discrete groups according to similarities in crop traits and environmental factors using clustering algorithms like K-means. With the helpof this clustering, groups of crops with comparable responses to environmental cues canbe identified, opening the door to more focused agricultural management techniques. Then, to predict crop yield, a logistic regressionmodel is developed using supervised learning techniques. After that, crop labels are predicted by training a logistic regressionmodel with supervised learning approaches basedon input parameters like temperature, rainfall, humidity, andsoil conditions. Metrics including classification accuracy, precision, recall, and F1-score are used to assess the model'sperformance and provide information.

DOI: 10.48175/IJARSCT-18185

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

In order to facilitate well-informed agricultural decision- making, the overall goal of this project is to offer a thorough framework for crop prediction in India that integrates data analytics, machine learning, and domain expertise. In the face of shifting environmental conditions and climatic uncertainty, stakeholders may maximize resource allocation, reduce risks, and improve agricultural productivity by utilizing the power of data-driveninsights. In order to identify underlying patterns and interactions between various variables, a thorough exploratory study of the dataset is required in the first phase of the project

The research aims to decipher complex relationships between important agronomic components including nitrogen, phosphorus, potassium, pH levels, and environmental variables using visualizations like histograms,heatmaps, and scatter plots.This exploration stage clarifies the distribution of critical nutrients among various crops as well as how they interact with climate conditions, providing important information for further modelling endeavors. Additionally, by using clustering techniques like K-means, the dataset can be divided into logical clusters according to similarities in crop traits andenvironmental factors. This clustering methodology makes it easier to identify different groups of crops that respond to environmental stimuli in a similar way. specific crop clusters through targeted agronomic interventions that maximize resource allocation andboost agricultural productivity.

Logistic regression is employed as a pivotal tool for crop prediction within the agricultural decision support system (ADSS). Logistic regression is a widely-used statisticalmethodfor binary classification tasks, but it can also be extended to handle multiclass classification, as is likely thecase here given the multiple crop labels involved. Logistic regression excels in situations where the relationship between the input features and the target variable (in this case, crop labels) is nonlinear, offering a flexible and interpretableapproach to classification. Initially, the dataset is split into training and testing sets usingthe train tests Split function from scikit-learn, a standard practice to assess the model's generalization performance. Following this, a logistic regression model is instantiated using the Logistic Regression class from scikit-learn, with various parameters specified to control the model's behaviors, such as max_iter, which defines the maximum numberof iterations allowed for model convergence during training. The model is then trained on the training data using the fit method, during which it learns the underlying patterns and relationships between the input features (e.g., temperature, humidity,rainfall, soil nutrients) and the target variable (crop labels). This process involves optimizing the model's parameters tominimize the difference between the predicted crop labels and theactual labels in the training data.

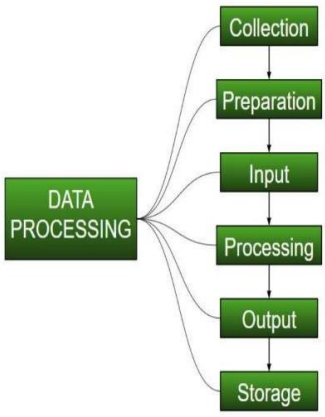

Figure 1: LOGISTIC REGRESSION

Once the model is trained, it is evaluated using the testing data to assess its performanceand generalization ability. Predictions are made on the testing data using the predict method, generating predicted crop labels based on the input features. Classification metrics such as precision, recall, F1- score, and accuracy are then computed using the classification report function from scikit-learn, providing valuable insights into the model's performance across different classes. Additionally, a confusion matrix is generated using the confusion matrix function to visualize the model's performance in predicting each class. This matrix provides a clear summary of the model's true positive, false positive, true negative, and false negative predictions, offering a comprehensive overview of its classification accuracy and errors. Overall, logistic regression serves as a fundamental component of the ADSS, offering a reliable and interpretable approach to crop prediction based on environmental and soil factors. Its simplicity and effectiveness make

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

it a valuable tool for empowering farmers with data-driven insights, enablinginformed decision-making and optimized agricultural practices.

II. PROBLEM DEFINATION

This project in agricultural data analysis aims to recommend crops based on environmental factors like temperature, humidity, and rainfall. Through machine learning and predictive analytics, it optimizes resource utilization while mitigating risks such as adverse weather and soil degradation. By fostering data-driven decision-making, it enhances agricultural productivity and resilience, contributing to global food security. The code imports libraries for data manipulation, visualization, and machine learning, analyses historical crop and weather data, and employs clustering and logistic regression for prediction. Visualization tools like histograms offer insights into data distributions, while interactive analysis enables dynamic exploration of crop requirements. Overall, the project amalgamates technology with agricultural expertise to address key challenges in crop recommendation and resourcemanagement.

