Crop Yield Prediction System: A Comparative Analysis of Random Forest and XGBoost Models

Sukhvinder Singh Bamber[1], Rajeev Kumar Dang[2], Naveen Dogra[3], Gurpreet Singh[4]

[1,3]Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University SSG Regional Centre, Hoshiarpur, Punjab, India.
[2]Mechanical Engineering, University Institute of Engineering & Technology, Panjab University SSG Regional Centre, Hoshiarpur, Punjab, India.
[4]Electronics & Communication Engineering, University Institute of Engineering & Technology, Panjab University SSG Regional Centre, Hoshiarpur, Punjab, India.

Abstract

The continuing demand for food calls for the agricultural sector to stride towards employing optimization strategies in the prediction of accurate yield to effect proper resource management. Therefore, this study seeks to uniquely identify a systematic machine learning (ML) algorithm for predicting crop yield across varied agricultural settings. Predictive models have been developed using data from various sources: weather patterns, soil characteristics, Satellite Images, and historical yield records by various algorithms such as Random Forest and XGB regressor. All these other means of solving models have been evaluated concerning where Deep learning models' behaviors are compared both from RSME numbers and R2 as an array of other analyses in their accuracy comparisons. Thus, proposed findings suggest that ML models consistently overcome cases of traditional statistical methods in accurately estimating crop yields. In effect, this study examined the real possibilities of ML for agriculture as well as insight into the most influential factors upon yield levels; this, as a consequence, facilitates further sustainable practices and food security.

Keywords: Machine Learning; Random Forest; Random Forest; XGBoost Regressor.

1. Introduction

With projections indicating a population rise beyond nine billion by 2050, agriculture finds itself under pressure like no other. On the other hand, the demand imposed many constraints on crop productivity within the framework of traditional agricultural systems, creating an impetus for radical technological innovations striving towards agronomic modernization for increased crop production. Disaster crop yield predictions provide the basis for effective agricultural planning, resource allocation, and risk management and are thus essential for the effective fulfillment of food security and sustainable supply.

The machine learning subdivision of artificial intelligence has recently started to offer crop yield forecasts with enhanced opportunities. They have enjoyed a lot of preferences, by providing that much-desired cutting-edge to `Agro' decision-making by exploring mountains of available data left unexplored. Areas of analysis for crop yield include climate conditions, soil properties, plots of agricultural land as viewed from satellite images, and historical yield data. Again, applied to a large dataset for a wider area, machine learning algorithms make more accurate predictions compared to traditional statistical methods.

Traditional agricultural yield forecasting methods involve regression and expert opinions that normally operate within limited information which may sometimes not be adequately generalizable or may have the future factors verified for the same. Such models, therefore, suffer from improper optimization and firmware mismatch concerning rapidly changing environmental situations. Whereas ML methods allow interoperability amongst many relations by integrating data of different types to bring out nuances of relationships amongst the differing parameters that affect crop yield. The model, for example, might allow the efficient integration of temporal and spatial data into precise seasonal variations to show how regional differences impacted predictions.

Implementation in ML using agricultural yield prediction is composed of several sub-processes, such as data collection, data cleaning, model selection, training, and evaluation. Data collection is the most vital task as it makes desirable possibilities for framing models with wide and appropriate datasets. This data will consist of climate and soil characteristics as well as cultivational techniques and socioeconomic factors. Following data collection, preprocessing provides various techniques such as normalization and feature extraction to improve the overall quality of the machine-learning classifiers. Machine learning algorithm selection should be conducted more judiciously in terms of agricultural context and type of data, Random Forests, SVM, or Neural Networks.

To further facilitate the evaluation of the modeling, a definition of the robustness and applicability of methods will have to be made. Metrics in which conventional measures, such as root mean square error and R^2, could be applied to evaluate correctness would also have to be undertaken. Further studies that compare against classical techniques would provide insights into the advantages ML techniques offer concerning yield prediction challenges. It has to be demonstrated, through research, how machine learning contributes toward the prediction of crop yield. The aim is to help nurture knowledge development using different datasets, forming a sound basis for decision-making tools for farmers, agronomists, and policymakers.

2. Literature Review

Authors Ranjani, et al. (2021) highlighted various approaches proposed in ML for crop yield prediction. In these attempts, it has been observed that the most widely preferred approach is the RF one. RF is preferred because its robust handling in classification and regression tasks reduces yield losses associated with environmental factors like climate and soil conditions. The studies compared RF with other models including DL and ELM when applied to predict specific crops, such as corn and coffee. DL models are often more accurate but computationally more intensive while ELM is helpful in smaller farm applications, offering better feature extraction as compared to RF. These results now show the power of ML for predictive agriculture, thus enabling a better crop selection with the maximization of yields under appropriate weather and climate conditions and other agricultural inputs. Authors Vanitha K. et al. (2024) called for the further improvement of agricultural crop yield predictions over different Indian states using ML models, such as Random Forest (RF), Gradient Boosting, and Linear Regression. The critical determining factor involved in the improvement of these ML models is the application of the Yeo-Johnson power transformation because this helps alter the non-normal data distribution; hence, their accuracy is improved through reduced skewness and stabilized variance. The experiments conducted in the review show that models including those transformations outperform the baselines both on yield prediction in a real setting with yields and on large diversified datasets when there is a rich complexity in the scenarios. This approach also underlines the need for complex preprocessing techniques required to provide accurate actionable insight in agricultural forecasting. As such, this model aims to promote further effective decision-making by stakeholders, enhancing India's agriculture and food security.

Researchers Sunil G. L. et al. (2022) discussed various techniques related to machine learning that are used for the prediction of crop yield. It compares supervised learning methods, which include classification and regression. The comparison highlights the effectiveness of algorithms like ANN, SVM, and Random Forest. In general, classification methods produce higher accuracy in multiple crop yields than regression. The paper also discusses unsupervised learning techniques such as clustering and association rule mining for pattern discovery in unlabeled data. The review suggests that possibly future work should explore coupling the advanced models as hybrid frameworks with deep learning and consider further work in unsupervised and hybrid techniques for accurate and robust prediction of crop yields. Arumuga A. R. et al. (2023) discussed machine learning approaches to the prediction of agricultural yields. In this paper, the emphasis will be on the numerous approaches that have been proposed for supervised and unsupervised learning methods that can apply to the subject matter. Given the complex climatic and geographic variations involved in crop forecasting, it has been seen that AI techniques are increasingly being adopted. The literature review, therefore, addresses the issue of efficiency in deep learning models like DNNs more so than traditional techniques used in handling complex data. Models like Random Forest, XGBoost, and Support Vector Regression are also tested to validate their accuracy level of prediction. Such a study highlights the integration of varied nature parameters and sophisticated algorithms for more accurate and reliable yield forecasting which helps farmers in making the right

decisions while minimizing agricultural losses. Kumar et al. (2023) cleared that ML has become important in optimizing crop selection and farming. The article discusses algorithms such as neural networks, decision trees, and ensemble models to enhance the capability of deciding agriculture. An analysis has been performed on datasets related to soil quality, climate conditions, and historic crop data by using ML techniques. The review also discusses real-time data acquisition, perhaps through IoT sensors and satellite imagery, to monitor crop health and predict the need for irrigation. It focuses on sustainable agriculture with an aim towards both yield improvement and environmental attenuation while supporting farmers' adaptation to climate variability through predictive analytics.

Researchers Krishna V. et al. (2022) focused on various machine learning techniques used to improve crop yield prediction. The author emphasizes the critical role of agriculture in supporting economies and meeting food demands, with an underlying tone being the role of technology in boosting productivity. Algorithms from the lists such as K-Means clustering, Random Forest, and linear regression will be dealt with and indicate better accuracy of predictions once adapted approaches based on existing ones are used. Specific studies include discussions on how the variation of variables, such as soil type, temperature, and rainfall, plays an instrumental role in predictive models with artificial neural networks being the most recommended. Of importance, the review emphasizes the constantly changing aspects of machine learning integration in agriculture, further able to improve climate and crop data in farmers' decision-making processes. Authors Devan K.P.K. et al. (2023) focused on various machine-learning approaches that are used for crop yield prediction and fertilizer recommendation. It reviews several algorithms, including Random Forest, Logistic Regression, RNN, LSTM, and SVM. The respective applications involving its use for crop yield prediction in the appropriate accuracy ranges are documented. Advanced models, such as deep reinforcement learning, CNN-RNN hybrid frameworks, and ensemble methods, are discussed for their potential to achieve better prediction accuracy. The outcome of these studies puts forward that the data on weather, soils, and crop characteristics can be used. Furthermore, it compares models in a review that claims this faces the problem of being computationally complex and involving holistic data. The review generally shows the evolution of machine learning in support of accurate and efficient agricultural decision-making.

