International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 11 (November 2025), PP 259-264

Forecasting State Wise Crop Yield in India

ABSTRACT
Agriculture is the backbone of India's economy, with millions of farmers dependent on timely insights into crop productivity for sustainable livelihoods and national food security. However, crop yields are highly sensitive to multiple interacting factors, including climatic variability, soil health, and input management. This project focuses on developing an intelligent, data-driven machine learning system for ***predicting crop yield across Indian states****, integrating historical agricultural, environmental, and meteorological data into a unified predictive framework.*

Date of Submission: 13-11-2025 Date of acceptance: 27-11-2025

The primary objective of this project is to build an accurate, interpretable, and scalable model that can forecast the yield of diverse crops based on factors such as area of cultivation, fertilizer and pesticide use, temperature, rainfall, humidity, soil nutrient composition (Nitrogen, Phosphorus, Potassium), soil pH, and seasonal patterns. The system leverages **supervised machine learning algorithms** - including **XGBoost**, **Random Forest**, and **Histogram Gradient Boosting** - to identify the complex, nonlinear dependencies between these input features and observed crop yields. Through model comparison and hyperparameter tuning,

the XGBoost model was found to deliver superior predictive accuracy and generalization, making it the foundation of the final deployed pipeline.

The dataset used for training and evaluation was compiled from authentic government and open data sources, encompassing multiple years and covering a wide variety of crops, states, and seasons. Preprocessing steps included data standardization, feature scaling, label encoding for categorical attributes (such as crop, season, and state), and normalization of temporal data using a **YearTransformer** to capture long-term yield trends. The final pipeline integrates all preprocessing and transformation steps with the trained model to ensure seamless, real-time prediction on new data.

To enable accessibility and practical usability, the model has been deployed as an **interactive Streamlit web application**. This app allows users - including farmers, agronomists, and policymakers—to input real-world parameters and instantly obtain yield predictions (expressed in quintals per hectare). The interface automatically handles encoding, scaling, and transformation, ensuring consistent predictions aligned with the trained pipeline. The application's intuitive design, coupled with visual data previews, provides transparency and ease of interpretation for end-users.

In conclusion, the proposed **Crop Yield Prediction System** demonstrates how advanced machine learning techniques can transform raw agricultural data into actionable intelligence.

Resources or Prerequisite

Data Resources

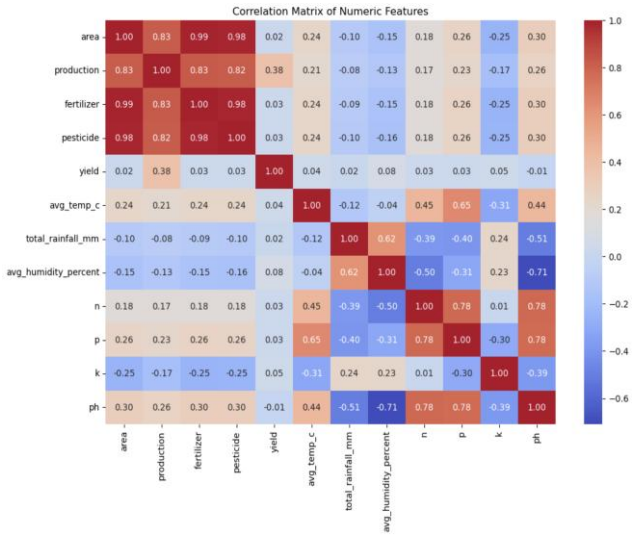
The project uses a combination of crop, soil, and weather datasets to build an accurate predictive model. The datasets were collected from **India.gov.in** (https://india.gov.in/category/agriculture-rural-environment/subcategory/agricultural-produce) and include:

1. crop_yield.csv – Contains historical yield data for major crops such as rice, wheat, sugarcane, and maize across Indian states, including attributes like year, crop, season, area, production, and yield.
2. state_soil_data.csv – Provides soil characteristics for different states, including NPK content (Nitrogen, Phosphorus, Potassium), pH value, and other soil fertility parameters.
3. state_weather_data_1997_2020.csv – Contains historical weather information such as average temperature, rainfall, and humidity from 1997 to 2020, enabling correlation analysis with crop yields.

These datasets are integrated and preprocessed in Python to create a unified dataset suitable for machine learning modeling. Preprocessing steps include handling missing values, encoding categorical variables, scaling numerical features, and feature engineering.