Utilizing machine learning algorithms like clustering and logistic regression, the project identifies patterns and trends within the data, facilitating the accurate prediction of crop yields under varying conditions. This predictive capability empowers farmers and stakeholders to make informed decisions regarding crop selection, planting schedules, and resource management strategies. Moreover, by visualizing data distributions and conducting interactive analyses, the project fosters a deeper understanding of the underlying factors influencing crop requirements and growth dynamics.

Beyond its immediate applications in agricultural decision- making, this project holds broader implications for global food security and sustainability. By optimizing crop recommendations based on environmental suitability and resilience, it contributes to the efficient utilization of land and resources, reducing wastage and environmental impact. Moreover, by promoting the adoption of data-driven approaches in agriculture, it catalyses innovation and resilience within the sector, paving the way for a more sustainable and food-secure future. In essence, this project represents a concerted effort to harness the power of data and technology in addressing critical challenges facing agriculture while ensuring the long-term viability of food production systems.

III. MOTIVATION

This initiative is driven by the urgent need to enhance agricultural techniques in light of the increasing global food demand and climate change. The project aims to reduce the risks associated with environmental unpredictability by improving crop productivity and resilience through the application of machine learning and data analysis tools. The intention is to provide farmers with practical insights so theycan make well-informed decisions, which will improve resource allocation and promote sustainable farming methods. Additionally, the initiative intends to solve issues related to food security and guarantee the provision of nutrient-dense food for future generations by selecting crops based on environmental adaptability. Adopting data-driven methods in agriculture promotes creativity and makes it easierto adjust to changing environmental circumstances, which eventually increases the resilience and sustainability of agricultural systems on Global.

IV. PROPOSED MODEL

This system predicts which crop should be cultivated based on the weather conditions. These are the steps that are employed in the process.

- Data Collection
- Data Preprocessing 3.Training and Testing
- Selecting Algorithm
- Prediction

V. PROPOSED ALGORITHM

The following suggests the pseudo code for the proposed crop prediction.

- Data Collection
- Data Preprocessing

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

- Handle missing values, outliers, and inconsistencies
- Normalize or scale numerical features
- Encode categorical variables if necessary
- Split the dataset into features (X) and target labels (y)
- Exploratory Data Analysis (EDA)
- Visualize the distribution and relationships betweenvariables
- Identify patterns and trends in the data
- Training and Testing
- Selecting Algorithms
- Model Training and evaluation
- Prediction

VI. MODELS USED

The models used in our project are:
1. Logistic Regression
2. K Means Clustering

LOGISTIC REGRESSION

For applications involving binary classification, a basic supervised learning approach is logistic regression. Logistic regression models the probability of a binary result depending on one or more predictor variables, in contrast to linear regression, which predicts continuous numerical values. It calculates the probability that a given observation falls into a given class, which is usually expressed as a binary result.

Logistic regression can be utilized in agricultural data analysis to forecast the probability of crops being appropriatefor particular environmental circumstances. Logistic regression models are able to determine the correlations between crop adaptability and environmental variables like temperature, humidity, and rainfall by examining historical crop production data in conjunction with these variables.

The logistic function, sometimes referred to as the sigmoid function, transfers input variables to a probability between 0 and 1, which is used by the logistic regression model to estimate probabilities. The method is able to classify data into discrete classes by thresholding this probability.

Stakeholders can evaluate the dependability and efficacy of crop recommendation models based environmental circumstances by training logistic regression models on labeled datasets and assessing their performance using measures like accuracy, precision, recall, and F1-score.

Logistic regression is a supervised learning algorithm used for binary classification tasks.It models the probability of a binary outcome (e.g., presence or absence of a certain event) based on one or more predictor variables. In this project, logistic regression is utilized to predict the probability of crops being suitable for specific environmental conditions. By analysing the relationships between environmental variables and crop suitability, logistic regression enables the algorithmto recommend crops tailored to specific environmental contexts.

K MEANS CLUSTERING

The tasks involving clustering, KMeans is a well-liked unsupervised machine learning technique. With each data point belonging to the cluster with the closest mean, or centroid, the algorithm seeks to divide the dataset into K clusters. It works in an iterative manner to reduce the inertia,or total squared distance, between data points and the centroids of each cluster.

