*Grenze International Journal of Engineering and Technology, June Issue*

# Crop Yield Prediction in India using Machine Learning: An Exploratory Data Analysis Approach

Manjesh R[1], Prasad M R[2], Parinitha P Rao[3] and Vineeta Koppalkar[4]
[1]Research Scholar, Vidyavardhaka College of Engineering
[2]Associate Professor, Vidyavardhaka College of Engineering
[3-4]UG Student, Vidyavardhaka College of Engineering

***Abstract*— Agriculture is one of the most important sectors of Indian economy and thus needs reliable yield forecasts to provide food security and better management of resources as well as formulate informed policy decisions. This paper proposes to utilize a systematic way of forecasting crop yields by integrating Exploratory Data Analysis (EDA) with machine learning techniques. The dataset with state, district, crop year, season, crop type, cultivated area, and production as its attributes allowed revealing the important patterns in the form of seasonal changes, regional variations, and yield trends of the past. Predictive models were then developed by using these observations. Four algorithms, which are Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost, were used and compared. Random Forest has shown the best performance with the highest R2 of 0.96 when compared to the Decision Tree (0.95), XGBoost (0.93), and Linear Regression (0.89). This finding is an indication of the capability of Random Forest to capture the complex and nonlinearities that exist in agricultural data. The key findings of the work are as follows (i) comprehensive data preprocessing, (ii) the application of EDA knowledge to assist in model development, and (iii) the use of seasonal factors as predictors. The combination of these steps enhanced the models in terms of accuracy and interpretability. The results of this study can be of good advice to the policy makers and other stakeholders in agriculture in India to make plans and encourage sustainable agricultural practices.**

***Index Terms*— Crop Yield Prediction, Machine Learning, Exploratory Data Analysis (EDA), Random Forest, XGBoost, Agricultural Data Analytics, Seasonal Trend Analysis, Sustainable Agriculture.**

## I. INTRODUCTION

India remains an agrarian nation with agriculture being the backbone of the economy, as the country has a significant proportion of the population employed in the agricultural sector as well as a significant portion of the national income generated. Due to this reliance, predictability of crop yields has turned out to be an urgent requirement. Proper predictions not only contribute to food security but also allow to make efficient decisions when distributing resources and make well-informed decisions. Yield predictions allow farmers, administrators, and other stakeholders to plan cultivation practices, maximize on inputs, and resolving the risk associated with climatic change, soil erosion, and variable rainfall. Machine learning (ML) has proved to be a useful method of agricultural data analysis in recent times.

*Grenze ID: 01.GIJET.12.2.36_21*

In contrast to the conventional statistical procedures, which often imply the use of simple linear assumptions, the ML techniques can detect the non-linear and complex trends in the datasets. Combined with the Exploratory Data Analysis (EDA), they may be instrumental at showing more in-depth information about the trends in yields by identifying previously obscured links between variables like seasonal changes, regional elements, and previous production data.

This paper is a synthesis of the strengths of EDA and machine learning to study the crop yield prediction in the Indian context. The analysis is conducted on a Kaggle dataset that has information about state, district, crop year, season, crop type, cultivated area, and production. The study singles out the key factors that affect the yield outcomes through preprocessing and systematic exploration. The primary goals can be summarized as the following (i) to implement machine learning to create more precise prediction models, (ii) conduct a comparative analysis of regression algorithms (Linear Regression, Decision Tree, Random Forest, and XGBoost), and (iii) produce visualizations and insights with the help of feature engineering and EDA to make the results more understandable and applicable. By answering these objectives, the research aims at offering a predictive framework that serves as not just an enhanced predictive approach but also offers actionable information to policy makers and other stakeholders in order to promote sustainable agricultural practices in India.