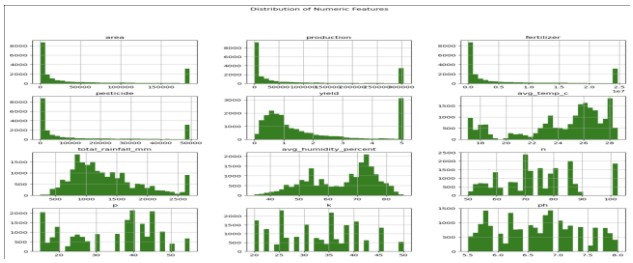
EDA Visualizations

Correlation matrix for numeric features



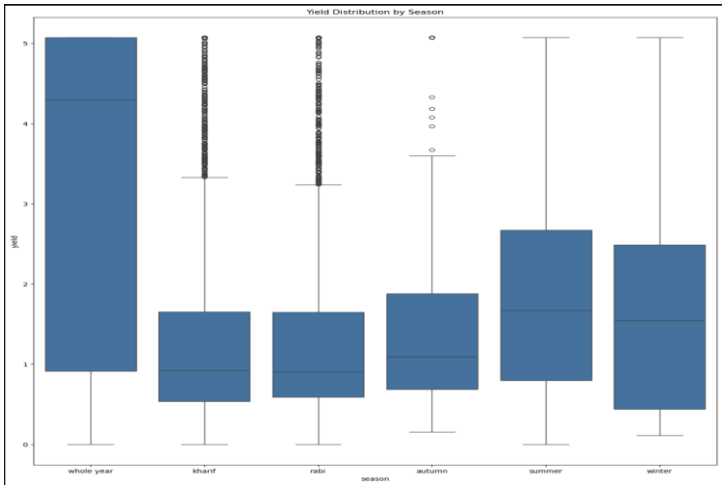
- Values close to +1 (red) → strong positive correlation: as one variable increases, the other tends to increase.
- Values close to -1 (blue) → strong negative correlation: as one variable increases, the other decreases.
- Values near 0 (white/gray) → little or no linear relationship.

Correlation matrix for numeric features

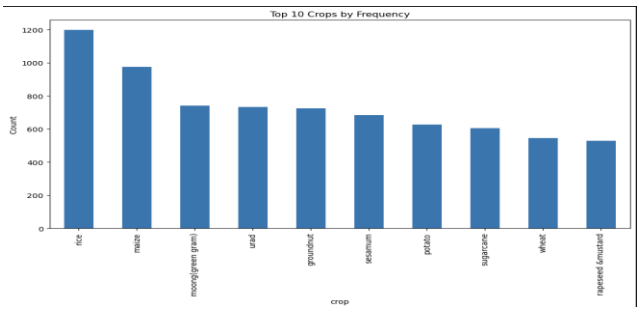


This code visualizes the distribution of key numerical features in the dataset using histograms. It helps identify the spread, skewness, and presence of outliers in variables like area, production, rainfall, and yield during EDA.

Boxplots to check outliers grouped by season

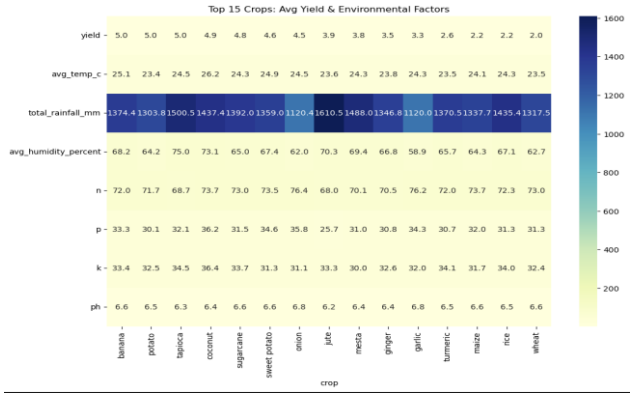


This code creates a **boxplot of crop yield across different seasons** to visualize how yield varies seasonally. It helps in **detecting outliers and comparing yield distribution** among Kharif, Rabi, and other seasons.
Top 10 Crops by Frequency