Pawar P. et al. (2023) included most of the machine learning and deep learning methods used for crop yield prediction and recommendation systems. Traditional models based on historical data and statistical models alone cannot capture complex, dynamic factors associated with agriculture. This review further emphasizes advanced models like Long Short-Term Memory Networks and Deep Reinforcement Learning toward better improvements in the prediction accuracy of those sequential data analyses. Moreover, Random Forest, Decision Trees, and neural networks-based studies proved the significance of both soil and climate factors. Hybrid models, wherein both machine learning and deep learning techniques are utilized together, have served well for facilitating crop yield and crop type decision-making accuracy in research. Although it is highly computation-intensive, such a model can easily optimize agricultural productivity and help farmers make decisions in different processes. Zhagparov Z. et al. (2021) comprised several approaches that employ machine learning for yield prediction in agriculture, placing more attention on sophisticated applications of algorithms as utilized internationally. A few of the models described are PRF, LR, and SVM, which have been found to possess significant accuracy in applications such as suitability in soils as well as yield forecasting in agriculture. Articles on the application of neural networks in seed physical property studies are reviewed based on alternative prediction techniques, notably RBNN. All these lead to the possibility that machine learning algorithms might have toward improved crop yield prediction; however, the authors claim limited realization of these prospects in Kazakhstan's agro-sector, so localization of the machine learning system is necessary. Rajkumar N. et al. (2023) identified various approaches used within recent literature on crop yield prediction by utilizing machine learning. In this regard, works like Zhu et al. (2018) have utilized hybrid approaches of data analysis to enhance the quality of yield prediction. Khosla et al. (2019) integrated the methods based on fuzzy logic to estimate the yield based on rainfall and temperature conditions. Tseng et al. (2021) presented hybrid models for assessing agricultural data concerning environmental factors. Alizamir et al. (2019) compared chemical constituents in terms of yield and also highlighted the effects of nutrients. Other studies focused on soil-specific parameters by conducting machine learning techniques for the suitability of crops. Collectively, these studies illustrate how factors such as rainfall, nutrients, and characteristics of the soil have a great influence on crop prediction models, and how machine learning can improve the accuracy of agriculture-driven decisions better.

Pandey S.M. et. al (2021) integrate various applied ML techniques into crop yield prediction systems to concentrate

on enhancing agricultural productivity and decision-making in the agricultural industry. Some of the yield-forecasting techniques reviewed included ANN, SVM, and RF. Major works highlighted are the ML models for soil and crop assessment using historical data, and systems predicting yield affected by climatic conditions like rainfall and temperature. GPS-based data combined with mobile applications provides location-based guidance toward enhanced usability and profitability for the end-users. The Random Forest algorithm is often perceived to provide better accuracy in predictive capabilities. It addresses some developments in ML-driven agricultural models, ultimately offering user-friendly decision-making tools for farmers. This is extremely challenging to update and maintain in relevance and improve user accessibility. Geetha M. et al. (2022) delve into several approaches and techniques for developing crop yield prediction models using ML and DL. Miriyala and Sinha, considering the complexity involved in yield prediction, proposed hybrid DL models with the incorporation of remote sensing approaches. Other research studies were focused on different ML algorithms, particularly Random Forest, CNNs, and LSTM networks. Russello and Nevavuori used CNN with satellite and UAV data to increase the predictive accuracy. LSTM networks, which are aware of their capability to handle time-series data, are used for crop yield application since it has persistence in the memory. In general, the literature showcases a tendency to deploy advanced DL models for accurate and scalable predictions while focusing on improving farmers' decisions and crop productivity in changing environments.

Krishna N.V.S. et al. (2023) explained different machine learning models used in predicting crop yield through a literature review. Their study reflected the accuracy of the Random Forest algorithm through Indian datasets, and it has been suggested to use it for crop prediction since it maintains high accuracy with variables related to yield like rainfall and temperature. Potnuru Sai, similar to this, pointed out the usefulness of Random Forest in handling crop yield analysis since it was able to build decision trees across the different data samples for an optimal response from the system under consideration. Vaishali Patil: Applied Random Forest for crop yield prediction in terms of rainfall and temperature parameters. No algorithmic comparisons. Sachin Deshpande: Theoretical applications of machine learning techniques to forecast; Devadatta has been able to demonstrate the utility of supervised learning applied towards achieving quite accurate crop predictions and proposes Random Forest as a good method in crop yield estimation. This literature argues that Random Forest is a suitable approach for precise crop prediction applications in Indian agriculture. Gadupudi A. et al. (2024) highlighted advancement in crop yield prediction models with machine learning (ML) and deep learning (DL) techniques. The earlier literature made use of ML techniques, like Random Forests, Decision Trees, and Support Vector Machines (SVM) for yield estimation to support crop decisions. Other research concentrated on deep learning models like LSTM and RNN which proved its capability for handling time-series data effectively under complicated conditions with successful results. In comparison studies, the RNN resulted as the model that was precise with a precision rate of 99.62% in crop yield prediction. These, of course, have their own set of challenges, especially at the primary stages of applying these on a large scale in agriculture: low availability of data and high cost, especially remain to be changed.

Kandan M. et al. (2021) reviewed crop yield prediction models from various data analysis techniques such as data mining, machine learning (ML), and deep learning (DL). The first studies used data mining techniques such as the decision tree to make predictions. However, all these models were unstable. Then ML techniques were brought into the application, notably Random Forest, Support Vector Machines, and K-Nearest Neighbors, to improve the accuracy with applications varying from soil analysis to climate data integration. Other recent approaches include deep reinforcement learning models for yield prediction, increasing flexibility, and improving performance over classical methods. Ensemble models using classifiers have also been considered to achieve high prediction accuracy. This is a very comprehensive review of the use of data-driven models for informed agricultural decisions, and there is also stress on challenges in adapting models to the different environmental variables. Kalimuthu M. et al. (2020) take applications of machine learning in agriculture and focus particularly on methods of crop prediction. This identifies different approaches to crop yield prediction using machine learning methods, including supervised learning algorithms like Naive Bayes and neural networks. Crop yield production forecasting based on the given parameters like soil conditions and climate, besides historical yield data, has been approached through research works published by Arun Kumar et al. and Nithin Singh. Other methods for yield improvement are regression and classification methods as suggested in some literature such as by Arun Kumar et al. 2018; and Medar & Ambekar, 2019. Other works cited include weather factors that are to be forecasted, like rainfall and temperature. The latter is one of the fundamental inputs used in crop prediction systems. This review confirms that proper data integration leads to maximum crop output and supports the idea of machine learning for dealing with present-day issues in agriculture.

Toomula S. and Pelluri S. (2022) focused on the increased application of machine learning in crop yield prediction and more specifically on models developed to improve accuracy, such as Kernel Extreme Learning Machine (KELM). Different techniques are presented by Abbas et al. and Suresh et al. and applied in predictive models; their studies took into consideration soil quality, climate conditions, and historical crop data as some of the factors that influence crop yield prediction. Techniques like random forest, neural networks, and support vector machines are employed to gain better accuracy and efficiency in the prediction tasks. Recently, research, as shown by Pant et al., points toward multiple parameter models in combination with both soil and weather variables, aimed to improve yield prediction. Using such advances, the KELM model used in this paper promises to obtain accurate crop predictions by generalizing over different datasets with data normalization and kernel transformation.

Shedthi S. B. et al. (2022) summarized different machine learning models devised for crop recommendation systems based on soil parameters, and it lists down several key studies employing different techniques of machine learning such as KNN, SVM, Random Forest, Decision Trees, and Ensemble Models to optimize crop selection. For instance, precision agriculture was discussed in Pudumalar et al., identifying suitable crop crops according to soil. SVM-based models, related to soil-classification models, based on their nutrient content were researched by Saranya et al. and Sharavani et al.; with relatively high accuracy. These primarily promote crop yields by matching the selected crops to the soil and environmental conditions. This now implies the potential of adding more data sources, like real-time sensor data and climate variables, to improve future model accuracy. Sharma A.K. et al. (2022) show the approaches that have been taken in IoT-based smart agriculture, especially in crop yield prediction. It establishes the role of machine learning and deep learning models, like CNNs, RNNs, and hybrid models, in developing better accuracy for yield predictions (Sharma & Rajawat, 2022). The paper reviews precision agriculture techniques that allow farmers to make informed decisions about agriculture by gathering environmental and soil data through IoT sensors. The work by Chlingaryan et al. (2018) and Nevavuori et al. (2019) through deep neural networks can well be utilized in enhancing crop management by making yield predictions and health monitoring of the plants with the aid of data coming from images and sensor interfaces. Hybrid models integrating machine learning as well as cloud computing, according to Agarwal et al. (2021), also are promising in the simplification of crop monitoring and management of resources, which can increase efficiency in systems of smart agriculture.