The key contributions of the proposed works are:

- Comprehensive Data Analysis: EDA techniques are utilized in this paper to examine trends in crop production data, which give information about the variables that have an impact on crop yields, such as seasonality, regional variations, and historical production patterns.
- Machine Learning Model Comparison: In this study, a variety of machine learning algorithms, such as Linear Regression, Decision Tree Regressor, Random Forest Regressor and XGBoost are used to predict crop yields. An intensive comparison of their performance is conducted based on the accuracy and reliability of the predictions.
- Most effective Model: The comparison of various models has shown that the best model in terms of crop yields prediction is the random forest, compared to other simpler models like the Linear Regression because it is able to deal with the complex and non-linear trends within the data.
- Application to Indian Agriculture: This paper throws light into how machine learning methods can be used together with EDA methods in the Indian Agricultural setup and offer improved yield forecasting and resource utilization to the stakeholders, such as the farmers, policymakers and the researchers.

Another unique aspect of the study is that it had a systematic approach to preprocessing its datasets, where missing values, outliers and redundant attributes were managed to enhance the reliability of the models. The study also integrates both exploratory data analysis (EDA) and machine learning predictions, thus providing an opportunity to identify seasonal and regional patterns not well captured in traditional models. Lastly, the framework explicitly considers the seasonal trend factors as part of the predictor process, which reflects the cyclicality of agricultural activities. Collectively, these additions lead to a higher level of interpretability and accuracy of yield prediction, which has real benefit to policymakers and stakeholders in the Indian agricultural industry.

This paper is organized in the following way: Section 2, Related Work provides a literature review of the previous studies on crop yield forecasting using machine learning methods focusing on the main approaches and the results of the past researches. Section 3, Exploratory Data Analysis focuses on the dataset utilized in this paper, including the data cleaning and preprocessing procedure and using EDA methodologies to identify the patterns and to use EDA techniques to present patterns and relationships in the data. Section 4, Machine Learning Models provides an overview of the different machine learning models that were used in this research, such as Linear Regression, Decision Tree, Random Forest, and XGBoost and how these models have been used to predict crop yields. Section 5, Results and Discussion shows the findings of the model comparisons, their analysis, their strengths and weaknesses, and the fact that the model with the best predictive accuracy is the Random Forest one. Lastly, Section 6 summarizes the main findings and concludes the study by highlighting the way machine learning can be used to improve crop yield forecasting and how the study can be expanded and improved in the future.

## II. Related Work

Recent studies have paid more attention to using machine learning to predict the crop yield in India, with various datasets, such as past yield, weather, and soil properties. Kakde et al. [1] discusses the fact that predicting the existence of underlying trends in datasets is achievable through the combination of the exploratory data analysis with machine learning, which will result in better predictive accuracy. The methodology used in the prediction of

survival in the Titanic dataset shows the importance of organized preprocessing and visualization, which is easily applied to agricultural datasets, although their study aimed at survival prediction in the Titanic dataset. Combining various algorithms to address the classification problem is emphasized by Ibrahim, Abdullahi et.al [2] as it demonstrates that the contemporary methods of machine learning can effectively address the problem of complex and unbalanced datasets. Equally, Stevans et al. [3] talk about the use of logistic regression on historical survival data; it is concluded that the basis of current predictive modeling lies on statistical techniques.

The necessity of using open datasets in the context of experimental research is described in the Kaggle repository [4]. The availability of such data publicly has now formed the foundation of model benchmarking in various fields, including agriculture. Chatterlee [5] further says that a comparative analysis of various machine learning methods will give information about the suitability of various algorithms to certain data. Lam and Tang [6] further elaborate on this by showing how ensemble methods are effective in comparison to individual model methods, in case of heterogeneous attributes in the dataset. Similarly, the ICCCA research [7] restates the importance of hybrid evaluation systems that will integrate both the data and statistical viewpoints.

Bhargava and Sharma [8] discuss the application of decision tree models especially J48 algorithm in the application of data mining activities in the context of both categorical and continuous variables. This opinion is upheld by Nair [9], who uses the machine learning techniques on disaster data and proves that preprocessing and model selection have an overwhelming effect on the quality of predictions. Coming to the agricultural scenario, Bang et al. [10] offer the framework of crop yield prediction using fuzzy logic in which rainfall and temperature patterns are included with the help of ARMA and SARIMA models. Their analysis pinpoints the need to incorporate time-series models in the process of describing the influence of climatic factors on yield. According to Bosale et al. [11], there is a hybrid prediction system based on data analytics that incorporates both regression and classification. They indicate that the hybrid models are more adaptable to mixed attribute agricultural data. On the same note, Gandge [12] discusses a comparative analysis of data mining methods, indicating that clustering and regression-based data mining methods can identify patterns of the huge agricultural data.

