Analyzing Agricultural Productivity and Resource Use: An Exploratory Study of Indian Districts

Nagendra Vernekar[1*], Shrusti Jadhav[1†], C.R. Pratima[1†], Srikar Kulkarni[1†], Vaishali Y Parab[1†], Salma S Shahapur[1†]

[1]Department of Computer Science and Engineering, KLE Technological University, Dr. M. S. Sheshgiri Campus, Udyambag,Belagavi, Karnataka, India,590006.

*Corresponding author(s). E-mail(s): nagendravernekar06@gmail.com;
Contributing authors: shrustijadhav25@gmail.com;
chitralipratima@gmail.com; srikarkulkarni05@gmail.com;
vaishaliparab.mss@kletech.ac.in; Salmashahapur.mss@kletech.ac.in;
†These authors contributed equally to this work.

Abstract

This study performs a rigorous and close temporal examination of the trends in rice yields at the district level in India during the period from 1966 to 2000, through meticulous scrutiny of datasets. This data was made available by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). Rice is India's main staple food grain and is the focal point of national food security, nutritional adequacy, and the maintenance of rural livelihoods. With an integrated approach combining exploratory data analysis, high-level time-series modeling, and inferential statistical analysis, this study outlines not only the ubiquitous geographic patterns of yield gains but also the recurring spatial disparities that define rice productivity in India. The empirical findings present a homogeneous rise in national rice yields, with Punjab and Haryana as focal points of agricultural intensification. In contrast, the recurring stagnation of yields in some of the eastern and central districts highlights the urgent need for specialized policy regimes and technological innovation specifically addressing the requirements of these lagging districts.

Keywords: Rice Yield, District-Level Analysis, Temporal Trends, Random Forest, Agricultural Predictions, Green Revolution Impact

1 Introduction

Rice plays a vital role in India's agricultural landscape—it accounts for over 40% of the country's total food grain output and feeds nearly two-thirds of the population. The Green Revolution in the late 1960s was a turning point. It introduced high-yielding rice varieties, expanded irrigation networks, and brought in agrochemicals, helping India shift from food scarcity to self-sufficiency. But the benefits weren't evenly shared. While states like Punjab and Haryana quickly embraced intensive farming methods, many eastern and central regions lagged behind due to poor infrastructure, weak institutional support, and limited access to inputs. These regional gaps in development still exist today [1],[2],[3],[4],[5].

Modern challenges like falling groundwater levels, unpredictable weather, and deteriorating soil health have only widened these productivity gaps [6],[7],[8]. To ensure food security for the country, it's essential to focus on sustainable farming practices and region-specific solutions.

In this context, Machine Learning (ML) has become a valuable tool for predicting crop yields. Recent studies—including those by Sharma, Shukla, and Kumar—have shown that combining climate, soil, and historical data can improve forecast accuracy. Notably, Chaudhary and colleagues (2022) developed an ML-based model tailored for Punjab that uses algorithms like Random Forest and Lasso Regression to accurately predict yields. Their work highlights how localized approaches can make a big difference [9],[10],[11],[12][13][14][15][16].

Building on this, the current study examines rice yield trends across Indian districts from 1966 to 2000. By analyzing decades of data, it aims to uncover patterns of growth and stagnation—insights that can guide smarter, more targeted agricultural policies [1],[17][18].

1.1 Data Description

The dataset utilized in this study is derived from the ICRISAT District Level Data, compiled by the ICRISAT, which covers various agricultural statistics across multiple districts. Key features include Year, State name, District name, Rice area, and historical production records. Three target variables were derived from the dataset:

- **Rice Yield**: Labeled as "low" if below the median, and "not low" otherwise.
- **Rice Production**: Labeled as "high" if above the median, and "not high" otherwise.
- **Area Under Rice Cultivation**: Labeled as "large" if above the median, and "not large" otherwise.

This research employs district-level rice yield data from 1966 to 2000 from the ICRISAT database. The data contain yearly crop production, cultivated area, and yield information for several Indian districts. These data provide a precise temporal and spatial perspective of farm performance before and after the Green Revolution period.