Phatangare S. et al. (2024) discussed numerous machine learning algorithms that have been applied to predict crop yield and price. This study focused on models such as Random Forest, Decision Tree, and CatBoost because they are the most utilized with agricultural complex data. Recent studies indicate that through the combination of IoT and machine learning, as combined in Extra Tree and Random Forest models, the yield can be predicted precisely. Other price prediction methods include other types, such as Naïve Bayes and KNN, which together consider factors like environmental and economic to help farmers decide the best crop to grow and the time to sell (Phatangare et al. 2024). The general agreement made by the literature review is that using machine learning for the optimization of agricultural productivity is demonstrated in these predictive models by how they assist farmers in making better decisions since data-driven information about storage and timing within sales is profitable and sustainable. Sharma A. et al. (2022) approach to predicting crop yields or making crop recommendations through machine learning has given an introduction to all possible kinds. Works such as the Paddy Yield Predictor sought to generalize yield prediction approaches across other crops, having more extensive soil parameters. Others showed modest degrees of reliability, whereby M5-Prime and k-nearest neighbor methods are considered in terms of predictive power. Random Forest scored over 75% accurate, whilst for Sorghum Yield Prediction, CNN scored 74.5%. A good number of models focus on some particular crops or regions, especially Southern India. The review emphasizes providing datasets and building flexible models for improved performance and feasibility over the implementation of crops across regions. From research during this time, Random Forest is much more accurate and suitable for nonlinear data, as seen in several other works.

Researchers Lagrazon P.G.G. et al. (2023) discussed the recommendation models of yielding along the predictions of patterns of Haywire Yield for machine-learning fetches and sorts by scope and performances of models which state that while installing particular methods, Random Forest and LSTM got the highest scores when one is predicting particular quality parameters like temperature and rain rainfall. Support Vector Machines and Gaussian Process Regression seem to thank commercial means for producing small ransom errors. Studies in Bangladesh and the Philippines utilize decision trees, neural networks, and ensemble models to study different crops under varying climate

conditions. Considering that, it is true that for small datasets, relatively the latter sees some level of limitation along the regionality suffers from the complexity control on models. In this respect, GPR seems to be much better balanced between generalization and accuracy. This is exactly where GPR then becomes most appropriate in a complicated agricultural state of affairs. Patel K. and Patel H.B. (2021) showed that supervised algorithms, such as Support Vector Machines, K-Nearest Neighbor, Random Forests, and Artificial Neural Networks, performed well in predicting suitable crops according to soil properties and environmental conditions. This study shows that when comparable investigations implemented any with machine learning at all dedicated to agriculture, most were limited to certain algorithms to highlight the huge demand for some comparative studies contrasting the various methods used to predict yield. The authors conclude that existing studies are focused on some algorithms and, therefore, there is a serious need for comparative studies examining different algorithms in predicting crop yield.

Authors Kumar J.A. et al. (2023). pinpointed the use of machine learning algorithms in crop yield prediction and mentioned Support vector machines, Artificial Neural Networks, and Random-forest algorithms. Developed to provide efficient methods to analyze all available data in agriculture (soil types, climate patterns, and crop types), this method will guarantee improved accuracy for yield prediction. These authors also pointed out the flaws in previous literature that do not contain any extensive comparison of algorithm performance in crop prediction. Above others, it emphasizes a blend of cognition in the machine learning setting with cognition in agriculture to deal with threats such as climate change volatility and case concerns in resource management to tackle the major, significant challenge of food security and sustainability in agricultural practice. Priyadharshini K. et al. (2022) illuminated one approach of deploying random forests and support vector machine models in forecasting crop yield through explorative views over some selected soil parameters and climatic variables. While the deployed system of these SVM and RVM models sought to inform crop yield monitoring through predictive systems, it was mentioned that climatic variables, along with historical data, should be merged for farmers to select higher-yielding crops. The issues related to kernel function retrieval and hyper-parameter tuning in SVM modeling were also addressed along the way. In conclusion, the emphasis was placed on the movement required to develop machine learning as critical to agriculture and climate change.

Researches Suresh N. et al. (2021) reviewed various methods from the literature concerning the use of machine-learning models, mostly using Random Forest, for crop yield estimation. Different models emphasized the fusion of climatic parameters, like temperature and humid status, to help in raising predictive accuracy; other literature cited efforts in the past, wherein machine learning was applied to data sets that were rather broad in vain attempts to improve the production of agriculture. Other topics hang on coverage, like the issue of data acquisition, and if there are true models that can make predictions of crop yields well and reliably, irrespective of variability in agricultural conditions. All in all, the reviewed literature seems to suggest that machine learning approaches, in particular Forest analog, are promising for crop yield predictions, and this is great for helping farmers reach better practices for agriculture in the world today. Padmavathi A. et al. (2024) discussed different machine learning algorithms for crop yield prediction and recommendation systems. There exist many statistical methods that are used for various agricultural data such as predicting soil quality, climate parameter type data, and many more. The most common algorithms used include a Random Forest, a Support Vector Machine, a Neural Network, and so many others. We discussed the applications of these models which can enhance crop productivity and on the other hand it can lead to the loss of the crop which is very essential from a farmer's perspective. We have also discussed the challenges that come along with applications such as noise in data, domain experts not being available, and so on. Literature suggests that machine learning is the key to the agricultural world and will lead to a very useful and profitable sector that will provide food to the whole world.

Gupta R.S. et al. (2024) traced the development in predicting yields in crops and indirectly the vegetation through using machine-learned techniques. There are tens of algorithms used, which apart from many typically machine-learned predictive methods use Manhattan or Karhunen-Loeve or truncated neural networks to predict crop yield with quantities such as soil quality types, weather types, and some former yield values. The authors believe that the use of these models has changed the performance of the forecasts and a tool for decision support for agriculture has to be developed. They all urge caution, however, that the model must be provided with many important layers such as the adaptability of the model to new fields for specific crops, with missing reports of data entry to record the real changes made during the growing season. Of these recommendations, we are seemingly all hopeful of getting their proposed

model and subsequently improving both crop productivity and sustainability of their agriculture. Rai S. et al. (2022) examined the use of machine-learning methodologies in crop yield forecasting that stressed the accuracy of forecasting in agriculture. It covered a variety of regression models that include Linear Regression, Decision Tree Regression, Lasso Regression, and Random Forest. These methods are said to analyze several elements, namely climate factors, soil conditions, and historical yield data. Random forest was discovered to possess better accuracy over other models generally; therefore, it has been favored for yield prediction. It also discusses challenges like data quality and the shortage of sufficient-sized data sets to improve the model reliability. From the literature reviewed, it is obvious that machine learning modalities offered much assistance in decision-making on administration in agriculture, packaging beneficial results to farmers and food security. Sajja G.S. et al. (2021) examined the use of machine-learning methodologies in crop yield forecasting that stressed the accuracy of forecasting in agriculture. It covered a variety of regression models that include Linear Regression, Decision Tree Regression, Lasso Regression, and Random Forest. These methods are said to analyze several elements, namely climate factors, soil conditions, and historical yield data. Random forest was discovered to possess better accuracy over other models generally; therefore, it has been favored for yield prediction. It also discusses challenges like data quality and the shortage of sufficient-sized data sets to improve the model reliability. From the literature reviewed, it is obvious that machine learning modalities offered much assistance in decision-making on administration in agriculture, packaging beneficial results to farmers and food security.

Extensive literature review revealed various machine learning techniques, which include Random Forest, Support Vector Machines, Neural Networks, and hybrid models for crop yield prediction but fails to provide enough practical insights into the implementation, preprocessing of data, architecture of the model, and real-world applications. Conversely, the paper addresses these limitations by coming up with an in-depth methodology including preprocessing of the data, and integrating the model with the real-world application through a friendly user interface. It compares Random Forest and XGBoost models in terms of accuracy and computational efficiency, showing XGBoost to perform better in practice. Moreover, the document bridges the gap between the presentation of theoretical discussions in the literature review by providing a clear system architecture, implementation details, and visualization techniques toward practical and actionable solutions providing a solid framework for improving agricultural planning and decision-making.

3. Proposed Methodology

The methodology proposed in this research work is an ML-driven Crop Yield Prediction model that helps in identifying the yield of different crops in different states, districts, seasons, and areas. For the implementation, researchers have used a dataset from Kaggle [32] which contains attributes like State_Name, District_Name, Crop, Season, Area, and Production. In this model, researchers have implemented two algorithms on the dataset i.e., Random Forest (RF) and XGBoost algorithm. Both algorithms have shown different results and accuracy scores. At last, the researchers have compared the accuracy of both algorithms and XG boost has shown the best accuracy. After the model is trained it is then integrated with Python to create the user input interface.