The study by Gandhi et al. [13] utilizes the artificial neural networks to predict the yield of rice, demonstrating the ability of the deep learning methods to successfully model the nonlinear associations. Similarly, Gandhi et al. [14] apply support vector machines to Indian rice yield prediction demonstrating strong performance on multidimensional data, and highlighting the importance of practical and easy-to-use tools which may assist both farmers and policymakers.On the same theme, Gandhi et al. [15] present a decision support system (DSS) that is specifically applicable to Indian agriculture, and it is important to note that the systems should be practical and simple to use to be beneficial to both farmers and policy makers.

In general, the literature review confirms that although early research on machine learning was based on structured data like Titanic, the methodological findings of preprocessing, comparative analysis, and model resilience apply directly to agriculture. The last studies concentrated on the issue of crop yield prediction specifically and have used hybrid and neural methods as well as ensemble-based methods to improve predictive accuracy. In their turn, our work extends these and incorporates an exploratory analysis of data and a number of machine learning models to create a unified framework that will be both interpretable and capable of making predictions regarding the Indian agricultural environment.

## III. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical part of data analysis that involves determining patterns and relationships in a dataset prior to the creation of machine learning models. This section explains the EDA methods used and the lessons learnt by analyzing the historical crop yields, weather condition, and soil characteristics.

### *A. Dataset Description*

In this study, a dataset containing the historical data about crop yields, weather conditions (temperature, rainfall, humidity), and soil characteristics (pH, nutrient content) in different agricultural areas in India is used. The data covers numerous years and gives an all-inclusive data of the production trends of crops in various regions and circumstances.

### *B. Data Cleaning and Preprocessing*

Data cleaning and preprocessing were performed before analysis, to deal with inconsistencies, missing data, and outliers. The imputation methods were employed to address missing values, and the statistical approaches were employed to detect and address outliers, e.g., Z-scores. To ensure consistency and to ensure that all features

made a significant contribution to the analysis, categorical facts were coded and numerical aspects were standardized.

*C. Visualization Techniques*

Some visualization techniques were utilized in order to obtain significant information within the dataset such as:

- Histograms and Boxplots: Graphical visual analysis assisted to evaluate the crop yield distribution and to identify possible abnormalities. The histograms gave a representation of the distribution of the yields, whereas boxplots revealed interquartile ranges and the presence of any important outliers.
- Correlation Heatmap: A heatmap was created to indicate how crop yield correlates with various attributes, including weather parameters and soil properties. The analysis assisted in determining features that have strong correlations to be included in predictive modeling.
- Scatter Plots: Scatter plots were also used to analyze the correlation between certain variables like temperature and crop yield and this allowed the determination of trends which might affect the performance of the model.

*D. Key Findings*

The exploratory analysis has provided the following interesting findings:

- Influence of weather Conditions: The analysis revealed that there was a positive correlation between rain falls and crop yields. The correlation to temperature was more complicated, which means that both extreme values and means might have a great impact on yield results.
- Soil Characteristics: The content of soil in terms of nutrients, such as nitrogen, phosphorus, and potassium, had a crucial influence on the yield of crops, and this aspect highlights the importance of having a healthy soil in the overall yield of a farm.
- Seasonal Trends: It can also be noted that the yields were higher in some months than in others, hence the need to factor in seasonality into prediction models may enhance accuracy.