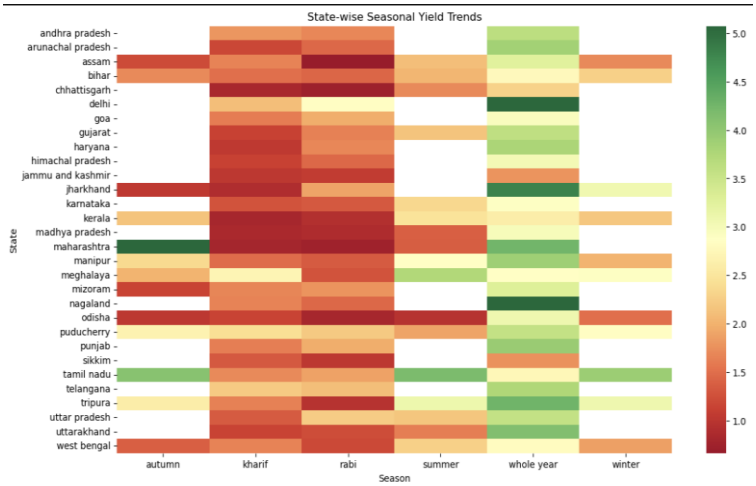


This code plots a bar chart showing the 10 most frequently occurring crops in the dataset. It helps quickly identify which crops are most common.

Top 15 Crops: Average Yield & Environmental Factors



This analysis groups crops by their average yield and associated environmental/soil factors like temperature, rainfall, humidity, and nutrients (N, P, K, pH). A heatmap visualizes the top 15 crops, making it easy to compare how different factors relate to yield. Banana, potato, and tapioca show the highest average yields, typically growing under moderate temperatures (23–26°C) and high rainfall (1300–1500 mm). Nutrient levels and soil pH also vary slightly across crops, indicating specific soil requirements for optimal growth. Overall, the visualization highlights key crops and the environmental conditions that favor higher yields.
State-wise Seasonal Yield Trends



This analysis calculates the average crop yield for each state across different seasons. The resulting heatmap visualizes state-wise seasonal yield trends, highlighting how productivity varies throughout the year. States like Delhi and Andhra Pradesh show high yields during the "kharif" and "whole year" seasons, while regions like Assam and Bihar have more consistent yields across multiple seasons. This helps identify which states perform best in specific seasons and informs targeted agricultural planning.

Interactive state-level dashboard showing

An interactive dashboard was created using Plotly to visualize crop yield trends across Indian states. The dashboard presents both seasonal and crop-based yield patterns for each state, helping to identify key crops and high-performing seasons.

The visualization includes:

- Seasonal Average Yield Chart – Displays the mean yield for each agricultural season (Kharif, Rabi, etc.) within a selected state.
- Season Summary Table – Shows the number of records available per season for that state.
- Top Crops by Average Yield Chart – Highlights the top 10 crops based on their average yield for the selected state.
- Crop Summary Table – Lists the most frequently cultivated crops and their occurrence count.

Machine Learning Modelling & Techniques Applied

Problem Statement

The goal is to predict crop yield using historical agricultural data, including crop type, state, seasonal factors, weather conditions, and soil nutrient levels. Machine learning models were applied to capture complex relationships between these features and crop yield.

Data Preparation

- The dataset was copied to a modeling dataframe df_model to avoid altering the original data.
- The target variable y was defined as yield.
- Columns like production were dropped since they are dependent on area and yield, preventing data leakage.
- Features were categorized for appropriate preprocessing:
    - Categorical Label Encoding: crop, state
    - One-Hot Encoding: season
    - Numerical transformations:
        - Log transformation for skewed features: area, fertilizer, pesticide, total_rainfall_mm
        - Standard scaling for temperature and humidity: avg_temp_c, avg_humidity_percent
        - Min-Max scaling for soil nutrients: n, p, k, ph
    - Year Transformation: Scaled to 0–1 range to normalize chronological effects.

Preprocessing Pipeline

- A ColumnTransformer was created to apply specific preprocessing steps to each feature type, ensuring reproducible and consistent transformations.
- Custom YearTransformer scales the year to a 0–1 range, preserving temporal information without biasing models due to raw year values.

Train/Test Split

- Data was split chronologically:
  - Training set: years ≤ 2017
  - Test set: years > 2017
- This approach mimics real-world prediction scenarios where future data is unknown during training.

Machine Learning Models Applied

Several regression models were trained to predict crop yield:

1. Linear Regression – baseline linear model for simple relationships.
2. Random Forest Regressor – ensemble tree-based model capturing non-linear patterns.
3. HistGradientBoosting Regressor – advanced gradient boosting with histogram optimization.
4. Support Vector Regressor (SVR) – kernel-based model for capturing complex non-linear relationships.
5. MLP Regressor – feedforward neural network with hidden layers (64, 32).
6. XGBoost Regressor – gradient boosting algorithm with hyperparameters:
   - n_estimators=1000, max_depth=8, learning_rate=0.05
   - subsample=0.8, colsample_bytree=0.9

Each model was integrated into a Pipeline with the preprocessor for streamlined training and evaluation.