3.1 Random Forest

Random Forest Random Forest is a machine learning algorithm that is used for both classification and regression tasks. It works on the concept of a decision tree and is made of multiple decision trees. The use of multiple decision trees is to increase the performance and accuracy of the model. In proposed model, researchers have used a Random Forest Regressor to predict the production or yield of the crop. The working of the model is described the flowchart in Figure: 1.

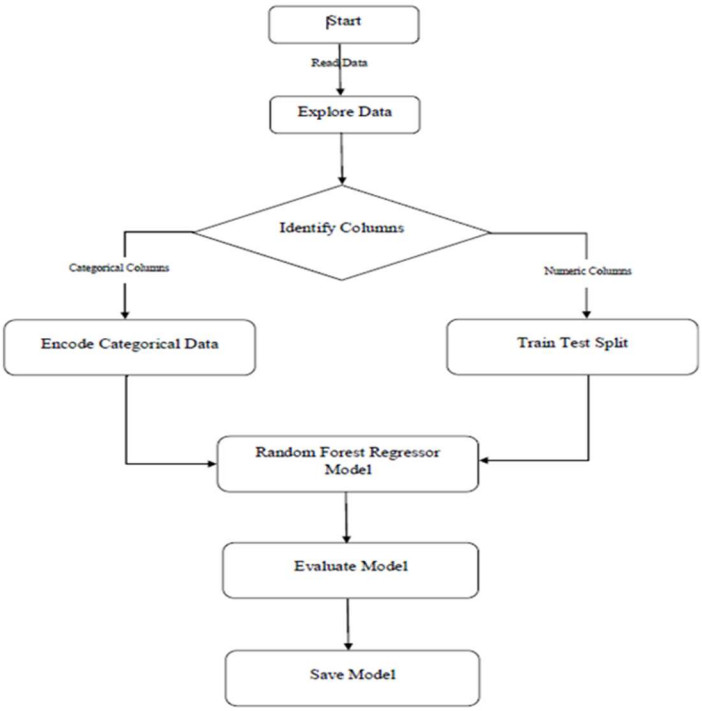

Figure 1. Flowchart - Random Forest Regressor.

Pseudocode for Random Forest:

Input:
- Dataset (D) with (N) data points.
- Number of trees (T), number of features (F).
For each tree (t = 1) to (T):
Sample (N) points with replacement from (D).
Build tree (T_t):
At each node:
Randomly select (F) features.
Choose the best feature and split the node.
Repeat until stopping criteria are met.
End loop.
Prediction for new instance (x'):
For classification: Output the majority vote of the trees' predictions.
For regression: Output the average of the trees' predictions.

3.2 XGBoost

XGBoost Regressor refers to a regression machine-learning algorithm built on gradient boosting that trains decision trees in a sequence to construct a strong prediction model. The algorithm comes with a host of benefits, including parallel processing, a regularizing technique that prevents overfits, and good handling of missing values, etc.; thus, this algorithm is great for big data. Being fast and accurate, easy tuning of its hyper-parameters makes XGBoost one of the frequently modified and popular algorithms for structured data applied in predictive modeling. In the proposed model, working of XGBoost Regressor has been discussed in Figure 2 below:

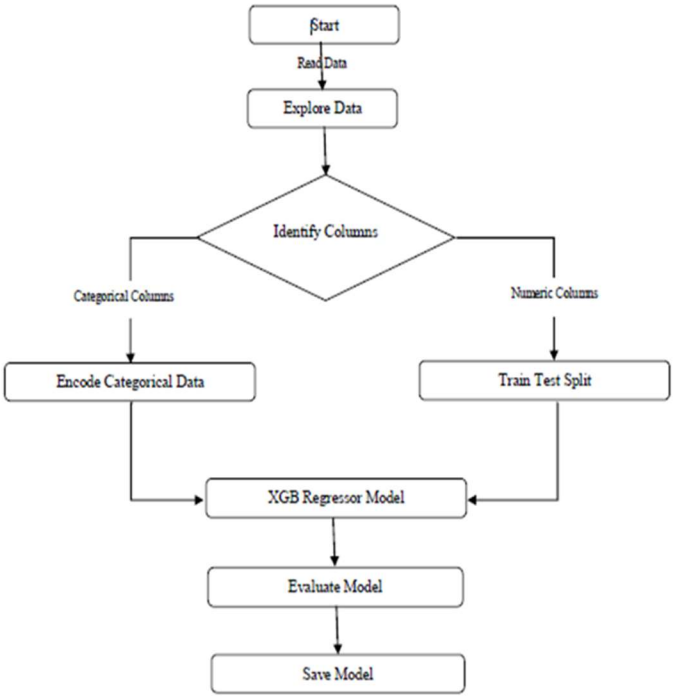

Figure 2. Flowchart - XGBoost Regressor.

Pseudocode for XGBoost:

***Step 1**: Start*
***Step 2**: Read Data*
data = read_data(file path)
***Step 3**: Explore Data*
explore_data(data)
***Step 4**: Identify Categorical and Numeric Columns*
categorical_columns, numeric_columns = identify_columns(data)
***Step 5**: Preprocess Categorical Data*
if categorical_columns:
data[categorical_columns] = encode_categorical_data(data[categorical_columns])
***Step 6**: Train-Test Split for Numeric Data*
X_train, X_test, y_train, y_test = train_test_split(data[numeric_columns], target_variable, test_size=0.2, random_state=42)
***Step 7**: Train XGBoost Regressor Model*
model = XGBRegressor()
model.fit(X_train, y_train)
***Step 8**: Evaluate Model*
predictions = model.predict(X_test)
evaluation_metrics = evaluate_model(y_test, predictions)
***Step 9**: Save Model*
save_model(model, 'xgb_regressor_model.pkl')
End

3.3 System Architecture

The system architecture consists of a data preprocessing module and a user interface for the farmers to input crop data.

The system architecture of the Crop Yield Prediction interface contains various components, starting with a data layer where we input data from a *CSV* file and then pre-process by dividing it into features *(X)* and target variables *(y)*. After that, these categorical features are encoded using an Ordinal encoder. Pre-trained machine learning models *Random Forest Regressor* and an *XGBoost Regressor* are loaded using the *pickle* library for the interface. For the user interface, *Streamlit* is used in this application, which enables users to input data or information such as State, District, Season, Crop, and Area. After the user inputs the data, it is then transformed to match the criteria that are used during the training of the model, and after that predictions are generated using the models. The system displays the predicted yield of the crop using the *XGBoost* model as it has the best accuracy. The result is shown in the *Streamlit* interface. The overall flow of the application embeds data preparation, model interface, and real-time user interaction through the interface as shown in Figure 3.

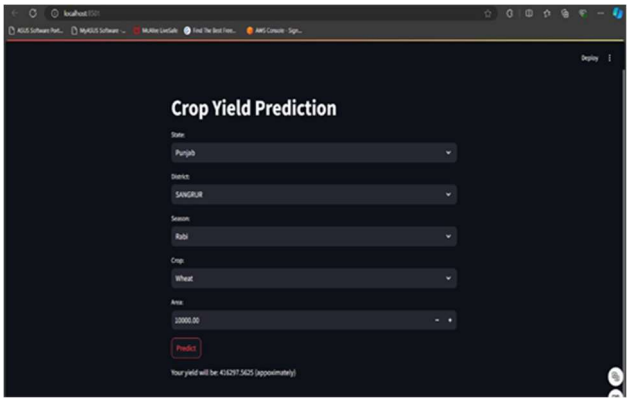

Figure 3: Interface of the Prediction Model

The backend architecture of the system starts with loading the crop data from a *CSV* file taken from Kaggle, followed by the inspection and filtering of data to check missing values and data types. It classifies the features into low-cardinality categorical columns and numeric columns, which are later processed using an Ordinal encoder. After encoding, the data is split into two sets i.e., training and testing sets using ***train_test_split.*** Now, two machine learning models, *Random Forest Regressor* and *XGBoost Regressor* are trained on the given dataset and their performance is evaluated on both training and testing. After that, the trains are saved for future use as a *.pkl* file using the *pickle* library. Outputs can be visualized as follows:

3.3.1 Scatter Graph

The scatter plot in Figure 4 shows the relationship between two variables: 'Crop' and 'Area'. The X-axis represents different crops that are numerically encoded. The Y-axis represents the area associated with each crop. The values are from 0 to 900000. In this graph most of the data points are in the lower region of the Y-axis showing many crops occupy small areas. Some crops between the range of 20-30 on the X-axis have extremely large areas that represent that major crop dominates large agricultural areas. Some areas Several areas, especially below 200,000 on the Y-axis, have dense clusters of data, indicating that many crops share a similar range in terms of area.: There are fewer crops beyond the 50 mark on the X-axis, and their corresponding areas seem lower overall, except for a few spikes around the 60 mark. The distribution of data shows gaps, particularly in the mid-range (30–50 crops), which may indicate missing data or less common crops being grown on large scales.