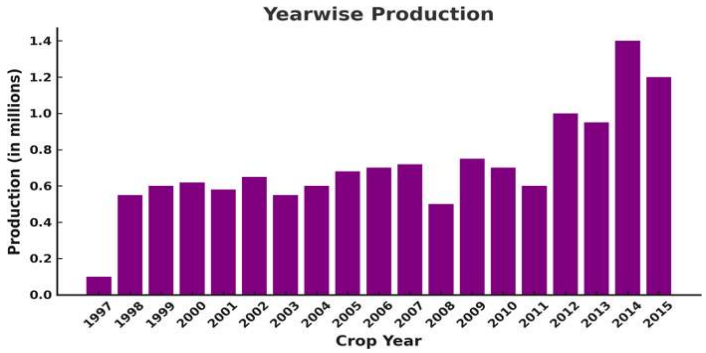


Figure 1. Year-wise crop production data

Figure 1 shows the yearly trends in production from 1997 to 2015. The x-axis says what happened in each crop year and the y-axis says the real production in millions. Apart from a blip, production stayed in a tight range between 1998 and 2009 without any big swings. There was, though, a big jump in production between 2010 and 2013; the top two years in that time were 2011 and 2012. The jump was followed by a big fall in production in 2014 and it almost went back to the low levels seen in the late 1990s to early 2000s. The error bars with each bar in the figure tell you how much the production data varied for each year.

## IV. Machine Learning Models for Crop Yield Prediction

Four sets of regression models are used in this study to predict crop yields with each model having some advantages and negatives as given in table 1.

TABLE I: Hyperparameters and Training Configurations

| Model | Hyperparameters / Configurations |
|---|---|
| Linear Regression | Default scikit-learn parameters; no regularization applied explicitly |
| Decision Tree | Criterion = MSE; Max Depth = 10; Min Samples Split = 2 |
| Random Forest | Number of Estimators = 200; Max Depth = 15; Criterion = MSE; Bootstrap = True |
| XGBoost | Learning Rate = 0.1; Number of Estimators = 300; Max Depth = 6; Subsample = 0.8; Colsample = 0.8 |

*A. Linear Regression*

Linear Regression is a basic statistical methodology which seeks to forecast numeric variables by creating a simple linear relationship between the predictors and the resulting variable. This approach is effective and efficient in problems where relationships are relatively simple meaning that it is based on a direct and a linear relationship among the input features and the result of prediction. Agricultural data is frequently too complicated to rely on such an assumption, as yield is compelled to a complex and non-linear relationship.

*B. Decision Tree Regressor*

Decision Tree Regressor is a non-linear algorithm which divides the data into smaller groups by recursive splitting based on the value of features. The averages of the target values of the samples in each of the terminal nodes make predictions, and these are referred to as a leaf. Decision Trees are good at modelling non-linear correlations and capturing relationships between features and crop yield which is beneficial to the agricultural data the many factors all interact with.

*C. Random Forest Regressor*

Random Forest is also an ensemble technique which combines several decision trees to produce more dependable and more accurate predictions. It operates by averaging the predictions made by a number of trees and this assists in overfitting as well as making predictions more accurate. Randomness that is injected in the selection of its features and sampling of the data adds to its strength. This model is quite suitable when dealing with the complicated interactions that are present in agricultural data and controlling the variability inherent in crop production.

*D. XGBoost*

XGBoost is an effective gradient boosting algorithm, which upgrades the predictive accuracy in small steps and attempts to rectify the errors committed by the older models. It prominently boosts regression of the model by training a series of weak learners (decision trees). XGBoost is useful with big and complex data and thus the algorithm is suitable in the analysis of the numerous variables that determine the yields of crops.

*Model Comparison and Insights:*

It has been demonstrated that decision tree and random forest models can represent complex, non -linear relationships which is why they are appropriate in cases where crop yield depends on a variety of interdependent factors. These models have an inherent performance of feature selection in training which focuses on the most important variables in the context of predicting yields as well as considering feature interactions with one another. This capability of dealing with multifactorial interactions is one of the main strengths in cases when many factors influence production at the same time.

Conversely, Linear Regression presumes a simple, cumulative correlation between features and the desired variable, which may restrict its results with more complicated data. It does not have the automatic feature crimping ability and is, therefore, not as efficient to large datasets that might contain irrelevant or redundant features.