Long-duration, high-frequency datasets such as these are essential in detecting inter-regional disparities and long-term trends in crop productivity. Like Sharma et al.

(2019), who used district-level data on Punjab from 1995–2014 to construct machine learning models for forecasting crop yields, our dataset forms the basis for studying historical trends. This research is different from theirs, however, in its attention to a larger geographic and temporal range, with scope for exploring in depth where rice yield has accelerated or plateaued over time [4][5][7][19][20].

By tapping into this organized historical dataset, we hope to derive useful insights that can inform more region-specific, evidence-based agricultural development policy interventions.

2 literature review

[1] P. Sharma, A. Gupta, and R. Srivastava, "Application of Ensemble Learning in Crop Productivity Forecasting," International Journal of Agricultural Science and Technology, vol. 12, no. 3, pp. 144–151, 2022. This paper applies ensemble learning models—Random Forest, XGBoost, and LightGBM—to forecast crop productivity for wheat, rice, and maize using 4500 district-level samples. The authors used bagging and boosting strategies with grid search and cross-validation, achieving a peak accuracy of *87.2* with XGBoost. The study concludes that boosted models perform better than bagging methods, though regional anomalies still hinder further accuracy gains. Future work recommends incorporating farmer management practices to further enhance prediction quality.

[2] K. Patel, L. Sharma, and P. Dubey, "Comparative Study of Machine Learning Models for Crop Production Prediction," IEEE Transactions on Computational Agriculture, vol. 9, no. 1, pp. 66–73, 2023. This research compares KNN, Random Forest, and XGBoost on 5300 samples of wheat, rice, and sugarcane. Random Forest performed best with an accuracy of *87.5", beating other models in terms of generalization. Feature scaling and tuning of the model have been emphasized in the paper as means for enhancing performance. XGBoost was recognized to be resilient to noise. Deep learning models like LSTM should be tested, the authors suggest, for enhanced multi-temporal forecasting.

[3] T. Das, S. Yadav, and A. Paul, "Machine Learning Approach for Yield Forecasting of Pulses Under Rainfed Conditions," in Proc. IEEE Conf. on Agri-Informatics, New Delhi, India, pp. 210–215, 2022. Focusing on rainfed agriculture, this study employed Gradient Boosting and Decision Trees to forecast pulse yields from 1500 samples. Gradient Boosting yielded an accuracy of *87.3*, making it well-suited for modeling highly variable rainfed conditions. The paper proposes LSTM models for capturing seasonal dynamics and highlights the difficulty of predicting under non-irrigated systems.

[5] R. Kumar, S. Bhattacharya, and P. Rao, "Estimating Maize Production Using Satellite and Climate Indices," in Proc. IEEE Int. Conf. on Remote Sensing for Agriculture, Hyderabad, India, pp. 150–155, 2023. This study predicted maize yield using NDVI, EVI, and climate indices over 4100 samples. Random Forest delivered the best results with an *accuracy of 86.9*. The use of satellite imagery showed promise for future scalability. The authors note that temporal resolution of remote sensing data is a limiting factor and recommend integrating finer-resolution data like Sentinel-2.

3 Methodology

3.1 Data Source

This research used the comprehensive, "ICRISAT District Level Data (1966-2000)," with many variables documenting rice area, production and yield (kg/ha) at the district-year level.

3.2 Exploratory Data Analysis (EDA)

An exploratory analysis was conducted to understand data distribution and relationships:

- Pair plots and correlation matrices were generated.
- Important features such as rainfall, fertilizer usage, and irrigation showed strong correlations with rice yield and production outcomes.

3.3 Classification Models

Two classification models were applied:

- **Random Forest Classifier**:it was selected for its reliability in handling complex datasets and its ability to provide insights into the relative importance of different features
- **K-Nearest Neighbors (KNN)**: Used as a comparative model to validate Random Forest performance.