When analysing agricultural data, KMeans clustering can be especially helpful in identifying discrete sets of environmental circumstances that are favourable to certain crop varieties. Through the examination of datasets that comprise characteristics like temperature, humidity, rainfall, and soil composition, KMeans can identify innate groups or clusters that signify ideal circumstances for particular crops. These clusters assist in providing important insights into the environmental preferences of various crop kinds.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

Furthermore, by helping with data exploration and visualization, KMeans clustering enables stakeholders to spotpatterns and trends in agricultural datasets. Farmers and policymakers can better comprehend crop-environment interactions and make more educated decisions about crop selection and management strategies by visualizing clusters and the environmental features that correspond with them. An unsupervised machine learning technique called KMeans is employed in clustering.

Based on how closely the data points resemble the cluster centroid, it divides the data into K clusters. In this study, clusters corresponding to various crop varietiesare identified by grouping similar data points representing environmental circumstances using KMeans. KMeans facilitates crop selection based on environmental parameters by grouping data points with comparable environmental features. This helps discover ideal circumstances for various crops.

V. IMPLEMENTATION AND METHODOLOGY

Data Collection:

Gather information on agricultural practices, including statistics on crop yield and environmental factors including temperature, humidity, rainfall, and soil composition. Government databases, IoT sensors placed in agricultural fields, and agricultural research institutes are a few places where this data can be found.

Preprocessing Data:

Address outliers, inconsistent data, and missing values to clean up the dataset. Scale or normalize numerical features to guarantee consistency and enhance model efficiency. If required, encode categorical variables using methods such as one-hot encoding. Divided the dataset into target labels (y) and features (X).

Training and Testing:

Use strategies like train-test split to divide the pre-processed dataset into training and testing sets. Define evaluation measures, including F1-score, recall, accuracy, and precision to calculate model performance.

Choosing Algorithms:

To comprehend the distribution and interactions between variables, perform exploratory data analysis (EDA). Select appropriate machine learning algorithms in accordance with the problem's nature. Logistic regression and KMeans clustering are two popular techniques for crop prediction based on environmental factors. Apply the chosen algorithms with Python packages such as scikit-learn.

Training Models:

Use the fit() method to train the selected algorithms on the training dataset. To maximize model performance, fine-tune hyperparameters using methods like grid search or random search. Determine the number of clusters (K) for KMeans clustering using methods like the elbow approach.

Model Evaluation:

Apply the previously created evaluation criteria to assess the trained models on the testing dataset. Examine the models' performance.

Prediction:

Apply the learned models to forecast fresh or unobserved data. Enter environmental variables into the logistic regression model to forecast the appropriateness of various crops in order to make crop recommendations. Use the trained KMeans model to cluster additional environmental data in order to determine the best condition.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

Iterative Refinement:

Try out various algorithms, feature engineering methods, andhyperparameter tuning approaches to improve the model development process iteratively. Gather fresh data on a regular basis and update models to keep them current and useful. .