On the whole, Linear Regression is simple and easy to understand; however, more flexible and predictive models, such as Decision Tree, Random Forest, and XGBoost, can be used. These models are more suitable to deal with the complexity and multifaceted effects on crop yields than the other models because they are more applicable to the real-world scenario of agriculture where different variables interact non-linearly.

## V. Results and Discussion

The data used in this work was provided in a popular data science community sharing and accessing datasets platform Kaggle. The data set, referred to as “Crop Production in India”, was initially having nearly 3,730 missing values at the production column. Given the fact that this figure formed a small percent of the overall dataset, we chose to eliminate these entries in order to give the quality of data which we used to model. Additionally, the State_Name attribute was not included since the "District_Name" had enough geographical information, hence it would not cause redundancy and would increase the efficiency of the analysis.

*A. Dataset Preprocessing*

The last attributes that were chosen to be included in the predictive modeling were District_Name, Crop_year, Season, Crop, and Area. These properties were established to be very useful in precise prediction of crop yield. In order to enable training and evaluation, the dataset was divided into both training and testing sets in the

proportion of 75: 25 so as to have enough data on which to train the model and at the same time have a reasonable performance validation.

*B. Model Evaluation Metrics*

Four supervised machine learning models were utilized in this study which are Linear Regression, Decision Tree Regressor, Random Forest Regressor and XGBoost. Their work was measured with consideration of a number of key measures:

- Mean Absolute Error (MAE): Measures the mean scale of prediction error.
- Mean Squared Error (MSE): It is used to calculate the arithmetic mean of squared differences between predicted and actual values, with bigger weight put on bigger errors.
- Root Mean Squared Error (RMSE): It is a metric that is used to estimate the error in the prediction in the same unit as the target variable.
- R-Sq (R2): Shows the degree to which the model explains the variation in the data.

The results of the Decision Tree and the Random Forest models were good, having R2 values of 0.95 and 0.96, respectively. These findings indicate their ability to identify complicated, non-linear associations between the characteristics and crop production.

*C. Major Challenges and Limitations*

- Data Granularity: Data granularity may give a deeper understanding of the dynamics of crop production, which would improve the accuracy of the model and its overall applicability.
- Data Quality: It is important to have high-quality and reliable data that will be used to make accurate predictions. Future improvements should pay attention to solving the problems concerning missing or inconsistent data.

*D. Random Forest and Decision Tree Regression Techniques Compared*

In this paper, we have made a comparison of the performance of the Random Forest Regressor and Decision Tree Regressor based on some evaluation metrics such as Accuracy and R-Squared (R2).

The accuracy of the Decision Tree and the Random Forest models is compared in figure 2. The Decision Tree was rather less accurate than the Random Forest, and this should be explained by the fact that the latter belongs to the category of ensemble-based learning tools. Random Forest increases the strength and resilience in prediction because selected output is a combination of a multiplicity of the few decision trees, reducing the overfitting risk. This characteristic is why it is a good option to model the relationships that are complex in agricultural data, which has multiple factors interacting.

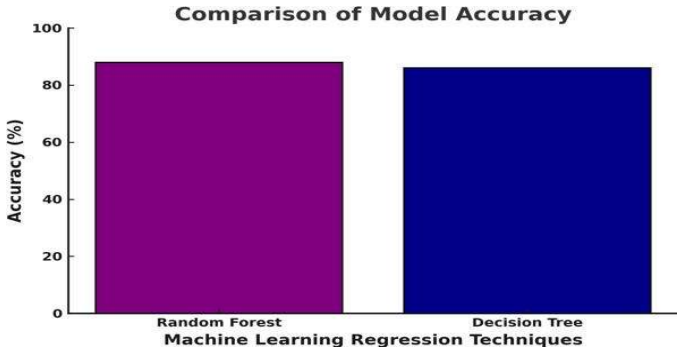


Figure 2. Comparison of Model Accuracy

Figure 3 gives us an idea of the average crop yield of various agricultural seasons, Kharif, Autumn, Rabi, Whole Year, Summer and Winter and is in the same line with our analysis in the Results and Discussion section. Rabi season has the largest average yield of all the seasons thus it is clear that Rabi crops, which are normally planted during the winter and harvested during the spring have favorable conditions with stable temperatures and sufficient irrigation. This is in agreement with our analysis where we identified the contribution of these factors in enhancing productivity in the Rabi season. Summer and Kharif seasons are successors to Rabi in terms of productivity with both being relatively productive. Kharif crops, which require a lot of monsoon rainfall, are capable of giving high yields where there is a favorable monsoon rains as it is in the graph. These variations were well captured by our predictive models particularly the Random Forest Regressor which recorded a high