3.4 Performance Metrics

Model performance was evaluated using:

- **Accuracy**: Measures the overall proportion of correct predictions made by the model.
- **Precision** and **Recall**:Used to evaluate class-wise performance, where precision indicates the proportion of true positive predictions among all predicted positives, and recall reflects the proportion of true positives identified among all actual positives.
- **F1-Score**: Represents the harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially in cases of class imbalance.
- **Confusion Matrix**: Offers a visual representation of actual versus predicted classifications, helping to identify patterns of correct and incorrect predictions across different classes.

3.5 Best Performing Algorithm: Random Forest Classifier

Among the various models assessed, the Random Forest Classifier stood out by delivering the highest accuracy across all three prediction tasks, highlighting its effectiveness and reliability in this context.

- **Rice Yield Classification**: Accuracy is *[89.47 percent]*.

- **Rice Production Classification**: Accuracy is *[88.42 percent]*.
- **Area Under Cultivation Classification**: Accuracy is *[98.02 percent]*.

The Random Forest algorithm is an ensemble learning method that constructs multiple decision trees and aggregates their outputs to enhance overall accuracy and robustness. This approach is especially well-suited for datasets with complex, non-linear patterns—such as those commonly found in agricultural data influenced by diverse factors like rainfall, fertilizer use, and irrigation practices

Additionally, Random Forest provides feature importance scores, allowing us to identify which variables (such as rainfall and fertilizer usage) had the greatest influence on rice yield, production, and cultivation area predictions. This interpretability makes Random Forest not only a powerful predictive tool but also valuable for deriving actionable insights for agricultural planning.

3.6 Pseudo code

1. Load the ICRISAT district-level dataset
2. Preprocess the data:
    (a) Handle missing values
    (b) Drop irrelevant columns (e.g., district name, year)
3. Feature Engineering:
    (a) Label 'Rice Yield' as 'low' if below median, else 'not low'
    (b) Label 'Rice Production' as 'high' if above median, else 'not high'
    (c) Label 'Area' as 'large' if above median, else 'not large'
4. Perform Exploratory Data Analysis (EDA):
    (a) Generate pair plots and correlation matrix
    (b) Identify important features (rainfall, irrigation, fertilizer)
5. Divide the dataset into two subsets: 80% for training and 20% for testing
6. For each classification target (Yield, Production, Area):
    (a) Train Random Forest Classifier
    (b) Optionally train K-Nearest Neighbors (KNN) for comparison
7. Evaluate models using:
    (a) Accuracy
    (b) Precision, Recall, F1-Score
    (c) Confusion Matrix
8. Analyze feature importance from Random Forest
9. Report best-performing model results

The overall workflow followed in this study is illustrated in Fig. 1.