VI. RESULTS

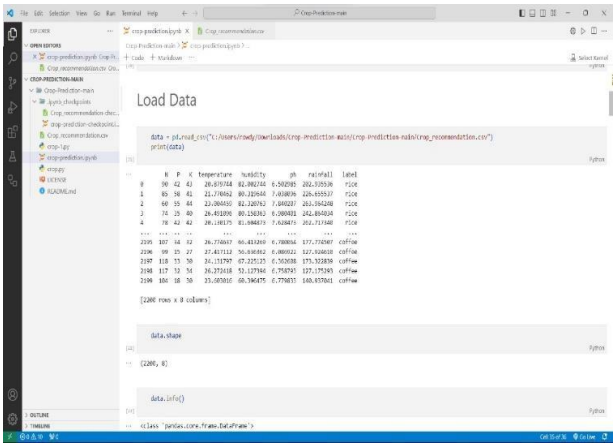

Figure 2:DATA SET

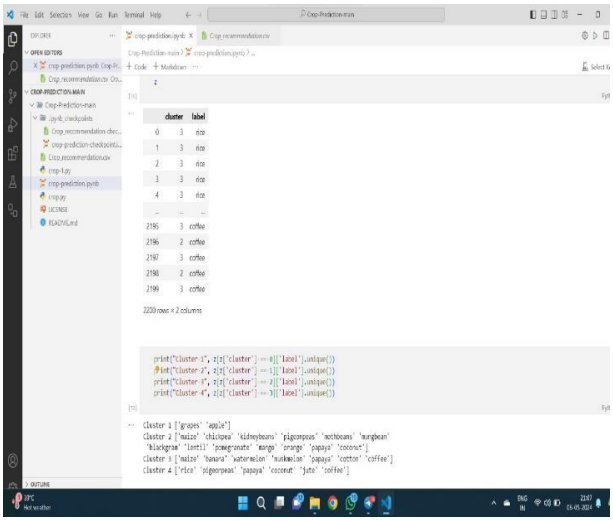

Figure 3:CLUSTER LABLES

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

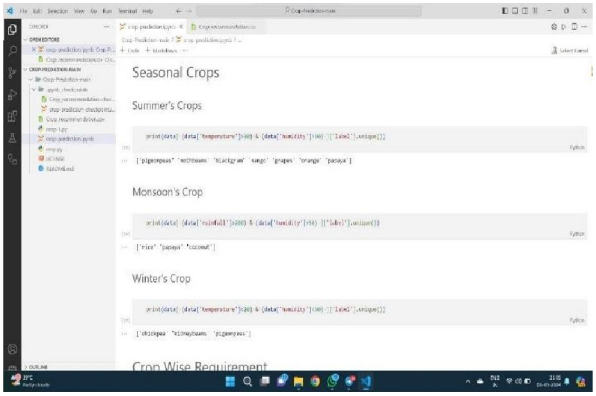

Figure 4: SEASONAL CROPS

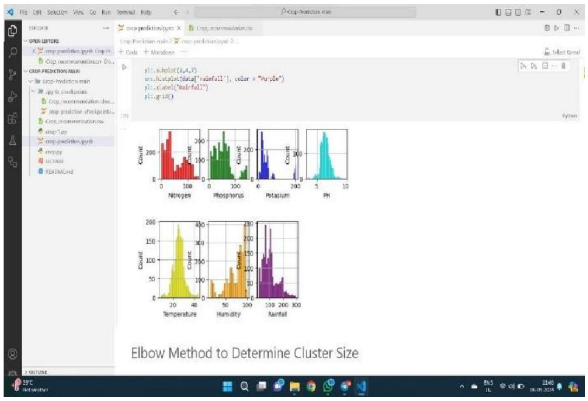

Figure 5:GRAPHS FOR INPUT VALUES

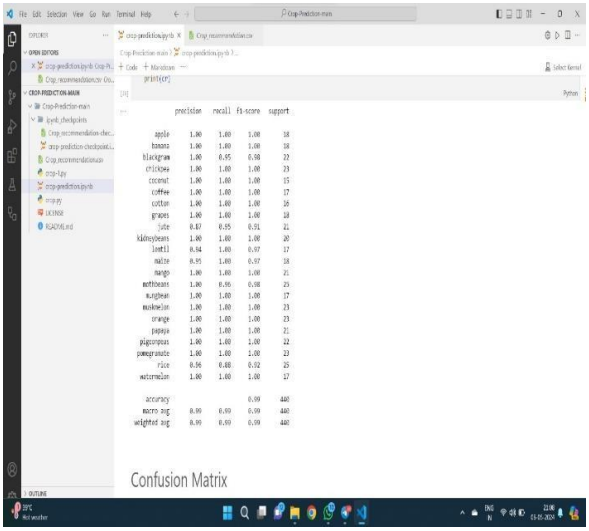

Figure 6:VALUES TO DRAW CONFUSION MATRIX

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

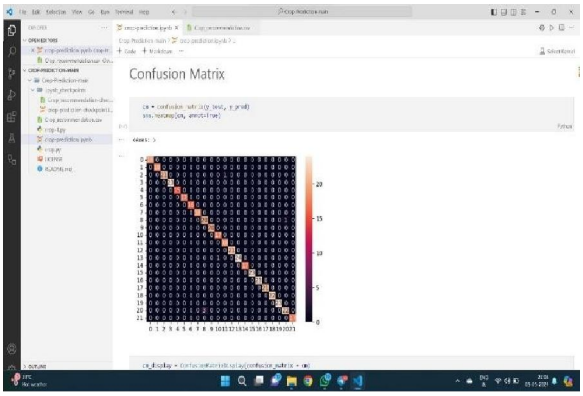

Figure 7: CONFUSION MATRIX

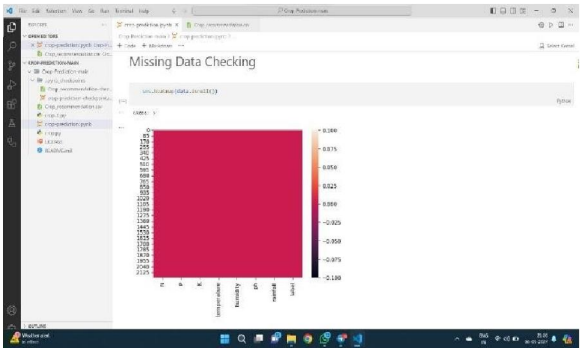

Figure 8: HEAT MAP

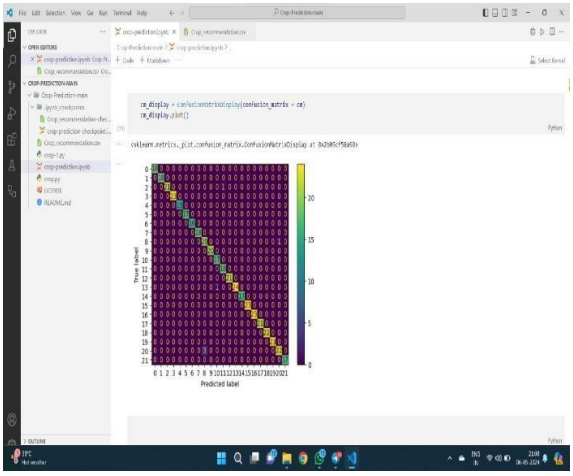

Figure 9: PREDICTION TABLE

VII. CONCLUSION

The Crop prediction using machine learning offers a thorough method for crop recommendation depending on environmental factors. Through the application of machine learning methods like logistic regression and KMeans clustering, the research offers insightful information on the best ways to cultivate crops. The initiative engages in thorough data preprocessing, exploratory analysis, and modeltraining to enable policymakers and farmers to make well-informed decisions. The research enhances agricultural production, sustainability, and resilience to changing climate

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

dynamics by predicting crop adaptability and identifying ideal climatic conditions. The initiative highlights the potential of data-driven approaches to transform contemporary agriculture and tackle global food security issues by providing stakeholders with practical advice.

REFERENCES

[1]. Aruvansh Nigam, Saksham Garg, Archit Agrawal[1]conducted experiments on Indian government dataset and it's been established that Random Forest machine learning algorithm gives the best yield prediction accuracy. Sequential model that's Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction. The paper puts factors like rainfall, temperature, season, area etc. together for yield prediction. Results reveals that Random Forest is the best classier whenall parameters are combined.