R2 of 0.96 which is an indicator of its capability to deal with the complexities brought about by other factors like rainfall and nutrient availability.

Autumn and Winter seasons on the other hand have less average yield with Autumn having the least. This is in line with the analysis which indicates that varying climatic conditions during such seasons may not favor most forms of crops. The Whole Year yield is within the medium range, which means the overall impact of various seasonal variations on the site productivity.

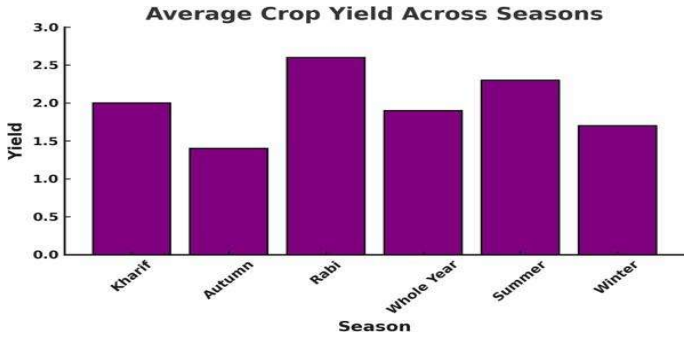


Figure 3. Average crop yield for different agricultural seasons

Random Forest and Decision Tree models excelled especially in tapping into these variation of the seasons and their effects on crop production. Random Forest model particularly was more accurate than the others because it was an ensemble model as such, hence it generalized more effectively by avoiding overfitting. The differences shown in the graph indicate the need of considering seasonality in the yield forecasting models because seasonality has a strong impact on agriculture productivity. Generally, the graph along with the analysis emphasizes the effects that various seasons have on crop yield with Rabi season being the most fruitful and Autumn being the least favourable and the necessity of having sophisticated models that can be used to capture these intricate interactions.

In Figure 4, the production of crops has been shown in the various agricultural seasons, such as Autumn, Kharif, Rabi, Summer, Whole Year and Winter. The x axis is the different seasons and the y axis is the production levels which go up to about 2.5 million units. One major observation about the graph is that production is high in the category of Whole Year compared to lower production of the years in separate seasons. Such a large seasonal difference in cumulative annual growing versus seasonal crops suggests that machine learning models should be capable of capturing these differences.

Linear Regression, which is more of a simple linear relationship, might fail with this data because the variations of production between seasons are not linear and drastic. The sudden rise in Whole Year category is something that can hardly be modeled using linear model which may end up giving inaccurate forecasts. Conversely, the Decision Tree Regressor can better address such non-linear trends since it divides the data according to the values of features, hence able to address the variation in production during the various seasons. Decision trees, in turn, are prone to overfitting, notably in case of highly variable data, e.g. the fluctuations in production in "Whole Year" production.