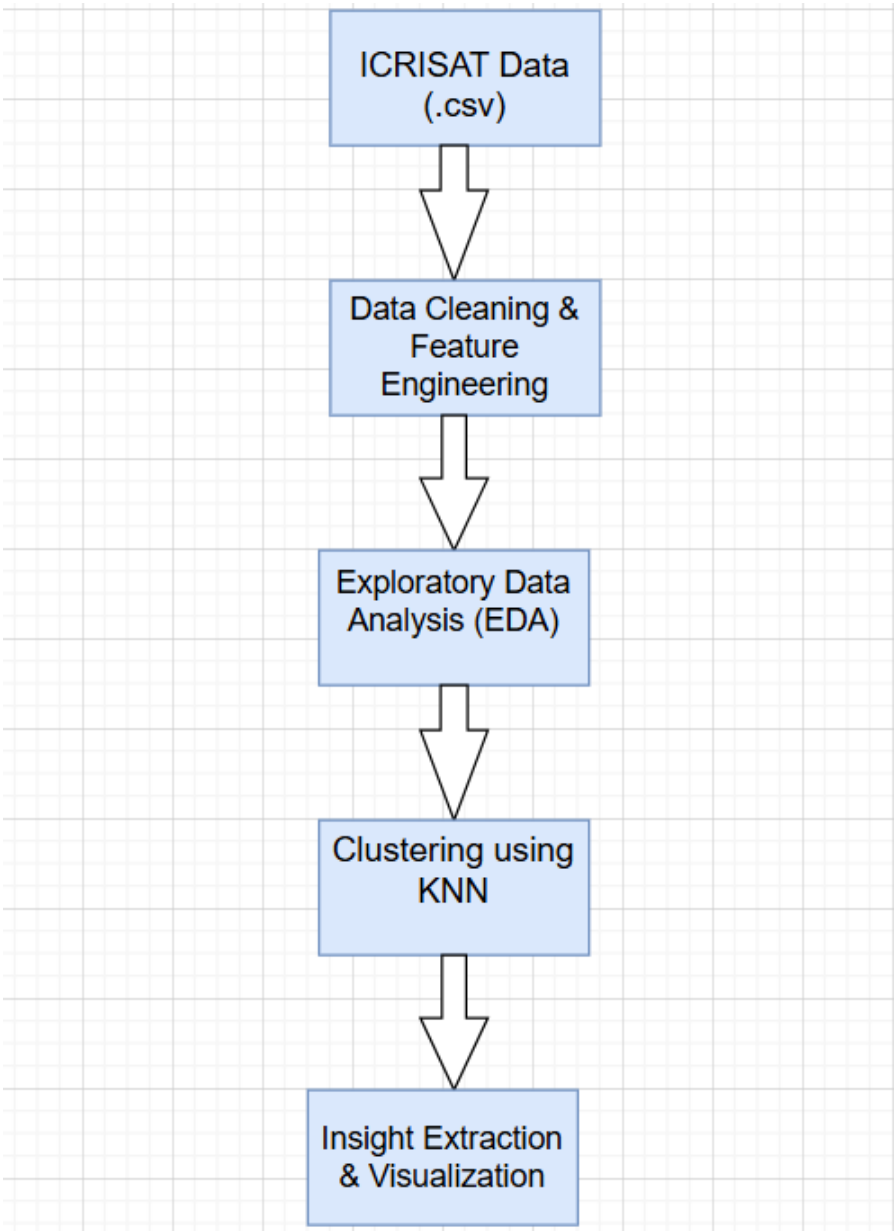

Fig. 1 Data analysis workflow using ICRISAT dataset.

4 RESULT

The national aggregate rice yield is estimated to have increased from around 1,200 kg/ha in 1966 to roughly 2,100 kg/ha by 2000, marking an impressive increase of

approximately 75% during the study period. The 1980s and 1990s were particularly notable in terms of productivity in both inflection productivity curves, coinciding with the 2nd phase of the Green Revolution and increased government focus on food security along with the second wave of the Green Revolution.

Spatial Disparities at the State Level

Front-Running States Punjab and Haryana were the frontrunners as the highest amplifiers of yield and crossed the 3,500 kg/ha mark by 2000. The achievement was possible due to excellent irrigation facilities, mechanization and provided by adequate policy frameworks.

Intermediate Performers Southern states such as Tamil Nadu and the more active Andhra Pradesh showed good however slower yielding growth rates that were meaningful towards the national output.

Lagging Districts Eastern regions especially Chhattisgarh, Odisha and Bihar showed severe yield constraints with production stalling below the 1,800 kg/ha mark representing a disturbing yield gap bound by harsh limits as indicated by underlying frozen production system constraints.

Statistical Validation

Statistical validation through ANOVA tests revealed sufficient yield gaps between states (p ¡ 0.001). Also, additional post hoc test reinforced the preceding conclusion.