[2]. Leo Bryman [2] is specializing in the accuracy and strength& correlation of random forest algorithm. Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better the answer forthe system. Random Forest used the bagging method to trained the data. To boost the accuracy, the randomness injected has to minimize the correlation while maintaining strength.

[3]. Balamurugan [3], have implemented crop yield prediction byusing only the random forest classifier. Various features likerainfall, temperature and season were taken into account to predict the crop yield. Other machine learning algorithms were not applied tothe datasets. With the absence of other algorithms, comparison and quantification were missing thusunable to provide the apt algorithm.

[4]. Mishra [4], has theoretically described various machine learning techniques that can be applied in various forecasting areas. However, their work fails to implement any algorithms and thus cannot provide a clear insight into the practicality of the proposed work

[5]. Gautam Naidu and M. Prasanna Kumar's[5],"A review on applications of machine learning techniques in agriculture" was published in 2019An in-depth summary of machine learning applications in agriculture, such as disease detection, pest control, and crop production prediction, is given in this article. It talks about several approaches, difficulties.

[6]. Peng Yu et al[6], (2020) published "Predicting Crop Yield, Nitrogen Use, and Nitrogen Concentration in Maize Using Remote Sensing Data": This study investigates the prediction of crop production, nitrogen usage, and nitrogen concentration in maize usingdata from remote sensing. It highlights the promise for remote sensing-based methods inagricultural monitoring by proving the viability andaccuracy of estimating crop yield using satellite photos andmachine learning algorithms.

[7]. Md. Shahriar Iqbal et al [7], (2021), "Predictive modelling for maize yield prediction using machine learning algorithms": The goal of this work is to apply machine learning algorithms to forecast maize yield by utilizing meteorological data. It highlights the effectiveness of ensemble learning techniques by comparing the performanceof several machine learning models, such as random forests, support vector machines