Predictive Modelling for Agricultural Crop yield: Synthetic Data Generation and Accuracy Evaluation

Pragati Tiwari[1], Prof. Mansi Kambli[2], Dr Bhakti Palkar [3]

*[1]Student (M. Tech Computer Science & Engineering), K.J. Somaiya college of Engineering, Mumbai, Maharashtra, India.

[2]Assistant Professor Department of Computer Science & Engineering, K.J. Somaiya college of Engineering, Mumbai, Maharashtra, India.

ABSTRACT - Modern farming operations are becoming increasingly reliant on predictive modelling techniques to optimize agricultural crop yields, which are critical for ensuring food security and maximizing agricultural productivity. Accurate crop yield predictions allow farmers, policymakers, and agricultural stakeholders to make informed decisions regarding resource allocation, crop management strategies, and long-term planning, which are essential for sustaining agricultural output and supporting global food systems. In this context, the ability to develop reliable and precise predictive models is of paramount importance. This research introduces an innovative approach to address this need by generating synthetic data specifically tailored for the creation and evaluation of predictive models aimed at estimating agricultural yields. The synthetic data generation process is meticulously designed to incorporate key agronomic variables, including actual and predicted crop yields, crop year, crop type, fertilizer and pesticide usage, and annual rainfall. By integrating statistical techniques with domain expertise, the proposed framework is capable of accurately simulating real-world agricultural conditions, resulting in datasets that are both realistic and versatile for model training and evaluation.

The study focuses on the rigorous evaluation of predictive models, with a particular emphasis on the performance of Random Forest regressors, a widely recognized machine learning technique known for its robustness and accuracy in handling complex datasets. The Random Forest models trained on synthetic data are systematically compared against those trained on real agricultural datasets to assess their predictive accuracy and reliability. The findings from this comparative analysis underscore the effectiveness of the proposed method in generating robust prediction models that can accurately estimate crop yields, thereby providing valuable insights for decision-making in farming practices. The ability to generate high-quality synthetic data and use it effectively in predictive modelling has significant implications for agricultural data science, offering a comprehensive framework for improving the precision and reliability of crop yield predictions. This work not only contributes to the advancement of predictive modelling in agriculture but also provides a powerful tool for decision support, enabling stakeholders to make data-driven decisions that enhance agricultural productivity and sustainability. By addressing the challenges associated with data availability and model accuracy, this research paves the way for future innovations in precision agriculture, where advanced modelling techniques are seamlessly integrated into farming operations to achieve optimal outcomes.

Keywords: ***Predictive modelling, Agricultural crop yield, Machine learning, Random Forest regressor, Decision support***

1.INTRODUCTION

Crop yield prediction is a cornerstone of modern agriculture, playing a crucial role in ensuring efficient resource management, optimizing decision-making processes, and ultimately supporting food security. Accurate yield predictions enable farmers, policymakers, and other stakeholders to make informed decisions about resource allocation, crop management strategies, and agricultural planning, which are vital for maximizing productivity and sustainability. In recent years, the advent of machine learning techniques has shown significant promise in enhancing the precision and reliability of these predictions, offering new ways to analyze complex agricultural data and extract valuable insights. However, the successful application of machine learning in agriculture is not without challenges. Key issues such as the selection of appropriate algorithms, the availability and quality of data, and the ability to generalize models across diverse agricultural conditions must be carefully addressed to harness the full potential of these technologies.

This research aims to tackle these challenges by developing and evaluating predictive models for crop yield estimation, focusing on how to effectively integrate machine learning into agricultural practices. By leveraging advanced algorithms and creating robust models, the study seeks to provide stakeholders with actionable insights that can lead to better decision-making, improved resource management, and more effective crop management strategies. Additionally, the research emphasizes the importance of sustainable farming practices, aiming to enhance crop yield estimation in a way that supports long-term agricultural productivity and environmental stewardship. Through this initiative, the ultimate goal is to contribute to the advancement of precision agriculture, where data-driven approaches are seamlessly integrated into farming operations, leading to more efficient and sustainable outcomes. This work not only addresses the immediate need for accurate crop yield predictions but also sets the stage for future innovations in agricultural data science, paving the way for more resilient and adaptive farming systems.

2.EXISTING SYSTEM

The current systems in agriculture that leverage machine learning algorithms like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes (NB) have been instrumental in transforming traditional farming practices. These systems are designed to provide critical crop cultivation recommendations, thereby assisting farmers in making informed decisions that enhance crop productivity and sustainability. By analyzing a wide range of agricultural data, including climatic conditions, soil nutrient levels, and other environmental factors, these algorithms help identify the most suitable crops for a given region or season.

Support Vector Machine (SVM) is a powerful classification algorithm used in agriculture to predict the best crop choices based on historical and real-time data. It works by finding the optimal boundary that separates different crop classes based on input features like temperature, rainfall, and soil composition. SVM is particularly effective in handling complex, nonlinear relationships within the data, making it a valuable tool for crop prediction in diverse agricultural environments.