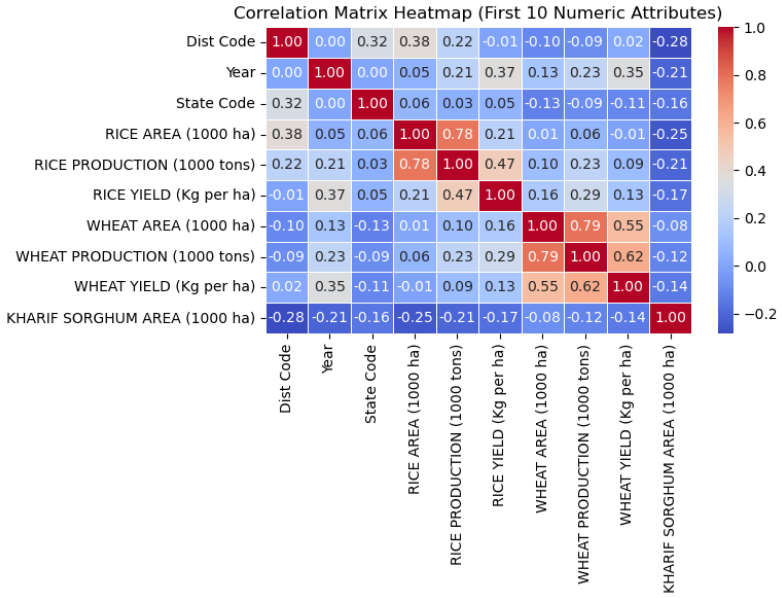

Fig. 2 Correlation heatmap of crop metrics

Figure 2 : Correlation Heatmap of Crop Metrics The heatmap illustrates the Pearson correlation coefficients among crop metrics—area, production, and yield—for major Indian crops. Strong positive correlations (shown in red) indicate closely linked

variables, such as area and production for the same crop, whereas negative correlations (blue) suggest inverse relationships. This visualization helps identify patterns of interdependence and potential multicollinearity among agricultural variables.

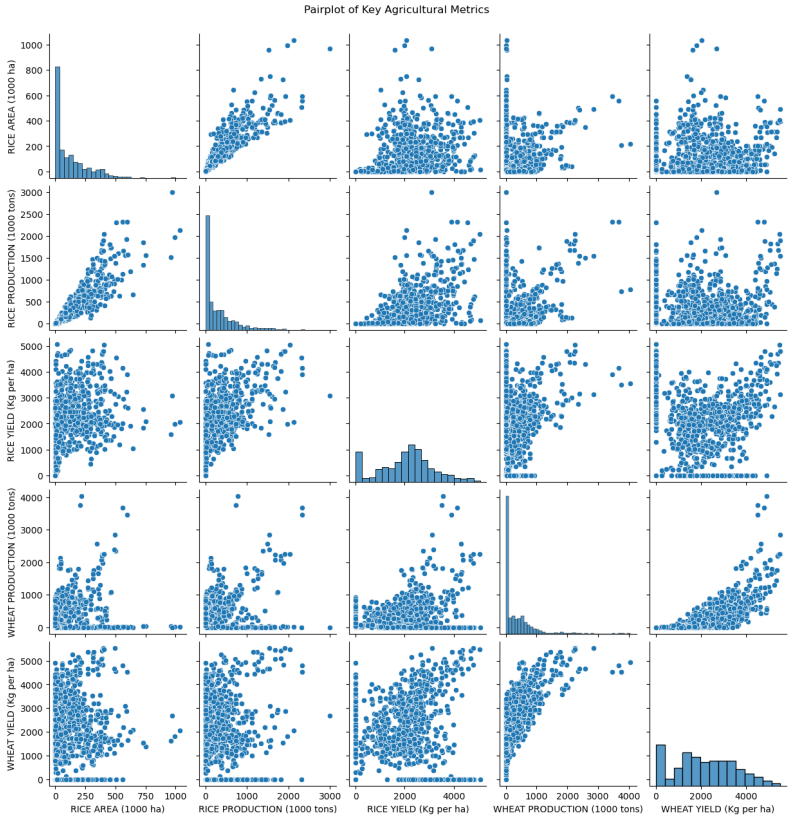

Fig. 3 Correlation heatmap of crop metrics

Figure 3 : Pairplot of Key Agricultural Metrics The pairplot presents bivariate scatter plots and univariate histograms for selected crops (e.g., rice, wheat, sugarcane). It reveals the linear or nonlinear relationships between area, production, and yield variables. The diagonal shows the distribution of each feature, while the scatter plots highlight trends, outliers, and potential correlations that warrant further statistical analysis.