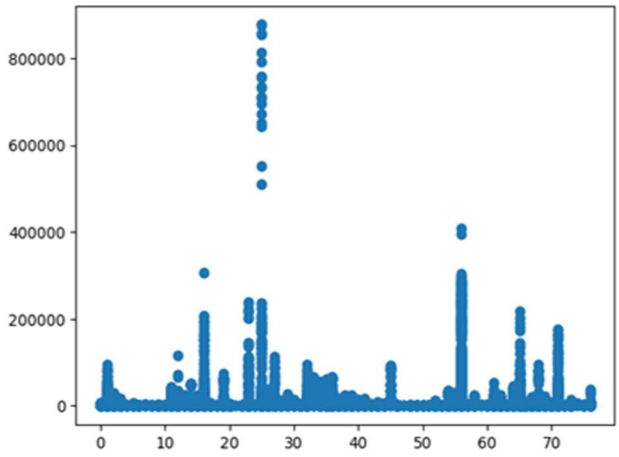

Figure 4: Scatter Graph between Crop & Area

3.3.2 Correlation Matrix Heatmap

Figure 5 below is a correlation matrix heatmap that describes the pairwise relationships of one set of variables in a data set.

Figure 5: Correlation Matrix Heatmap: Showing the pairwise relationships of one set of variables in a data set.

The color scheme on the heatmap graphically depicts the strength and direction of the associations between the variables. All the diagonal elements have a perfect correlation with a value of 1.00 because the correlation of every variable with itself should be perfect. The negative correlation of -0.47 between "State_Name" and "District_Name" shows that the two variables are moderately inversely related in the sense that while one increases, the other tends to decrease. "State_Name" is very weakly correlated to all other variables including "Crop_Year" at 0.02, "Season" at 0.16, "Crop" at 0.09, and "Area" at -0.10, indicating that all these pairs share a very little linear relationship. The "Distrct_Name" variable contains a weak negative correlation with all the other variables other than "Crop" and "Season" for which it's nearly zero, showing that "District_Name" is not strongly linearly related to other variables. The "Crop_Year" has a very weak negative correlation with "Season" (-0.06) and with "Area" (-0.02), and insignificant correlations with either "Crop" or "District_Name." "Season" is positively correlated with "Crop" (0.03) but close to no correlation with "Area" (-0.08), "Crop_Year" or "District_Name." "Crop" has close to no correlation with other variables, which include "Area" (0.03) and "Season" (0.03); therefore, the correlation of "Crop" with other variables

in this dataset is minimal. Very poor correlations with all other variables, bearing a slight negative relationship with "State_Name" at -0.10 and "Season" at -0.08. The matrix generally suggests that most of the variables are not strongly correlated, and most parts of the heatmap are represented as light blue except for the inverse between "State_Name" and "District_Name." The color scale on the right reaches from -1.0 (strong negative correlation) to +1.0 (strong positive correlation) and white means no correlation.

3.3.3 Histogram

This histogram displays the distribution of the variable "Area" for a given data set of the graph as depicted in the figure 6 below:

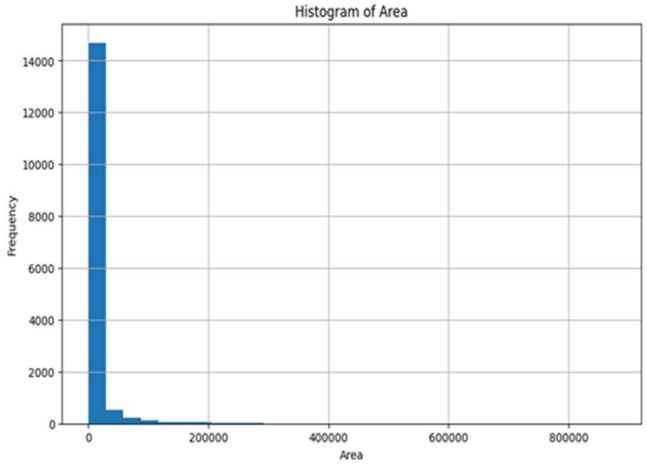

Figure 6: Histogram displaying the distribution of Area.

The horizontal axis of the histogram represents the range of values for "Area" that in effect are bunched together on the left side of the graph. The vertical axis of the histogram is the frequency number of observations that fall in each bin, or range, for the "Area". Most of the values are small or nearly zero. The points of the data are near zero. The histogram of the distribution is positively skewed or right skewed. Since most of the values are small and few of them are relatively so much larger, it makes the right tail of the distribution. The outliers are the few bins that stretch way out to the right on the x-axis but of much fewer frequencies than most. The tail extends right and therefore areas of some of the bins are much larger compared to others but still sparse. The highest frequency bin corresponds to an "Area" value which is very low and reflects that the dataset contains a large number of small areas. There is a sharp decline in frequency as the "Area" is increased and the frequency of large areas is negligibly small as compared to those for smaller areas. This distribution is applied generally for data in which values, like "Area," behave as power-law or exponential decay and many observations fall under fewer categories and fewer large ones. Because this distribution will have a large separation between the peak that occurs near zero and the remaining frequencies that are more to the right, it would be very skewed. Such a dataset might need to be transformed (for example, log-scale) depending upon the analysis that is desired.

3.3.4 Scatter Plot

Figure 7 is a scatter plot that shows the comparison of actual to predicted crop production values. Using this plot, one can analyze model performance. Here, on the x-axis, there are actual production values; on the y-axis, there are predicted production values. Each blue dot represents a data point, for which position is given by actual production along the x-axis and by predicted production along the y-axis.

The red dashed line is the line of perfect prediction since the actual values are matched to the predicted values exactly (that is y = x). The closer points get to this line, the better the predictions are because this means that the value predicted

for the data point is pretty close to the real value of the data point. Points far away from this line mean larger prediction errors.

The scattering of the points away from the red line is a sense of how well the model works. The closer a cluster is to the line, the more accurate the model is. The farther the points are from the line, the lesser the accuracy. From this plot, we can see that most points are lying near the line. It means, roughly speaking, the function behaves pretty well except for some outliers that show deviations, which have appeared in the prediction by the model and the real values. Hence, this type of visualization will be helpful if one wants to rapidly conclude the reliability of the model in crop production predictions.

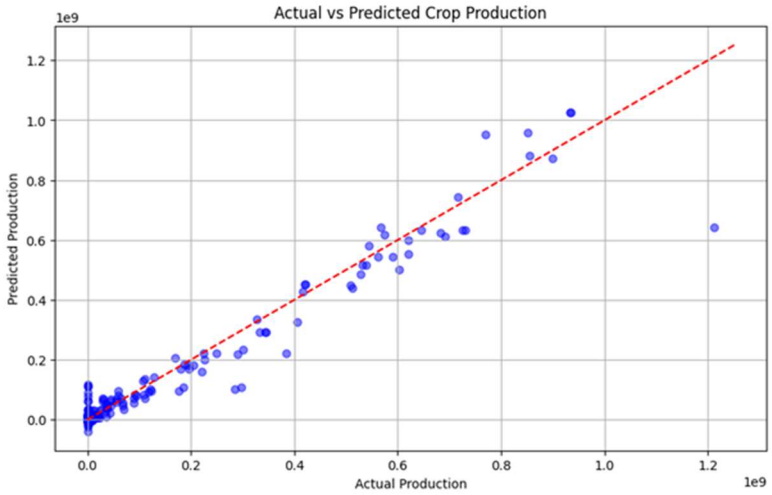

Figure 7: Scatter Plot showing the difference between Actual Crop and Predict Crop

3.3.5 Count Plot of State Name

Figure 8 is a horizontal count plot of the "State_Name" variable. Each bar in the histogram is a name of a state. The length of each bar is equal to the count, or frequency, of that state in the set of data. The state with the largest count is at the top, and the state with the smallest count is at the bottom.