K-Nearest Neighbors (KNN), on the other hand, is a simpler, yet effective algorithm that classifies crops by comparing the current conditions of a farm with those of other similar farms in the dataset. By identifying the 'nearest neighbors'—farms with similar climatic and soil conditions—KNN recommends crops that have historically thrived under those conditions. This approach is especially useful in regions where detailed soil and weather data may be limited, as it relies on the principle of similarity to guide crop selection.

Naive Bayes (NB), a probabilistic classifier, is another popular algorithm in existing agricultural systems. It calculates the probability of each crop being the best choice based on the presence of certain features like soil pH levels, moisture content, and historical yield data. Despite its simplicity, Naive Bayes is effective in agricultural settings where the assumption of feature independence holds true, making it a practical tool for farmers seeking quick and reliable crop recommendations.

These existing systems have made significant strides in helping farmers optimize their crop choices by considering factors that were previously difficult to quantify and analyze. By providing data-driven recommendations, these algorithms have enabled farmers to maximize their yields, reduce resource wastage, and adopt more sustainable farming practices. However, while these machine learning models have been successful in offering valuable insights, they are often limited by the quality and quantity of available data, as well as by their ability to generalize across different agricultural conditions. This has led to a growing need for more sophisticated models and methods, such as the use of synthetic data and advanced algorithms like Random Forest, to further enhance the accuracy and applicability of crop yield predictions in diverse farming environments.

2.1 Drawbacks of Existing System

1.Single Factor Focus: Most existing systems consider only one factor either the weather or the soil, when determining crop suitability. This limitation reduces the accuracy and effectiveness of the recommendations.

2.User Friendliness: Despite numerous proposed solutions, there remain unresolved issues in developing a user-friendly application for crop recommendation.

3.Limited Crop Range: The systems do not comprehensively address the need for a wider range of crops that can be cultivated throughout different seasons, limiting their practical utility for farmers.

4.Technology Accessibility: Many systems do not leverage the most accessible technology to provide direct advisory services to even the smallest farmers at the micro-level of their crop plots.

5.Integration and Adaptation: Existing models may not be easily adaptable to different regions without significant adjustments, limiting their broader applicability across various geographic areas.

3. PROPOSED SYSTEM

The proposed system aims to significantly enhance agricultural decision-making by broadening the range of crops that can be cultivated across different seasons, thereby addressing one of the most pressing challenges farmers face—deciding which crop to grow at any given time. Unlike traditional systems that provide crop recommendations based primarily on either climatic factors or soil nutrient levels, this model takes a more comprehensive approach by integrating a wide array of environmental factors, including rainfall, temperature, area, season, and soil type, to deliver more precise and actionable insights. The ultimate goal of this system is not only to recommend the most suitable crops for cultivation but also to maximize crop yield and profitability by offering tailored guidance on the best agricultural practices.

A key innovation of this system is its use of the Random Forest algorithm, a powerful machine learning technique known for its robustness and accuracy in predictive modeling. The system utilizes Random Forest to analyze the input data provided by the user—such as soil type, farming area, and other relevant environmental factors—and to forecast crop yield for various crops under consideration. By doing so, the system can identify the optimal crops to plant based on current and predicted environmental conditions, ensuring that farmers make informed decisions that enhance both productivity and profitability. The system goes a step further by providing detailed recommendations on how to increase crop yield, including the most profitable crops to cultivate in a specific area and the ideal schedule for fertilizer application. By predicting the optimal timing for planting and fertilizer use, the system helps farmers align their practices with the best possible outcomes, thereby improving efficiency and reducing the risk of crop failure. This approach not only takes into account the immediate needs of the crop but also considers market demand, ensuring that farmers are growing crops that are likely to yield the highest returns.

In practice, a user of this system would begin by inputting essential information about their farming area, such as the type of soil and the environmental conditions prevalent in their region. The system then processes this data, leveraging the Random Forest algorithm to predict the expected yield for various crops. Based on this analysis, the system suggests the most lucrative crop to grow in the given conditions and provides a tailored fertilizer application schedule designed to optimize growth and yield. This level of detailed, data-driven guidance empowers farmers to make decisions that are not only scientifically sound but also economically advantageous, leading to more sustainable and profitable farming practices overall. By enhancing the precision and scope of crop recommendations, this proposed system offers a significant improvement over existing models, providing farmers with the tools they need to navigate the complexities of modern agriculture effectively.

3.1 Advantages of Proposed System

1.Informed Crop Selection: The proposed system provides farmers with recommendations on the most profitable crops to cultivate based on environmental factors, helping them make more informed decisions.

2.Increased Crop Yield: By offering instructions on how to increase crop yield and suggesting optimal fertilizer application times, the system helps maximize agricultural productivity.