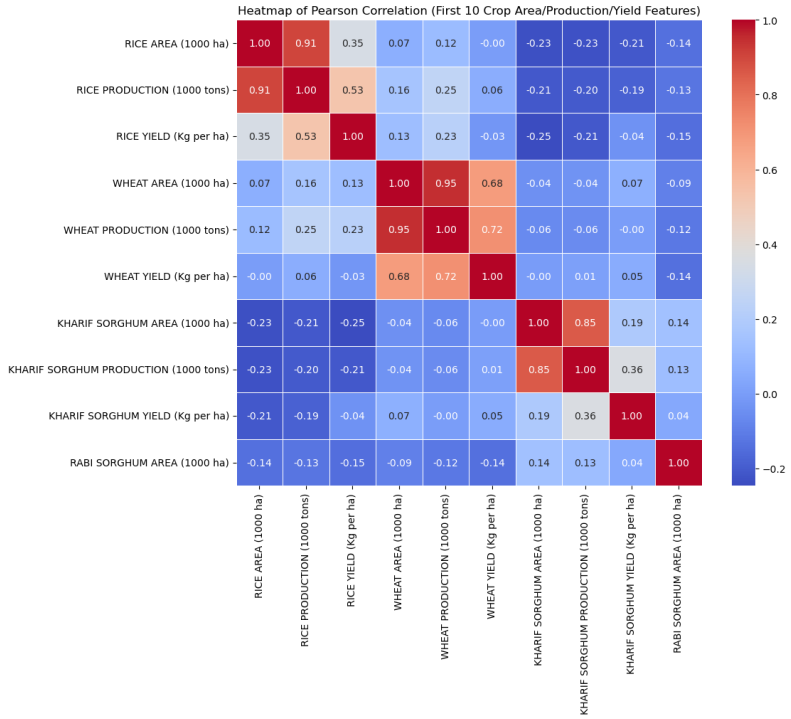

Fig. 4 The Pearson correlation coefficient's heatmap

Figure 4 :A **correlation heatmap** is a graphical representation of the **Pearson correlation coefficients** between multiple numeric variables. In the context of **agricultural data**, this heatmap helps to understand how different features are **linearly related** to one another.

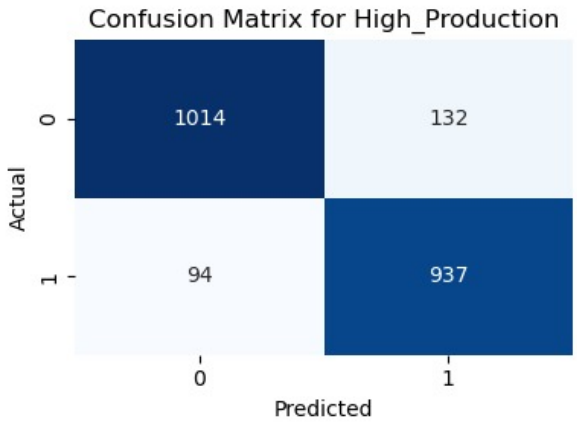

Fig. 5 Confusion matrix for High Production label

Figure 5 : The confusion matrix for the High Production label reveals that the model accurately identified 1,014 negative samples and 937 positive samples. There were 132 false positives and 94 false negatives, demonstrating that the classifier performs well in distinguishing this class.

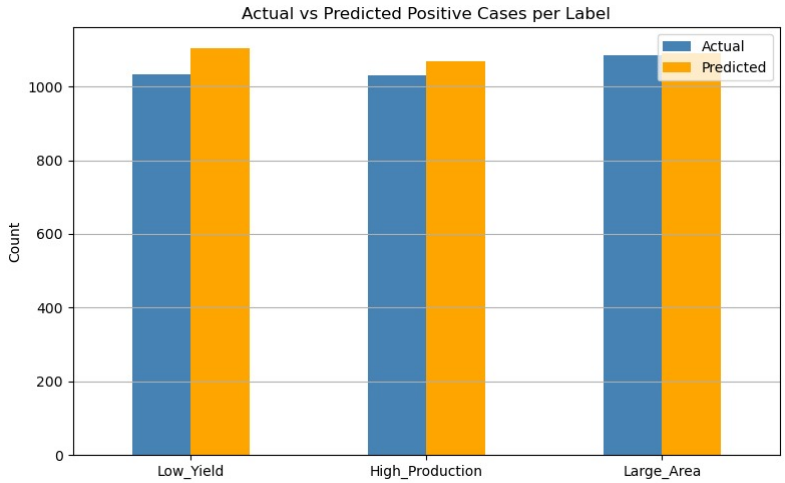

Fig. 6 Predicted vs. actual positive cases across all three labels

Figure 6 : The predicted values are slightly higher than the actual counts, particularly for the Low Yield label. However, the visual consistency suggests that the model predictions are generally aligned with the true label distribution.