Model Evaluation

- Models were evaluated using R^2 score on both training and test sets.
- Results Summary:

Model	R^2 Train	R^2 Test
XGBoost	0.997	0.922
RandomForest	0.990	0.903
HistGradient	0.899	0.850
MLPRegressor	0.506	0.403
LinearRegression	0.202	0.151
SVR	0.070	-0.027

- XGBoost outperformed all models, achieving the highest R^2 on the test set (0.922), indicating strong predictive performance with minimal overfitting.
- Seasonal factors (season_whole year, season_winter) and crop type are the most influential predictors.
- Soil nutrients (n, p, k, ph) also contribute significantly, highlighting their role in yield determination.

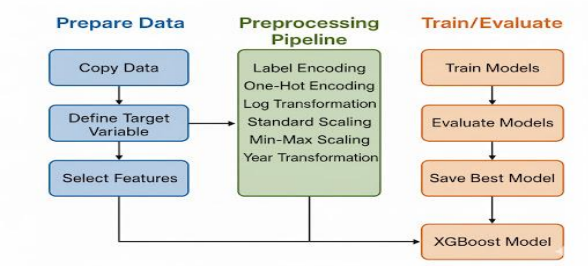
Model Deployment

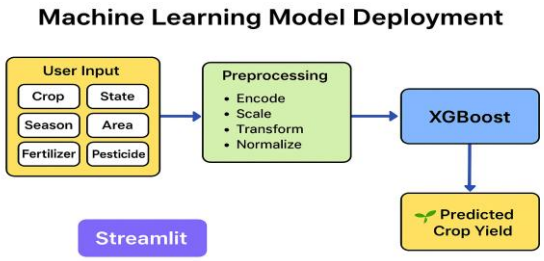
- The best-performing XGBoost model, along with the preprocessing pipeline, was saved as a pickle file (best_model_XGBoost.pkl) for future predictions and integration into applications.

Summary

- A comprehensive machine learning pipeline was built, from data preprocessing to model evaluation.
- XGBoost was identified as the most effective model for predicting crop yield.
- The pipeline captures non-linear relationships, handles categorical and numerical transformations, and provides interpretable feature importance.
- This approach ensures a reproducible, scalable, and accurate yield prediction system suitable for practical agricultural applications.

Model Building Flow Diagram





Key Insights of Project

- **Data Overview & Patterns**
  - Crop yield varies significantly across states and seasons.
  - Certain crops dominate the dataset, showing higher frequency and production.
- **Feature Importance**
  - Features like area, fertilizer, avg_temp_c, and total_rainfall_mm strongly influence yield.
  - Seasonal and state-level variations also play a significant role in prediction.
- **Outlier & Data Quality Analysis**
  - Outliers exist in yield, area, and production, affecting model accuracy.
  - Proper data cleaning and scaling improved model stability.
- **Model Performance**
  - Tree-based models like Random Forest and XGBoost performed better than linear models.
  - Hyperparameter tuning and cross-validation enhanced predictive

Project GitHub Link

https://github.com/swaroopkumaraduru/Swaroop-Kumar-Aduru-Forecasting-Crop-Yield-in-India

BIBLIOGRAPHY

1. India Crop Production. https://www.kaggle.com/datasets/thedevastator/statewise-crop-production-in-india-a-statistical
2. Crop Production Trends https://www.kaggle.com/code/abhijitdahatonde/crop-production-trends

REFERENCES

[1]. Overview of Regression Models and How to Determine the Best Model for Data : https://journaljsrr.com/index.php/JSRR/article/view/2452/5060

[2]. Technology and Management for Social Innovation (IATMSI), Gwalior, India,21–23 December 2022Ranjan, P.; Garg, R.; Rai, J.K. Artificial Intelligence Applications in Soil & Crop Management. In Proceedings of the IEEE Conference on Interdisciplinary Approaches in

[3]. Gehlot, A.; Sidana, N.; Jawale, D.; Jain, N.; Singh, B.P.; Singh, B. Technical analysis of crop production prediction using MachineLearning and Deep Learning Algorithms. In Proceedings of the International Conference on Innovative Computing, Intelligent.Communication and Smart Electrical Systems (ICSES), Chennai, India, 24–25 September 2022;

[4]. Vivek, S.; Ashish, K.T.; Himanshu, M. Technological revolutions in smart farming: Current trends, challenges & future directions. Compute. Electron. Agric. **2022**.

[5]. Mamatha, J.C.K. Machine learning based crop growth management in greenhouse environment using hydroponics farming techniques. Meas. Sens. 2023, 25, 100665.