3.Resource Optimization: The system aims to make the best use of available resources, including water and soil nutrients, which can lead to more sustainable farming practices.

4.Real-Time Analysis: The proposed system includes real-time crop analysis, enabling farmers to make timely and effective decisions about their farming practices.

5.Economic Benefits: By increasing crop yields and optimizing resource use, the system can help improve the economic well-being of farmers, contributing to higher profits and better living standards.

4. LITERATURE SURVEY

Recent research has increasingly focused on leveraging machine learning methodologies for crop yield prediction, highlighting their potential to enhance agricultural decision-making and optimize productivity. Reddy et al. [1] explored the application of machine learning algorithms to predict crop yields by analyzing historical data on crop production, climatic conditions, and soil parameters, demonstrating significant improvements in prediction accuracy. Similarly, Sharma et al. [2] employed a combination of regression and deep learning techniques, integrating climatic data, soil characteristics, and crop management practices, which resulted in highly accurate yield forecasts. Nigam et al. [3] examined the effectiveness of various machine learning algorithms, including Support Vector Machines (SVM) and Random Forest, emphasizing the importance of selecting appropriate features and algorithms tailored to specific crops and regions for reliable predictions. In a study focused on Indian agriculture, Nishant et al. [4] utilized machine learning techniques to analyze the impact of factors such as rainfall, temperature, and soil type on crop yields, demonstrating the technology's potential to assist farmers in optimizing production. Medar et al. [5] further highlighted the complexity of yield prediction in Indian agriculture, advocating for the use of ensemble methods, which combine multiple models to improve prediction accuracy. Aruna Devi et al. [6] developed a machine learning-based system for both crop selection and yield prediction, incorporating features such as soil properties, climatic conditions, and crop characteristics, and demonstrated its effectiveness in aiding farmers' decision-making. Krishna et al. [8] analyzed various machine learning algorithms for crop yield prediction, showcasing how these methodologies can address the challenges posed by regional variability and multiple influencing factors. R. J et al. [7] focused on using machine learning algorithms to predict crop yields by leveraging data on climatic conditions and agricultural practices, highlighting the potential of these tools in improving prediction accuracy. Kavita and Mathur et al. [7] explored crop yield estimation in India using machine learning, demonstrating the effectiveness of these techniques in handling diverse datasets and regional variability. S. V and Padyana et al. [9] applied machine learning to predict crop yields based on geographical and climatic data, emphasizing the importance of location-specific data in achieving accurate predictions. Mondal and Banerjee et al. [10] used deep learning techniques to develop an effective crop prediction model, illustrating the potential of advanced algorithms in enhancing prediction accuracy. Prashant et al. [11] implemented deep learning for crop yield prediction in Indian districts, underscoring the value of these methods in addressing regional agricultural challenges. Kuriakose and Singh et al. [12] employed LSTM deep learning networks for Indian crop yield prediction, showing the advantages of these models in capturing temporal dependencies in agricultural data. Finally, Ajaykumar and Madhavi et al. [13] reviewed crop yield prediction using deep learning and machine learning algorithms, providing insights into the strengths and limitations of various techniques in different agricultural contexts.

5. METHODOLOGY

This project is designed to revolutionize the decision-making process in agriculture by determining the best crops for planting based on a detailed analysis of soil characteristics combined with farmer-provided information. Leveraging a meticulously curated dataset, managed and updated by system administrators, the project integrates both historical agricultural data and synthetically generated data to ensure that the predictions are not only accurate but also adaptable to a wide range of scenarios. The heart of this project lies in its use of advanced predictive modelling techniques, with a particular emphasis on Random Forest regression, a machine learning method renowned for its ability to handle complex, high-dimensional data and produce robust predictions. The project's core functionality revolves around analysing the diverse input variables provided by farmers—such as soil type, climate conditions, and regional specifics—to forecast future crop yields and production trends. This forecasting capability is critical for helping farmers select the most appropriate crops for planting, tailored to their unique environmental conditions. By integrating synthetic data with real historical records, the system enhances the reliability of its predictions, offering insights that are relevant not only for current conditions but also for future agricultural seasons across different states and regions.