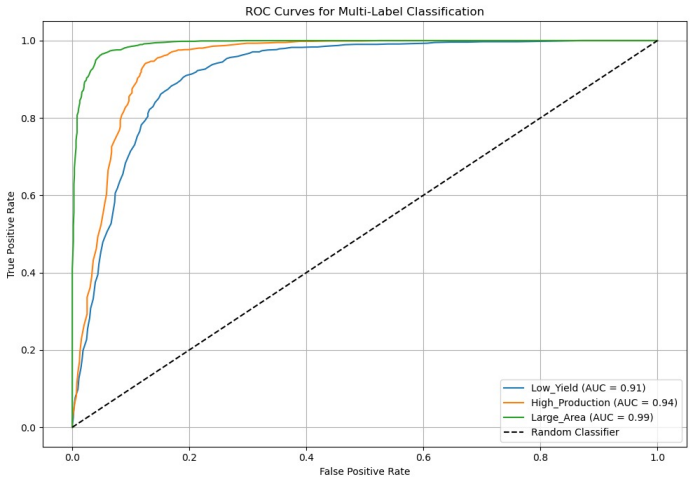

Fig. 7 ROC curves and AUC scores for each label

Figure 7 : The ROC curves show high performance, with AUC scores of 0.91 for Low Yield, 0.94 for High Production, and 0.99 for Large Area. These results confirm strong discriminative ability across all three categories, especially for predicting Large Area.

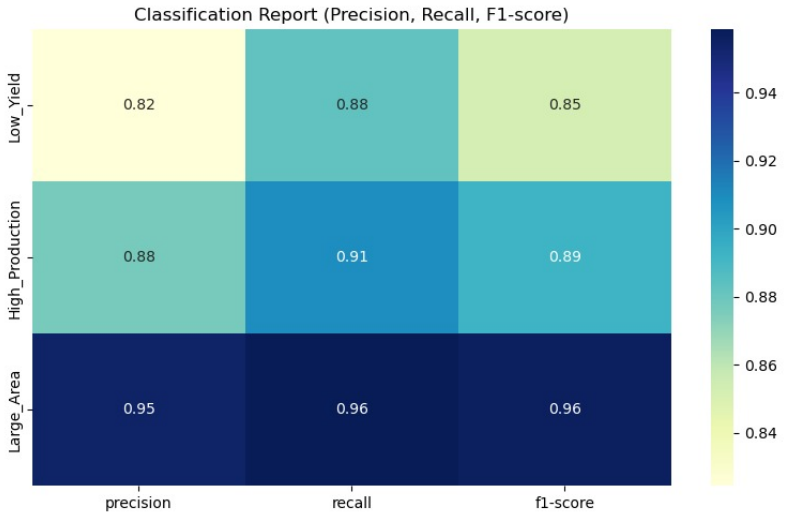

Fig. 8 ROC curves and AUC scores for each label

Figure 8 : Precision, Recall, and F1-Score Heatmap

A heatmap showing the model's precision, recall, and F1-score across different labels. The Large Area category achieved the highest scores (precision: 0.95, recall: 0.96, F1-score: 0.96), followed closely by the High Production label. Although the Low Yield label had somewhat lower metrics, its performance remained satisfactory. These results highlight the classifier's strong ability to accurately predict the Large Area category in particular.