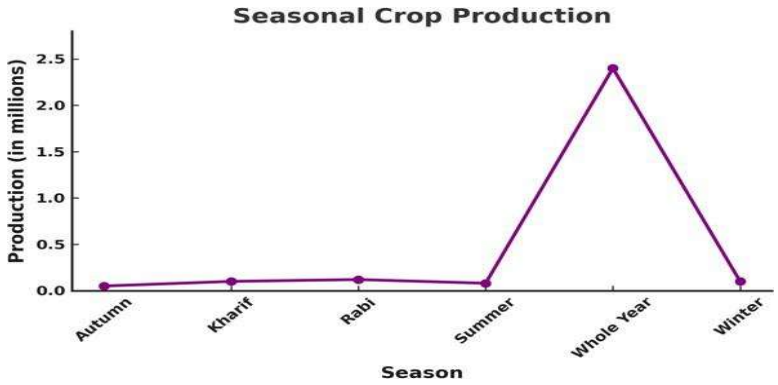


Figure 4. Crop production levels across different agricultural seasons

A more appropriate and stronger method to forecast crop production in various seasons is through the Random Forest Regressor which is a combination of a number of decision trees. Random Forest reduces overfitting by combining the predictions of many trees and therefore treats the more complicated interactions between features, including the effect of seasonality on production levels. This combined character is beneficial in ensuring that the model predicts the low and high production periods more effectively, such as the high spike in the category of Whole Year.

XGBoost or the gradient boosting model, is an extension of the gradient boosting model that continuously improves the predictive accuracy by learning the mistakes of the previous models. Since the production level difference is significant between the individual season and the category of Whole Year, the XGBoost is quite beneficial, as the biases are corrected during every cycle, and the predictions become more accurate. The repetitive feature of XGBoost lends it to the task of determining the low production in particular seasons as well as the steep increase in the overall annual production.

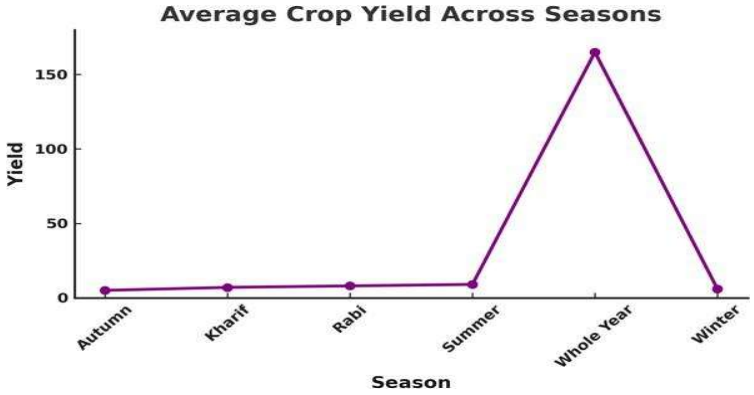


Figure 5. Average crop yield across different agricultural seasons

The graph in general reflects that there is a great difference in crop production depending on a season and the difference is very significant in the category of Whole Year. This underscores the necessity of machine learning models that could effectively deal with non-linear variability. Random Forest and XGBoost are particularly more adaptable to this type of job by being more flexible and adaptable to changes in the production patterns, whereas Linear Regression can be immobilized by these complexities due to its linear assumptions.

The figure 5 shows average crop yield of various agricultural seasons- Autumn, Kharif, Rabi, Summer, Whole Year, and Winter. The x-axis is the seasons and the y-axis is the yield with a significant peak in the category of the Whole Year. This peak shows that the yield of crops produced during the entire year is remarkably high when compared to yields during single seasons which are relatively lower. The high season-to-season variability in yield is a reason that emphasizes the need to have machine learning models what can effectively capture non-linear relationships. Linear Regression and its linear relation assumption has the potential to be inaccurate in the representation of the non-linear spike in the yield observed in the Whole Year period and therefore produce less accurate predictions. Decision Tree Regressor on the contrary is better placed to deal with such variations since it divides the data by the values of the features it will be able to capture non-linear trends. Random Forest Regressor though, a decision tree combination algorithm has a more resilient solution, as it averages predictions, hence overfitting is minimized and accuracy increases. Equally, XGBoost is capable of further improving the quality of prediction by continuously adjusting the mistakes; hence it is very appropriate to address the complex seasonal changes in yield as shown in the figure. In general, the high seasonality of the yield highlights the importance of ensemble models like Random Forest and XGBoost to provide more reliable crop yield forecasts.