Furthermore, the project's model undergoes rigorous validation through various evaluation metrics to assess its performance and reliability. These metrics ensure that the predictive models maintain high accuracy and consistency, making them a dependable tool for agricultural planning. The insights generated by this model go beyond simple crop recommendations; they provide a comprehensive decision-support system that enables farmers to optimize their planting strategies, thereby improving overall agricultural productivity and sustainability. By equipping farmers with precise, data-driven insights, this project contributes to more informed and strategic agricultural practices, ultimately leading to higher yields, better resource utilization, and a more resilient farming industry. This approach not only addresses the immediate needs of farmers but also supports long-term agricultural sustainability by promoting practices that are aligned with both economic and environmental goals. These insights play a crucial role in assisting farmers to make well-informed decisions that directly enhance agricultural productivity and sustainability. By providing data-driven recommendations on crop selection, planting schedules, and resource management, the system empowers farmers to optimize their operations, leading to increased crop yields and more efficient use of resources such as water, fertilizers, and land. The ability to predict future crop yields and production trends enables farmers to plan ahead, reducing the risks associated with unpredictable weather patterns, market fluctuations, and other external factors. This proactive approach not only improves the profitability of farming practices but also promotes sustainable agriculture by encouraging practices that are environmentally friendly and resource-efficient. By adopting these insights, farmers can achieve a balance between maximizing output and maintaining the health of their land for future cultivation, contributing to the long-term sustainability of the agricultural sector as a whole.

5.1 Data Collection

The dataset used for this analysis contains detailed information on the crop Arhar/Tur across multiple years and states in India. It is sourced from Kaggle and includes key variables such as Crop, Crop Year, Season, State, Area, Production, Annual Rainfall, Fertilizer, Pesticide, and Yield. The dataset, in Comma Separated Value (CSV) format, covers the years 2000 to 2020, providing insights into the agricultural patterns and yield for Arhar/Tur in different regions like Assam, Karnataka, Meghalaya, Uttar Pradesh, and Uttarakhand. Each entry in the dataset records the specific year, season, and state, along with the area harvested, production volume, annual rainfall, amount of fertilizer and pesticide used, and the resulting yield. This rich dataset facilitates a comprehensive analysis of the factors influencing the yield and production of Arhar/Tur, enabling the development of predictive models for future yield and production.

Crop	Crop_year	season	state	Area	Production	AnnualRainfall	Fertilizer	Pesticide	Yield
Arhar/Tur	2000	Kharif	Assam	7280	5159	1965.5	714677.6	1892.8	0.714347826
Arhar/Tur	2000	Kharif	Karnataka	582763	263533	1213.3	57209843.71	151518.38	0.492083333
Arhar/Tur	2000	Kharif	Meghalaya	863	641	6258.8	84720.71	224.38	0.78
Arhar/Tur	2001	Kharif	Assam	7236	5125	1824.7	739012.68	1881.36	0.709130435
Arhar/Tur	2001	Kharif	Karnataka	482100	147437	1002.9	49236873	125346	0.37625
Arhar/Tur	2001	Kharif	Meghalaya	818	631	4241	83542.34	212.68	0.793333333
Arhar/Tur	2002	Kharif	Assam	7013	4961	1973.6	663920.71	1753.25	0.710869565

Table 1: filtered data.csv

The Table 1 presents a portion of the dataset utilized in the project to predict the yield of the crop Arhar/Tur using a Linear Regression model. This dataset contains 100 entries, each detailing specific information for different years, states, and seasons. Key features include the crop type (Arhar/Tur), crop year, season (e.g., Kharif), state (e.g., Assam, Karnataka, Meghalaya), area cultivated, total production, annual rainfall, fertilizer and pesticide usage, and the resulting crop yield. These variables are essential for developing a model that predicts crop yield based on historical data, and the table serves as a representative example of the CSV file used in the project.

5.2 Pre-Processing (Null Removal)

Pre-processing is a critical step in ensuring the successful operation of any data-driven application, especially in the field of predictive modelling for agriculture. The data gathered from various sources, such as weather stations, soil sensors, and historical agricultural records, often comes in raw form, and this raw data is rarely clean or ready for immediate use. It typically contains conflicting information, redundancies, or missing and incomplete data, all of which can severely impact the accuracy and reliability of predictive models if not addressed properly. To mitigate these issues, a thorough pre-processing phase is essential, where the data undergoes cleaning, transformation, and normalization to ensure it is suitable for analysis.

One of the most important tasks during this pre-processing phase is handling missing values, particularly in features containing numerical data. Missing data can occur for various reasons, such as sensor malfunctions, data entry errors, or inconsistent data collection practices. If not properly handled, missing values can lead to biased predictions, reduce the model's accuracy, and ultimately undermine the effectiveness of the application.A widely accepted and effective method for managing missing values is to substitute them with the mean (average) of the available data within the same feature. This approach is straightforward yet powerful: by calculating the mean of the non-missing values in a numerical feature, you obtain a representative value that can be used to replace the missing entries. This method assumes that the missing data points are missing at random and that the mean provides a reasonable estimate of what the missing values might have been. By replacing the missing values with the mean, you maintain the integrity of the dataset and ensure that the model has a complete set of inputs for training and evaluation.

This process of null removal through mean substitution helps in smoothing out the dataset, making it more uniform and free from gaps that could lead to erroneous predictions. It also preserves the overall distribution of the data, as the mean is a central measure that reflects the general trend of the feature. By addressing missing data in this manner, the pre-processing phase ensures that the subsequent steps in the modelling process are built on a solid foundation, leading to more accurate, reliable, and robust predictive models that can be effectively used to enhance agricultural decision-making.