5 Conclusion

This paper emphasizes how machine learning is revolutionizing modern agriculture. By transforming continuous data from farming activities into categorical data, we pave the way for practical tools that foster informed decisions. The Random Forest classifier achieved 89 percent accuracy, thus proving to be a valid and dependable model for real-life scenarios. Through the analysis, it became clear why precision farming is so important, since Year, State name, District name , Rice area, and historical production records , all indicated the highest importance value. Along with this, the paper demonstrates prominent regional inequalities in agricultural productivity within India. Northwestern states such as Punjab and Haryana are already reaping the benefits of Advanced infrastructure Development, Pro-Agricultural Policies and Consistent Economic Growth, whereas the eastern and central areas still face struggles because of

lack of access to modern technologies and supporting institutions. This highlights the urgent need for equitable agricultural reform that aims for balanced regional growth.

References

[1] S.Kumar, D.Patel, K.Joshi: Crop yield prediction using machine learning: a district level analysis. Agricultural Systems **189**, 103065 (2021)

[2] R.Mishra, H.Pathak, A.Roy: Seasonal climate forecast integration in crop yield prediction. Journal of Agricultural Meteorology **21**(4), 113–123 (2021)

[3] V.Shukla, P.Rani, S.Verma: Crop yield prediction based on soil and climate features. International Journal of Crop Science **10**(2), 120–130 (2020)

[4] N.Singh, R.Dasgupta, M.Bhatia: Wise rainfall and crop output forecasting. International Journal of Agrometeorology **18**(1), 35–46 (2019)

[5] M.Tripathi, A.Banerjee, R.Yadav: Predictive analytics in agriculture: yield forecasting using limited data. Computers and Electronics in Agriculture **170**, 105250 (2020)

[6] R.Kumar, S.Bhattacharya, P.Rao: Estimating maize production using satellite and climate indices. Remote Sensing of Environment **245**, 111821 (2020)

[7] P.Mehra, S.Banerjee, A.Choudhury: Rainfall driven yield forecasting in rice using support vector machines. International Journal of Agricultural Science **12**(1), 25–34 (2020)

[8] V.Sharma, N.Khan, R.Dubey: Machine learning based yield prediction using weather and soil data. Agricultural Systems **198**, 103322 (2021)

[9] F.Alam, S.Raza, M.A.Khan: Machine learning based prediction of crop production under drought conditions. Agricultural Water Management **234**, 56–67 (2020)

[10] A.Bose, D.Mukherjee, T.Saha: Climate-driven crop yield prediction using neural networks. Neural Computing and Applications **32**(10), 6003–6013 (2020)

[11] T.Das, S.Yadav, A.Paul: Machine learning approach for yield forecasting of pulses under rainfed conditions. Field Crops Research **250**, 107776 (2020)

[12] V.Desai, P.Mishra, K.Sharma: Crop yield estimation using machine learning on multi-year climate data. Agricultural and Forest Meteorology **292–293**, 108154 (2020)

[13] K.Joshi, A.Iyer, T.Mishra: Using weather variables for crop production forecast. Agricultural Forecasting Journal **9**(3), 210–220 (2020)

[14] A.Nair, S.Verma, H.Pande: Analysis of agricultural data for predicting crop health and yield. Agricultural Systems **178**, 102699 (2020)

[15] K.Patel, L.Sharma, P.Dubey: Comparative study of machine learning models for crop production prediction. Expert Systems with Applications **169**, 114312 (2021)

[16] S.Patil, P.Deshmukh, A.Jagtap: Predicting soybean yields with machine learning techniques. Journal of Agricultural Informatics **11**(2), 87–95 (2019)

[17] M.Patel, P.Das, J.Rana: Predicting rice production in coastal districts. Journal of Environmental Informatics **15**(1), 45–56 (2018)

[18] P.Thakur, A.Reddy, V.Narayan: Comparative study of machine learning models for sugarcane yield prediction. Computers and Electronics in Agriculture **175**, 105583 (2020)

[19] R.Sharma, S.Patel, Kumar: Agricultural yield prediction. Journal of Agricultural Data Science **12**(4), 321–333 (2019)

[20] P.Sharma, A.Gupta, R.Srivastava: Application of ensemble learning in crop productivity forecasting. Computers and Electronics in Agriculture **187**, 106264 (2021)

Appendix:

- Dataset Source: ICRISAT
- Software: Python (Pandas, Scikit-learn, Matplotlib)
- Code and models: Available upon request