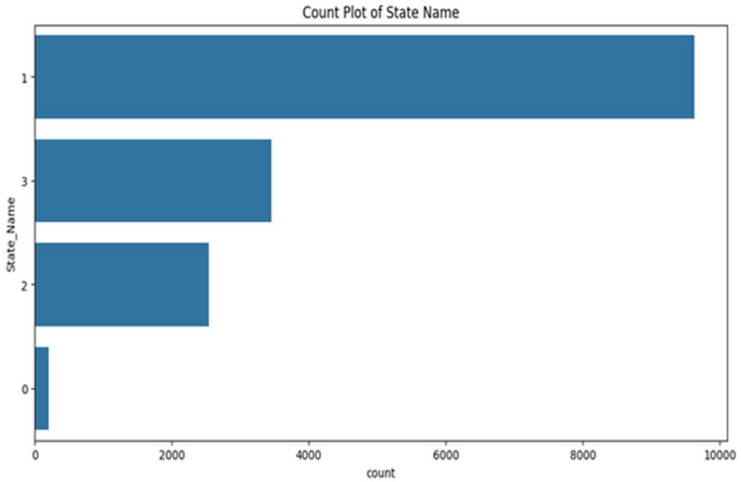

Figure 8: Count Plot Graph of State Name

The y-axis is a list of state names. In this graph, however, the labels appear to be missing or overwritten with number

references (0, 1, 2, 3, etc.). This is likely an error in the labels or perhaps in the plotting function; these need to be corrected to print the actual state names. The longest bar, labeled as "1", is approximately 10 times larger than the rest, implying that this state appears mostly in the dataset, and the count is almost 10,000. This indicates a significant preference for the data set for the given state. Again the second and third states labeled, "3" and "2", respectively, have counts which are quite high but thousands fewer than the top state. Their bars are all much shorter, about 4,000 to 6,000 in count, showing a large drop-off in frequency compared to the top state. The rest of the states (including the one labeled "0") have considerably lower counts, and the smallest bar is so short that it is almost invisible, suggesting that it had a very low number of occurrences. Going by this plot, the state name distribution seems to be very skewed. One state is much better represented in the dataset than any other state. This may raise concerns about bias in sampling or even regional concentration in data collection. This count plot can be very useful in the categorical distribution of the "State_Name" variable, wherein it becomes easy to discern which states occur the most and which are less common. This kind of visualization would be very useful in various decision-making procedures where the frequency of occurrences by state matters- some examples include regional strategies for marketing, resource allocation, or population studies. The graph could easily be enhanced if the actual names of the states were displayed on the y-axis for better interpretability. Including percentages or even an exact count on the bars would give more clarity to the viewer. From this graph, it is very easy to see one state has more frequency than all the others in the data set which can help in understanding the distribution of data in a geographic sense. The imbalance could influence the analysis; therefore, one would be justified in the study of why these certain states are represented much above or below.

3.3.6 Count Plot of District Name

This bar chart would be a horizontal bar chart that demonstrates the occurrence count of different district names found in a certain set of data. The plot is ordered to support viewing by positioning the district with the highest count at the top of the chart and the district with the lowest count at the bottom as depicted in the figure 9 below:

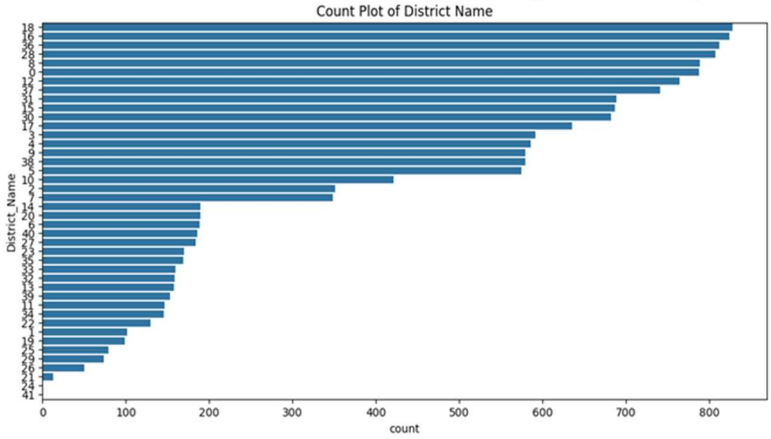

Figure 9: Count Plot Graph of District

The x-axis would represent the "count" for every district that is, how often each district name would appear in the data. The y-axis has the various names of districts listed and ordered in frequency. The district that occurs most frequently goes at the top of the graph, with a count of more than 800, while that at the bottom of the graph has the lowest count with less than 100. This graph also makes it possible for one to easily identify which districts occur frequently in the data and which are rarely found. The count value is the length of each bar, thus easily graphing the number of occurrences of district names for comparison. It appears that one has a count of around 850 and that many had much lower counts in the 100-200 range. It was constructed using Seaborn's count plot function. This automatically generates a count-based bar plot for categorical data. The order parameter plots in the order of frequency rather than the default alphabetical or unsorted order. This kind of plot is highly useful in finding out if there's a trend or a pattern for categorical data. Set the size of the figure to 12x6 so even smaller counts of districts will still fit and be legible.

3.3.7 Count Plot of Season

The following Figure 10 is a horizontal bar plot of counts of occurrence across different categories of "Season" in some datasets. Every bar in this graph corresponds to some season, and the length of the bar is an indication of how often that season occurs in the data.

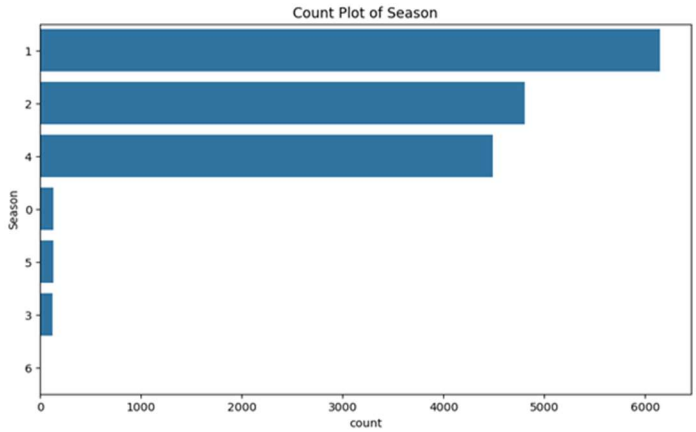

Figure 10: Count Plot of Season

The x-axis plots a count of occurrences; the y-axis shows the names of different seasons. Thus, the most present season in the dataset is Season "1," with more than 6000 occurrences, thus appearing most. Season "2" with close to about 4000 counts, then the near counts of around 3500 counts of Season "4." There are a few other seasons that occur much less frequently. The three seasons "0", "5", and "3" are extremely rare in that their counts are much smaller than the top three. It will indicate the commonness of which season appears most frequently in this dataset and which are rare. I have used the count plot function from Seaborn. It is very suitable for plots of categorical data, and it presents the distribution across categories. In the callings of the function, I've utilized the order parameter to plot the seasons in descending as per their counts. A long bar, as in this case, indicates an important disproportion between the incidence of seasons. In Seasons "1," "2," and "4," most of the data is concentrated. This is a wonderful type of plot to quickly get an intuition for how categories are distributed, say, in categorical variables such as "Season ". The overall positioning and size of Figure 10x6 are such that even the less common season counts are still visible and read better.

3.3.8 Count Plot of Crop

It is a horizontal bar graph (Figure 11) that shows counts of occurrences for some different crop types in a dataset. Along the y-axis are the crops and along the x-axis is the count.

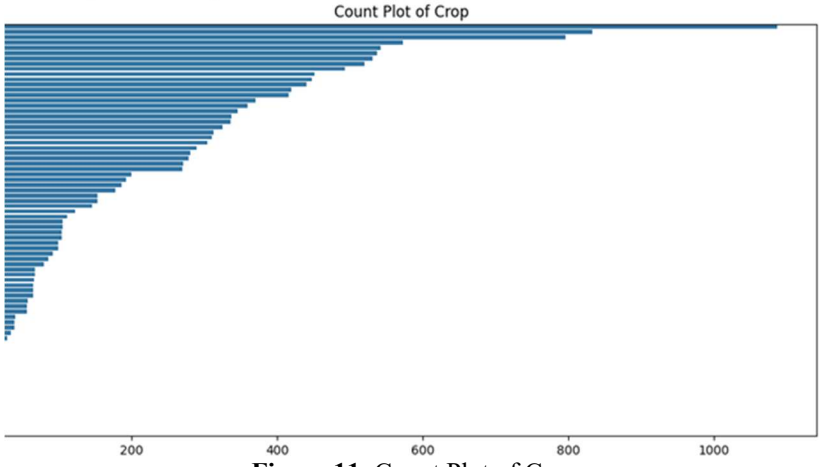

Figure 11: Count Plot of Crop

This graph shows the count of occurrences for the various crop types along the y-axis while the x-axis contains a list of the different crops arranged according to their count of occurrences. The top crop by count is over 1000 counts so that crop is the most representative crop in the dataset. The crops falling down the graph are progressively rarer, and many crops in hundreds of counts of about 100 or less. This plot is generated using the count plot function from Seaborn. The order parameter means that the crops are ordered with the highest-count values, in descending order. The longest bar represents the crop of the highest occurrences and the smallest bars at the bottom are those crops of very low occurrences. The plot reveals an extreme skewing in the distribution as most of the data crops are hugely dominated while only a few crops appear heavily. This type of visualization helps in finding the trends of a crop occurrence and which crop is more dominant in the dataset. This figure size is 12x6 so that one can read even the crops with fewer counts in them so that the overall view of all ranges of crop distribution is clear. For example, this table may be useful for agricultural or environmental studies to identify what kind of crops are planted more frequently in a certain place. Visualization: In the table, information about different crops is presented in such a way that a comparison between different types of crops can be done to come up with a conclusion about which crop occurs the most and which crop occurs the least in this dataset.