5.3 Comparison of Linear Regression and Random Forest Models

A comparative analysis of the Linear Regression and Random Forest models for predicting crop yield based on historical agricultural data. The models were evaluated using various performance metrics, and the results are summarized below.

Linear Regression Model: Linear Regression is a statistical approach that models the relationship between a dependent variable and one or more independent variables using a linear equation. It is a simple and interpretable model but may not capture complex nonlinear relationships in the data.

- R-squared: 0.4139
- Mean Absolute Error (MAE): 0.0819

Random Forest Model: Random Forest is an ensemble learning technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It is effective in handling complex, non-linear relationships between variables.

- Mean Squared Error (MSE): 0.000209
- Root Mean Squared Error (RMSE): 0.0145
- Mean Absolute Error (MAE): 0.0084
- R-squared (R2): 0.9939

The Random Forest model demonstrates significantly better performance compared to the Linear Regression model. It has a much higher R-squared value, indicating a better fit to the data. The lower MAE, MSE, and RMSE values also reflect its superior accuracy in predicting crop yield.

5.4 Crop Forecasting using Random Forest (RF) algorithm.

The Random Forest (RF) algorithm is a highly effective machine learning tool that has gained prominence in crop forecasting due to its ability to handle complex datasets and deliver reliable predictions. The RF algorithm operates by constructing a multitude of decision trees during the training phase, where each tree is built on a random subset of features and samples from the dataset. The final prediction is derived by aggregating the outputs of these individual trees, typically through majority voting for classification tasks or averaging for regression tasks. This ensemble approach helps to mitigate the risk of overfitting, making RF particularly robust and versatile in handling agricultural data, which often involves numerous interdependent variables.

In our study, we employed the RF algorithm to predict crop yield, focusing on key factors that significantly influence agricultural productivity, such as precipitation, perception, temperature, and production. The dataset utilized for training the RF model was comprehensive, incorporating a variety of features including crop type (specifically Arhar/Tur), crop year, season, state, area, production, annual rainfall, fertilizer usage, pesticide usage, and yield. By leveraging this rich dataset, the RF model was able to capture the intricate relationships between these variables, leading to highly accurate predictions of crop yield.The training of the RF model yielded impressive results, as evidenced by the low error rates and high accuracy metrics achieved. We assessed the model's performance using several key evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). These metrics provided a quantitative measure of the model's accuracy, with low MSE and RMSE values indicating that the model's predictions were close to the actual observed values, and a high R² value demonstrating that the model effectively explained the variance in the yield data. The strong performance of the RF model underscores its potential as a reliable tool for crop forecasting.

To further illustrate the practical application of the RF algorithm in agricultural planning, we extended our analysis by generating synthetic data for the state of Assam, projecting future crop yields over the coming years. The synthetic data was designed to simulate realistic agricultural conditions based on historical trends and key environmental factors. The RF model's predictions indicated a positive trend of increasing yields over time, offering valuable insights for future crop management strategies. This forward-looking analysis highlights the utility of RF in not only understanding current agricultural trends but also in anticipating future outcomes, thereby supporting proactive decision-making.

By integrating the RF algorithm into the crop forecasting process, we provide farmers with a powerful tool that delivers accurate yield predictions. These predictions are instrumental in optimizing resource allocation, such as determining the optimal use of fertilizers and pesticides, as well as refining crop management practices to maximize productivity. The ability to forecast crop yields with precision enables farmers to make informed decisions that enhance their agricultural efficiency and productivity, ultimately contributing to improved food security and economic stability. The successful application of the RF algorithm in this study demonstrates its value as a cornerstone of predictive modelling in agriculture, empowering farmers with the insights needed to thrive in an increasingly data-driven agricultural landscape.

6.RESULTS

The results of the predictive modelling indicate significant potential for forecasting future crop yields. The use of Random Forest regression has enabled accurate predictions of crop production for different states and future years. Synthetic data generation has been instrumental in achieving these results, with high reliability confirmed through evaluation metrics such as MSE, RMSE, MAE, and R2. These predictions are valuable for proactive decision-making in agriculture, aiding in efficient resource allocation and risk management. Integrating emerging technologies with these predictive models promises further advancements in agricultural forecasting and sustainability.