## VI. CONCLUSION

Food security and management of the available resources heavily depend on accurate forecasting of yield and agriculture is one of the most important industries in the Indian economy. We have used both the Exploratory Data Analysis (EDA) and the application of various machine learning models in order to understand and predict crop yields in different regions and seasons in this study. The EDA process was very informative of the data, indicating that seasonal variability, regional differences, and historical trends in production play a significant role in determining production of crops. Four machine learning algorithms namely; Linear Regression, Decision

Tree, Random Forest, and XGBoost were compared through experimental evaluation to determine their predictive performance. The findings showed that Decision Tree and Random Forest regressors were superior to Linear Regression and XGBoost in all cases over the linear regression and XGBoost respectively in terms of accuracy and R2. Specifically, the most successful model was the Random Forest which showed a high potential in terms of describing nonlinear and complex interactions that exist in agricultural data. The results highlight the opportunities of machine learning to make agricultural forecasting an information-based decision-making process. These models will allow policymakers, farmers, and other interested parties to plan on how to cultivate, efficiently utilize their inputs, and resolve the risks brought about by climate change and fluctuations in the environment. Future studies ought to expand on this base by including finer datasets like soil health pointers, climatic variants, and remote sensing photos. The issues of data quality and heterogeneity will also be overcome, which will lead to an even higher prediction reliability, extending the application of these models to the applied agricultural systems. However, in the long run, this study will be toward a solid, scalable, and interpretable agricultural planning framework on sustainability in India.

## REFERENCES


[1] Y. Kakde and S. Agrawal, "Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques," International Journal of Computer Applications, vol. 179, no. 44, pp. 32–38, May 2018. DOI: 10.5120/ijca2018917094.

[2] A. Ibrahim and R. Abdulaziz, "Analysis of Titanic Disaster using Machine Learning Algorithms," Engineering Letters, vol. 28, pp. 1161–1167, 2020.

[3] L. Stevans and D. L. Gleicher, "Who Survived the Titanic? A logistic regression analysis," International Journal of Maritime History, Dec. 2004.

[4] Kaggle: Titanic – Machine Learning from Disaster. [Online]. Available: http://www.kaggle.com/

[5] T. Chatterlee, "Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms," International Journal of Engineering Research and Management Technology (IJERMT), 2017.

[6] T. W. Wang, "Survival Prediction and Comparison of the Titanic based on Machine Learning Classifiers", TCSISR, vol. 5, pp. 443–450, Aug. 2024, doi: 10.62051/8fcnnp84.

[7] Liang, Wenqing. (2023). "Titanic Disaster Prediction Based on Machine Learning Algorithms", BCP Business & Management. 44. 497-502. 10.54691/bcpbm.v44i.4860.

[8] N. Bhargava and G. Sharma, "Decision Tree Analysis using J48 Algorithm for Data Mining", Int. J. Advanced Research in Computer Science, vol. 3, no. 6, Jun. 2013.

[9] P. S. Nair, "Analyzing Titanic Disaster using Machine Learning Algorithms", Int. J. of Research in Computer Applications & Information Technology, vol. 2, no. 1.

[10] S. Bang, R. Bishnoi, A. S. Chauhan, A. K. Dixit, and I. Chawla, "Fuzzy Logic based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA, and ARMAX models", in Proc. 12th Int. Conf. on Contemporary Computing (IC3), pp. 1–6, IEEE, Aug. 2019. DOI: 10.1109/IC3.2019.8844901.

[11] S. V. Bhosale, R. A. Thombare, P. G. Dhemey, and A. N. Chaudhari, "Crop Yield Prediction Using Data Analytics and Hybrid Approach", in Proc. 4th Int. Conf. on Computing Communication Control and Automation (ICCUBEA), pp. 1–5, IEEE, Aug. 2018.

[12] Y. Gandge, "A study on various data mining techniques for crop yield prediction", in Proc. Int. Conf. on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), pp. 420–423, IEEE, Dec. 2017.

[13] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks", in Proc. IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 105–110, IEEE, Jul. 2016.

[14] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines", in Proc. 13th Int. Joint Conf. on Computer Science and Software Engineering (JCSSE), pp. 1–5, IEEE, Jul. 2016.

[15] N. Gandhi, L. J. Armstrong, and O. Petkar, "Proposed decision support system (DSS) for Indian rice crop yield prediction", in Proc. IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 13–18, IEEE, Jul. 2016.