3.3.9 Residual Scatter Plot

Figure 12 is the scatter plot of the residuals, that is, the difference between the actual and predicted values, against the predicted values of crop production. The x-axis gives the predicted value for each data point; the y-axis gives the residual, and purple dots represent each of the data points. The red dashed line at zero shows where residuals would be if the model's predictions were perfect. Ideally, these should be randomly distributed around that zero line, which would mean no systematic error in the predictions. These points above the line represent over-predictions, meaning the predicted value is greater than that which has occurred and those points below indicate under-predictions.

This is a clustered residual plot around the zero line, indicating that this model is not heavily biased towards over- or under-prediction. However, there are larger residuals, or outliers, indicating that the model sometimes makes significant errors. This helps evaluate the consistency of the model and if there exist certain patterns within the residuals that could, for instance, indicate model biases or suggest areas for improvement. It would be a random scatter around the zero line, more or less, indicating that the model is well-fit without systematic errors.

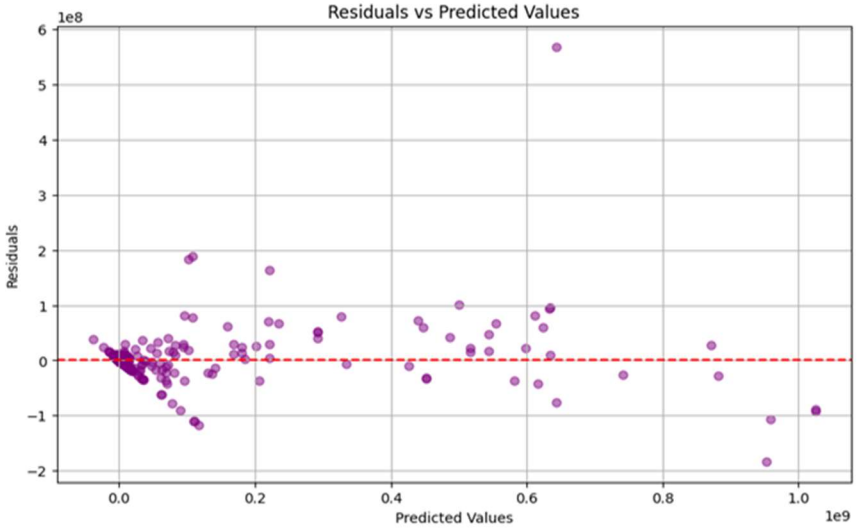

Figure12: Scatter Graph showing Residual vs Predicted Values

4. Results & Discussions

After training the different machine learning models and testing them on the given dataset users can predict the yield of the crop. The difference in the test accuracy between both the models i.e. Random Forest and XGBoost for the crop prediction systems shows that XGBoost exceeds Random Forest as shown in Figure 13. The Random Forest and XGBoost achieve a test accuracy of 87.93% and 95.51% respectively. The difference of approximately 8.09% in

accuracy indicates that XGBoost has a better ability to generalize and capture underlying patterns in crop yield data. This also tells us that XGBoost's higher precision makes it a preferred model for crop yield recognition systems and yield forecasting. This study shows the potential and precision of XGBoost to enhance agricultural planning.

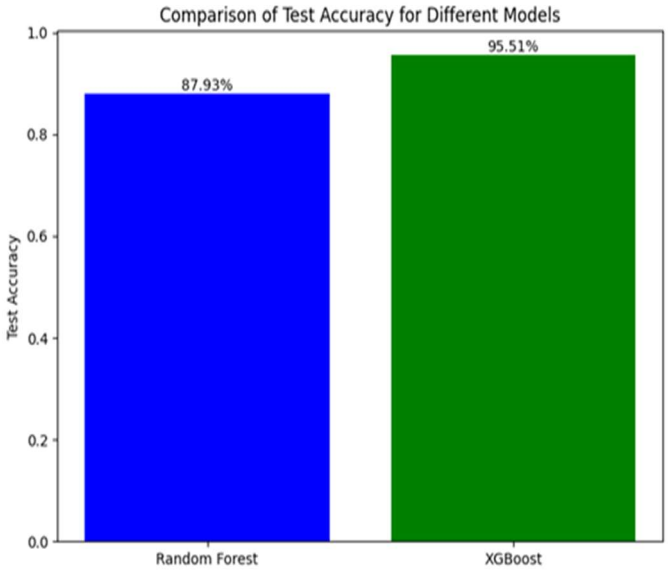

Figure 13: Comparison Between Test Accuracy of Random Forest & XGBoost Model

After integrating the XGBoost model with the prediction system, the system gives the yield prediction of the crop based on the user's input.

5. Conclusion

Proposed work highlights the importance of ML algorithms in the field of agriculture. In this paper, these algorithms have been proven by crop yield prediction systems using Random Forest and XGBoost. As it is predicted that the population of the world will rise beyond nine billion by 2050 so because of it the agriculture sector will face various challenges to meet the food demand. Proposed study shows ML models, especially XGBoost, that have markedly exceeded the traditional statistical method in precisely predicting crop yields by gaining a test accuracy of 95.51 %. The Random Forest has also shown an accuracy of 87.93%. The integration of various data sources such as weather patterns, and soil characteristics helps very much in increasing the prediction power of the models. The study shows the importance of data quality and availability, indicating the need for specific datasets of regions to optimize the model performance.

Moreover, the difference between the analysis of the two algorithms highlights the XGBoost's ability to analyze complex patterns in crop yield data making it a more reliable and trusted tool for agricultural decision making. The user-friendly interface developed in this research helps farmers to use these predictions in real time and helps them to make informed decisions before growing crops. While the research shows the rapid enhancement achieved through ML models, it also discusses the existing challenges, such as data interpretation and the importance of transparency in model operations. In the future, researchers should focus on resolving these challenges, by improving the methods of data collection and developing AI-integrated tools that help farmers traverse the complexities and challenges of agricultural planning.

In conclusion, the knowledge acquired from this research not only contributes to the academic discourse of technology that is used in agriculture but also has practical implementation for the farmers. The study illustrates the crucial and essential role of machine learning in modern agriculture.

6. Future Scope

Future Scope Research Directions of Machine-Learning-Based Crop Yield Prediction Systems: Model Random Forest and XGBoost: The direction wherein the predictive accuracy, usability, and the scope of applications of the models could be taken to higher levels while expanding the research in some of the directions outlined below. The paper focuses mainly on efficiency, and most of the work has been around data integration. Future developments in crop yield prediction systems will scale data sources, integrating real-time information from IoT sensors and advanced remote sensing techniques including SAR and hyperspectral imaging. Insights should be timely and granular with hybrid models combining traditional machine learning variants, such as Random Forest and XGBoost, with deep learning techniques, including CNNs and LSTMs. Improving model interpretability will make use of tools like SHAP and LIME. This improvement increases user understanding of the predictions and, therefore well-informed decision-making. Transfer learning and domain adaptation to various regions and crops strengthen scalability. User-friendly tools like mobile applications can create a system that reaches all farmers. The inclusion of socioeconomic factors as well as factors based on climate change will strengthen the models and maintain their relevance in the changing agriculture landscape. Lastly, addressing data privacy, security, and ethical concerns will be critical for equitable technology adoption. Together, these directions will build robust, adaptable, and ethical systems that are necessary for sustainable agriculture and food security.

References

[1] Ranjani J., Kalaiselvi V.K.G., Sheela A., Deepika S.D. & Janaki G. (2021). Crop Yield Prediction Using Machine Learning Algorithm. *4th International Conference on Computing and Communications Technologies (ICCCT)*, IEEE, pp. 611-616. DOI: 10.1109/ICCCT53315.2021.9711853.