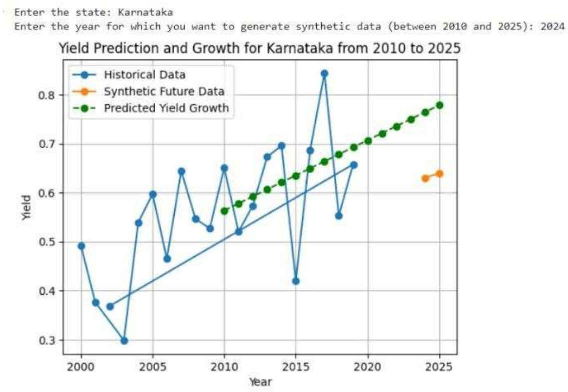
6.1 Linear Regression Model

The linear regression model was employed to predict crop yield based on the provided dataset. The model was trained using features such as crop type, crop year, season, state, area, production, annual rainfall, fertilizer usage, pesticide usage, and yield. The performance of the linear regression model is summarized by the following metrics:

R-squared: 0.4139202308848312

Mean Absolute Error: 0.08193042235172884

The results indicate that the linear regression model explains approximately 41.39% of the variance in the crop yield data. While the error metrics (MAE, MSE, RMSE) are relatively low, there is room for improvement in the model's predictive accuracy.



```
Enter the state: Karnataka
Enter the year for which you want to generate synthetic data (between 2010 and 2025): 2024
```

Fig 1: Prediction of yield using Linear regression model

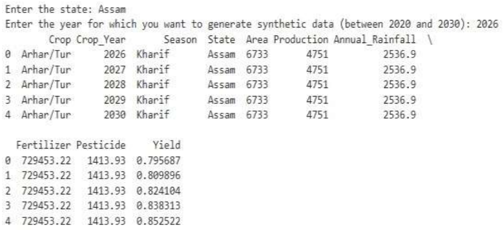
In the Fig 1 above:

- The **blue dots** represent the historical data.
- The **orange dots** represent the synthetic data generated for future years.
- The **green line** indicates the predicted yield growth over time.

The historical data provides a basis for the model, while the synthetic data allows us to project future yields, demonstrating the model's ability to predict trends and growth in crop yield. The green line shows a clear trend, which helps visualize the expected increase or decrease in yield based on the model's predictions.

Overall, the combination of historical data, synthetic data, and predicted yield growth in Figure 1 provides a comprehensive view of the model's capabilities. The historical data underpins the model's training, the synthetic data extends its predictive horizon, and the green line offers a visual representation of expected yield changes. Together, these elements demonstrate the model's effectiveness in forecasting crop yield trends and its potential for supporting informed agricultural decision-making.

6.2 Random Forest Regression Model

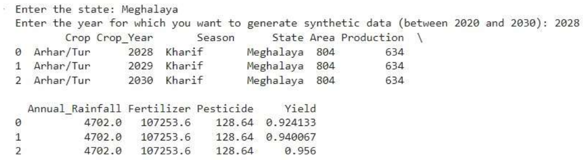


```
Enter the state: Assam
Enter the year for which you want to generate synthetic data (between 2020 and 2030): 2026
       Crop Crop_Year      Season  State  Area Production Annual_Rainfall  \
0  Arhar/Tur      2026  Kharif      Assam  6733       4751          2536.9
1  Arhar/Tur      2027  Kharif      Assam  6733       4751          2536.9
2  Arhar/Tur      2028  Kharif      Assam  6733       4751          2536.9
3  Arhar/Tur      2029  Kharif      Assam  6733       4751          2536.9
4  Arhar/Tur      2030  Kharif      Assam  6733       4751          2536.9

  Fertilizer Pesticide     Yield
0  729453.22   1413.93  0.795687
1  729453.22   1413.93  0.809896
2  729453.22   1413.93  0.824104
3  729453.22   1413.93  0.838313
4  729453.22   1413.93  0.852522
```

Fig 2: Yield prediction using RF Regressor model

In Fig 2, The predicted data for Assam provides an overview of future agricultural performance for the crop Arhar/Tur. Key factors such as area, production, annual rainfall, fertilizer use, and pesticide application are considered, indicating a steady increase in yield over the predicted years.

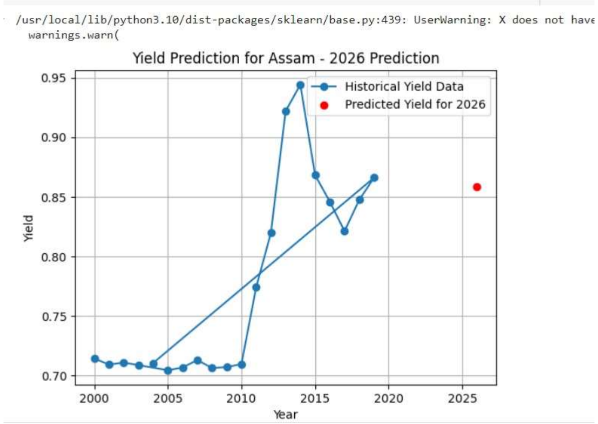


```
Enter the state: Meghalaya
Enter the year for which you want to generate synthetic data (between 2020 and 2030): 2028
       Crop Crop_Year       Season      State Area Production  \
0  Arhar/Tur      2028  Kharif       Meghalaya  804        634
1  Arhar/Tur      2029  Kharif       Meghalaya  804        634
2  Arhar/Tur      2030  Kharif       Meghalaya  804        634

  Annual_Rainfall Fertilizer Pesticide     Yield
0          4702.0   107253.6    128.64  0.924133
1          4702.0   107253.6    128.64  0.940067
2          4702.0   107253.6    128.64     0.956
```

Fig 3: Yield prediction using Random Forest Regressor model

In Fig 3 For Meghalaya, the predicted data also focuses on the future agricultural performance of Arhar/Tur. The data showcases consistent values for area, production, and input factors, with a noticeable increase in yield over the forecasted period.



```
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have
  warnings.warn(
```

Fig 4: Yield Prediction for Asam – 2026 Prediction

Fig 4, The yield prediction graph for Assam illustrates historical yield data in blue and the predicted yield for a future year in red. The forecast suggests a positive trend in yield growth, indicating potential improvements in agricultural output.

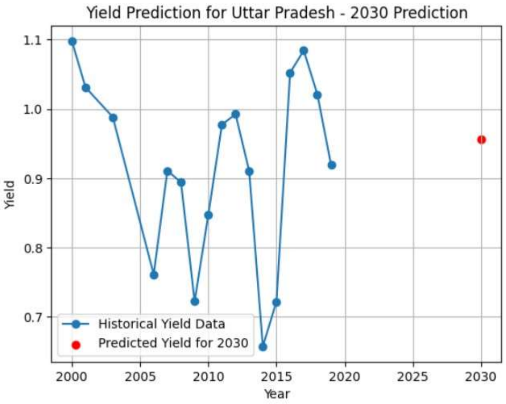


Fig 5: Yield Prediction for Uttar Pradesh – 2030 Prediction

In Uttar Pradesh, the yield prediction graph in Fig 5 shows historical yield data in blue and the predicted yield for a future year in red. The prediction points to a substantial increase in yield, highlighting the effectiveness of the model in forecasting future agricultural performance.

7.CONCLUSIONS

Predictive modelling, particularly using Random Forest regression, provides valuable insights into future crop yield trends. Synthetic data generation techniques facilitate accurate forecasting of crop production for specific states and future years. Evaluation metrics such as MSE, RMSE, MAE, and R2 aid in assessing model performance and prediction reliability. This approach is essential for proactive decision-making in agriculture, enabling better resource allocation and risk mitigation. Collaboration with stakeholders and the integration of emerging technologies offer opportunities for further enhancing predictive modelling solutions in agriculture. By leveraging these advanced techniques and fostering partnerships, the agricultural sector can develop more robust strategies to address challenges and optimize crop production, ultimately contributing to greater food security and sustainability. Furthermore, the incorporation of synthetic data generation techniques has enhanced the model's predictive capabilities, allowing for more robust and reliable predictions. By augmenting the dataset with synthetic data points, we have expanded the model's training set and improved its ability to generalize to unseen data. This approach holds promise for addressing the challenges posed by limited historical data and increasing the model's resilience to fluctuations in agricultural conditions. Overall, the findings of this study underscore the potential of machine learning algorithms, specifically Random Forest, in revolutionizing agricultural practices by providing timely and accurate predictions of crop yields. By empowering farmers and policymakers with actionable insights, these predictive models can contribute to optimizing resource allocation, mitigating risks, and enhancing overall agricultural productivity.

REFERENCES

[1] D. J. Reddy and M. R. Kumar, "Crop Yield Prediction using Machine Learning Algorithm," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2021, pp. 1466-1470, doi:10.1109/ICICCS51141.2021.9432236.

[2] P. Sharma, P. Dadheech, N. Aneja and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," in *IEEE Access*, vol. 11, pp. 111255-111264, 2023, doi:10.1109/ACCESS.2023.3321861

[3] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms," *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2019, pp. 125-130, doi:10.1109/ICIIP47207.2019.8985951.

[4] P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, "Crop Yield Prediction based on Indian Agriculture using Machine Learning," *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, 2020, pp. 1-4, doi:10.1109/INCET49848.2020.9154036.

[5] R. Medar, V. S. Rajpurohit and S. Shweta, "Crop Yield Prediction using Machine Learning Techniques," *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, 2019, pp. 1-5, doi:10.1109/I2CT45611.2019.9033611.
M. Aruna Devi, D. Suresh, D. Jeyakumar, D. Swamydoss and M. Lilly Florence, "Agriculture Crop Selection and Yield Prediction using Machine Learning Algorithms," *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2022, pp. 510-517, doi:10.1109/ICAIS53314.2022.9742846.

[6] R. J, V. K. G. Kalaiselvi, A. Sheela, D. S. D and J. G, "Crop Yield Prediction Using Machine Learning Algorithm," *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, India, 2021, pp. 611-616, doi:10.1109/ICCCT53315.2021.9711853.