[2] Vanitha K., Priya P. R., Surya G., & Rashmi M. (2024). Enhancing Predictive Accuracy for Agricultural Crop Yields in Indian States Using Power Transformation in Machine Learning Models. *10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, pp. 2403-2408. DOI: 10.1109/ICACCS60874.2024.10716829.

[3] Sunil G. L., Nagaveni V and Shruthi U. (2022). A Review on Prediction of Crop Yield using Machine Learning Techniques. *2022 IEEE Region 10 Symposium (TENSYMP)*, IEEE, DOI: 10.1109/TENSYMP54529.2022.9864482.

[4] Arumuga A. R., Saraswathi T., Meeha D. and Sugeerthi G. (2023). A Machine Learning-based Agricultural Yield Forecasting System for Predicting Crop/Plant Yield before Planting the Crops/Plants. *12th International Conference on Advanced Computing (ICoAC)*, IEEE, DOI: 10.1109/ICoAC59537.2023.10249364.

[5] Kumar T. S., Azeez P., Arunprasad S., Kumar B., Eniyan S. and Sushanth, P. (2023). Crop Selection and Cultivation using Machine Learning. *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, IEEE, DOI: 10.1109/ICCEBS58601.2023.10448940.

[6] Krishna V., Ramar K., Reddy T., Hariharan S., Harsha S. and Prasad B. (2022). Analysis of Crop Yield Prediction using Machine Learning algorithms. *2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pp. 1-5, doi: 10.1109/CISCT55310.2022.10046581.

[7] Devan K. P. K., Swetha B., Uma Sruthi P. & Varshini S. (2023). Crop Yield Prediction and Fertilizer Recommendation System Using Hybrid Machine Learning Algorithms. *12th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 171-175, doi: 10.1109/csnt.2023.33.

[8] Pawar P., Shinde V., Raut A., Suke S., Kolpe K. & Manna A. (2023). An Automated Combined System for Crop Prediction and Yield Prediction using Deep Hybrid Learning Technique. *7th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1-5, doi: 10.1109/ICCUBEA58933.2023.10392068.

[9] Zhagparov Z., Buribayev S., Joldasbayev S., Yerkosova A. and Zhassuzak M. (2021). Building a System for Predicting the Yield of Grain Crops Based On Machine Learning Using the XGBRegressor Algorithm. *IEEE Smart Information Systems and Technologies (SIST)*, Nur-Sultan, Kazakhstan, pp. 1-6, doi: 10.1109/SIST50301.2021.9465938.

[10] Rajkumar N. & Mukunthan M. A. (2023). Efficient Crop Yield Analysis Prediction in Modern Agriculture System using Machine Learning Algorithm. *International Conferences on Data Science, Agents, and Artificial Intelligence (ICDSAAI)*. doi: 10.1109/ICDSAAI59313.2023.10452646.

[11] Pandey S. M., Ramesh P. K., Anmol B. R. A., Rohilla K. & Shaurya K. (2021). Crop Recommender System Using Machine Learning Approach. *Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE. DOI: 10.1109/ICCMC51019.2021.9418351.

[12] Geetha M., Suganthe R. C., Latha R. S., Anju R., Sastimalar K. & Shobana P. (2022). Deep Learning Based Yield Prediction Model to Predict the Yield of Paddy in Cauvery Delta Region. *Proceedings of the 2022 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE. DOI: 10.1109/ICCCI54379.2022.9740944.

[13] Krishna N. V. S., Prudhvi B. V., Neeraj P., Deepthi V. H., & Surya B. (2023). Machine Learning Algorithms for Crop Yield Prediction in Real-Time Scenarios. *4th International Conference on Signal Processing and Communication (ICSPC)*, pp. 377-381, IEEE, https://doi.org/10.1109/ICSPC57692.2023.10126011

[14] Gadupudi A., Rani R. Y., Jayaram B., Sharma N., & Deshmukh J. K. (2024). An Adaptive Deep Learning Model for Crop Yield Prediction. *2nd International Conference on Computer, Communication and Control (IC4),* pp. 1-7, IEEE. https://doi.org/10.1109/IC457434.2024.10486733

[15] Kandan M., Niharika G. S., Lakshmi M. J., Manikanta K., & Bhavith K. (2021). Implementation of Crop Yield Forecasting System based on Climatic and Agricultural Parameters. *International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)*, pp. 207-211, IEEE, https://doi.org/10.1109/ICISSGT52025.2021.00051

[16] Kalimuthu M., Vaishnavi P. and Kishore M. (2020). Crop Prediction Using Machine Learning. *3rd International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, pp. 926-932, doi: 10.1109/ICSSIT48917.2020.9214155.

[17] Toomula S. and Pelluri S. (2022). Design of Kernel Extreme Learning Machine Based Intelligent Crop Yield Prediction Model. *International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 694-701, IEEE, doi: 10.1109/ICACRS55517.2022.10029093.

[18] Shedthi S. B., Anusha A., Shetty R, Alva B. D. and Shetty A. D. (2022). Machine Learning Techniques in Crop Recommendation based on Soil and Crop Yield Prediction System – Review. *International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 230-234, doi: 10.1109/AIDE57180.2022.10078849.

[19] Sharma A. K. and Rajawat A.S. (2022). Crop Yield Prediction using Hybrid Deep Learning Algorithm for Smart Agriculture. *2nd International Conference on Artificial Intelligence and Smart Energy (ICAIS-2022)*, pp. 330-335, IEEE, doi:10.1109/ICAIS53314.2022.9743001.

[20] Phatangare S., Borhade B., Laddha A., Atram P. and Bambal S. (2024). A Data-Driven Approach to Crop Yield and Market Price Prediction. *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2024)*, pp. 805-810, IEEE, doi:10.1109/I-SMAC61858.2024.10714807.

[21] Sharma A., Tamrakar A., Dewasi S. and Naik, N. S. (2022). Early Prediction of Crop Yield in India Using Machine Learning. Region 10 Symposium (TENSYMP), pp. 1-6, IEEE, https://doi.org/10.1109/TENSYMP54529.2022.986490.

[22] Lagrazon P. G. G. and Tan J. B. Jr. (2023). A Comparative Analysis of the Machine Learning Model for Crop Yield Prediction in Quezon Province, Philippines. *12th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 115-120. https://doi.org/10.1109/csnt.2023.24

[23] Patel K. and Patel H. B. (2021). A comparative analysis of supervised machine learning algorithm for agriculture crop prediction. *4th International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1-6. https://doi.org/10.1109/ICECCT52121.2021.9616731

[24] Kumar J. A., Parimala N. and Pitchai R. (2023). Crop Selection and Yield Prediction using Machine Learning Algorithms. *2nd International Conference on Augmented Intelligence and Sustainable Systems (ICAISS 2023)*, IEEE, DOI: 10.1109/ICAISS58487.2023.10250548.

[25] Priyadharshini K., Prabavathi R., Brindha Devi V., Subha P., Mohana Saranya S. and Kiruthika K. (2022). An Enhanced Approach for Crop Yield Prediction System Using Linear Support Vector Machine Model. *International Conference on Communication, Computing and Internet of Things (IC3IoT)*, *Chennai, India*, pp. 1-6. doi: 10.1109/IC3IOT53935.2022.9767994.

[26] Suresh N., Ramesh N.V.K., Inthiyaz S., Poorna Priya P., Nagasowmika K., Kumar H., K.V.N., Shaik M. and Reddy B.N.K. (2021). Crop Yield Prediction Using Random Forest Algorithm. *7th International Conference*

on Advanced Computing & Communication Systems (ICACCS), pp. 279-283, IEEE, DOI: 10.1109/ICACCS51430.2021.9441871.

[27] Padmavathi A., Gupta A. and Prakash K. B. S. (2024). Crop Recommendation and Yield Prediction Using Machine Learning Based Approaches. *5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, pp. 302-309, doi: 10.1109/ICRTCST61793.2024.10578531.

[28] Gupta R., Padmawar T. S., Kumar D., Ray D. and Kadam P. (2024). Significance of Machine Learning in Crop Yield Prediction. *2nd World Conference on Communication & Computing (WCONF), Raipur, India*, pp. 1-7, doi: 10.1109/WCONF61366.2024.10692141.

[29] Rai S., Nandre J. and Kanawade B. R. (2022). A Comparative Analysis of Crop Yield Prediction using Regression. *International Conference on Intelligent Technologies (CONIT)*, pp. 1-6. DOI: 10.1109/CONIT55038.2022.9847783.

[30] Sajja G. S., Jha S. S., Mhamdi H., Naved M., Ray S. and Phasinam K. (2021). An investigation on crop yield prediction using machine learning. *3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021)*, pp. 916-921, IEEE, https://doi.org/10.1109/ICIRCA51532.2021.9544815.