[7] M. Kavita and P. Mathur, "Crop Yield Estimation in India Using Machine Learning," *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2020, pp. 220-224, doi:10.1109/ICCCA49541.2020.9250915.

[8] V. Krishna, T. Reddy, S. Harsha, K. Ramar, S. Hariharan and B. A, "Analysis of Crop Yield Prediction using Machine Learning algorithms," *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2022, pp. 1-4, doi:10.1109/CISCT55310.2022.10046581.

[9] S. V and A. Padyana, "Machine Learning based Crop Yield Prediction on Geographical and Climatic Data," *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2021, pp. 186-191, doi:10.1109/ICIIP53038.2021.9702556.

[10] A. Mondal and S. Banerjee, "Effective Crop Prediction Using Deep Learning," *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Pune, India, 2021, pp. 1-6, doi:10.1109/SMARTGENCON51891.2021.9645872.

[11] P. Prashant, K. Ponkshe, C. Garg, I. Pendse and P. Muley, "Crop Yield Prediction of Indian Districts Using Deep Learning," *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2021, pp. 250-255, doi:10.1109/ICIIP53038.2021.9702573.

[12] S. M. Kuriakose and T. Singh, "Indian Crop Yield Prediction using LSTM Deep Learning Networks," *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2022, pp. 1-5, doi:10.1109/ICCCNT54827.2022.9984407.

[13] K. Ajaykumar and S. Madhavi, "Review on Crop Yield Prediction with Deep Learning and Machine Learning Algorithms," *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2022, pp. 903-909, doi:10.1109/ICIRCA54612.2022.9985016.

BIOGRAPHIES OF AUTHORS

	Pragati Tiwari is currently pursuing her M. Tech in Computer Science and Engineering at K.J. Somaiya College of Engineering, Mumbai. She has a strong foundation in Machine Learning, Data Science, Computer Networks, and Data Analysis. Prior to her postgraduate studies, Pragati completed her Bachelor of Engineering from the University of Mumbai. She has already made significant contributions to academic research, having published two research papers. Her first paper, titled Virtual Ally: Campus Navigation System Using Tableau, was published in the International Journal of Computer Engineering in Research Trends (Vol. 9, Issue 03, 2022, E-ISSN: 2349-7084). The second paper, Locker Management System, was published in the Journal of Emerging Technologies and Innovative Research (Vol. 7, Issue 7, July 2020, E-ISSN: 2349-5162), a UGC-approved journal with a 7.95 impact factor calculated by Google Scholar and Semantic Scholar. In addition to her research accomplishments, Pragati has completed various projects in the field of Computer Science and Engineering, showcasing her expertise in the domain.
	Prof. Mansi Manoj Kambli, currently serving as a faculty member in the Department of Computer Engineering at K. J. Somaiya College of Engineering, Vidyavihar, Mumbai, has an extensive teaching career spanning over 20 years and is presently pursuing her PhD. She holds a Master of Engineering (ME) degree from Mumbai University (TSEC, Bandra), completed in 2010 with a 76%, and a Bachelor of Engineering (BE) from K.J. Somaiya College of Engineering, completed in 2001 with a 61%. Her research interests include Remote Sensing, Image Processing, Machine Learning, and the Internet of Things (IoT). Prof. Mansi has authored two significant publications: Physical Computing and IoT Programming (TechMax, 2017) and IoT Technologies (TechKnowledge, 2023). In addition to her academic achievements, she has supervised 4 postgraduate theses and guided 15 undergraduate project groups. Early in her career, she gained practical experience as a Trainee Engineer at Hind Rectifiers Ltd for one year. Prof. Mansi has delivered a wide range of courses, including Image Processing, Machine Learning, Soft Computing, Digital Design, and IoT, further enriching her contribution to the field. She can be reached at prasadinipadwal@somaiy.edu or mansi.mk@somaiya.edu.
	Dr. Bhakti Palkar is a distinguished academic and Associate Professor at K. J. Somaiya College of Engineering, Somaiya Vidyavihar University, Mumbai. She holds a Ph.D. in Computer Engineering and has made significant contributions to the field of medical image processing, geographic information systems, and sentiment analysis. Her research interests lie in the application of image fusion techniques, vector quantization algorithms, and deep learning for various real-world problems, especially in medical imaging. Dr. Palkar has published numerous papers in reputable international journals and conferences, including works on color image segmentation, fusion of multi-modal images, and deep learning techniques for medical image processing. Notable publications include "Image Fusion Techniques: A Review" and research on the fusion of lumbar spine images using Kekre's wavelet transform. She has also explored novel approaches to AI-driven applications, such as automatic colorization of grayscale videos and fraud detection in auto insurance. Dr. Palkar is recognized for her mentorship, guiding students in complex engineering concepts and fostering innovation in her field. Her work continues to influence advancements in both academic research and practical applications in healthcare and technology.