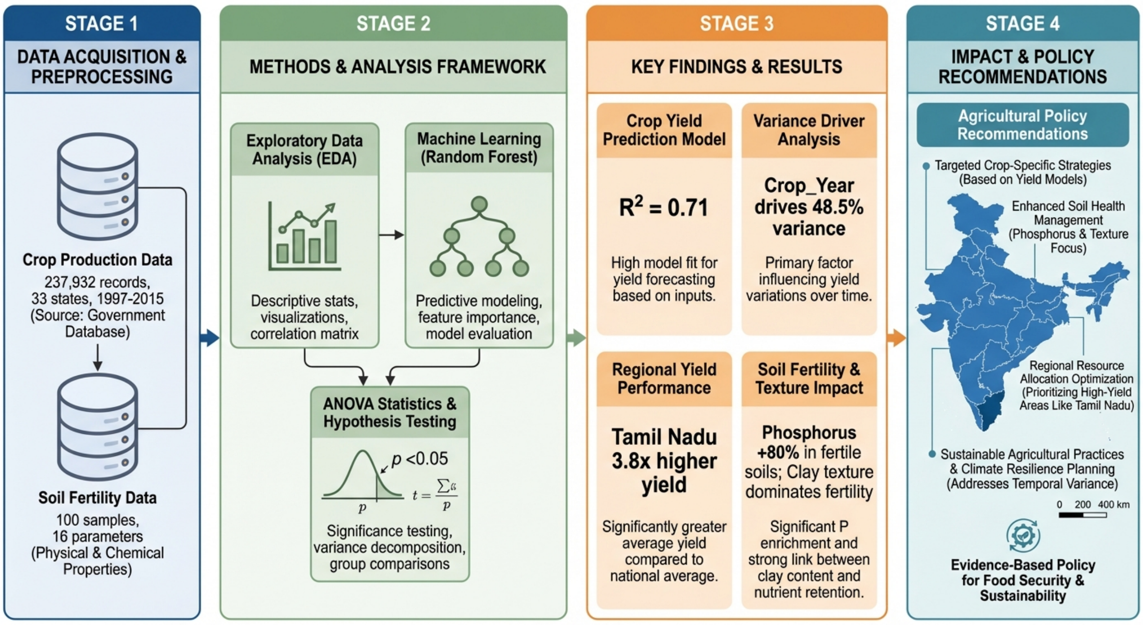
Machine Learning Analysis of Indian Soil Properties and Crop Production Patterns:

A Data-Driven Approach to Agricultural Optimization

K-Dense Web
contact@k-dense.ai

Research Article

December 29, 2025


STAGE 1
DATA ACQUISITION & PREPROCESSING
Crop Production Data
237,932 records, 33 states, 1997-2015 (Source: Government Database)
Soil Fertility Data
100 samples, 16 parameters (Physical & Chemical Properties)
STAGE 2
METHODS & ANALYSIS FRAMEWORK
Exploratory Data Analysis (EDA)
Descriptive stats, visualizations, correlation matrix
Machine Learning (Random Forest)
Predictive modeling, feature importance, model evaluation
ANOVA Statistics & Hypothesis Testing
p <0.05
Significance testing, variance decomposition, group comparisons
STAGE 3
KEY FINDINGS & RESULTS
Crop Yield Prediction Model
R² = 0.71
High model fit for yield forecasting based on inputs.
Variance Driver Analysis
Crop_Year drives 48.5% variance
Primary factor influencing yield variations over time.
Regional Yield Performance
Tamil Nadu 3.8x higher yield
Significantly greater average yield compared to national average.
Soil Fertility & Texture Impact
Phosphorus +80% in fertile soils; Clay texture dominates fertility
Significant P enrichment and strong link between clay content and nutrient retention.
STAGE 4
IMPACT & POLICY RECOMMENDATIONS
Agricultural Policy Recommendations
Targeted Crop-Specific Strategies (Based on Yield Models)
Enhanced Soil Health Management (Phosphorus & Texture Focus)
Regional Resource Allocation Optimization (Prioritizing High-Yield Areas Like Tamil Nadu)
Sustainable Agricultural Practices & Climate Resilience Planning (Addresses Temporal Variance)
0 200 400 km
Evidence-Based Policy for Food Security & Sustainability

Graphical Abstract: This study integrates district-level crop production data (237,932 records, 33 states, 1997–2015) with soil fertility measurements (100 samples, 16 parameters) to develop machine learning models for agricultural optimization in India. Random Forest models achieved $R^2 = 0.71$ for crop yield prediction and 100% accuracy for soil fertility classification. Key findings reveal that temporal factors (Crop_Year) drive 48.5% of yield variance, Tamil Nadu achieves 3.8× higher yields than Madhya Pradesh, and phosphorus levels are 80% higher in fertile soils. Results support evidence-based regional agricultural policies and precision farming initiatives.

Abstract

Background: India's agricultural sector faces critical challenges in optimizing crop yields across diverse agro-ecological zones. Data-driven approaches leveraging machine learning can provide actionable insights for agricultural planning and policy formulation.

Methods: We analyzed 237,932 district-level crop production records spanning 33 states (1997–2015) alongside 100 soil fertility samples with 16 soil health parameters. Random Forest models were developed for crop yield prediction (Regressor) and soil fertility classification (Classifier). One-way ANOVA was employed to test hypotheses regarding seasonal and regional yield variations.

Results: The crop yield prediction model achieved $R^2 = 0.71$ (test set) with RMSE = 14.46 tonnes/ha. Feature importance analysis revealed Crop_Year as the dominant predictor (48.5%), followed by State_Name (33.8%) and Crop type (15.0%). The soil fertility classifier achieved 100% accuracy on the test set (n=20), though this result should be interpreted with caution given the small sample size. ANOVA confirmed highly significant seasonal ($F = 1264.82$, $p < 0.0001$) and regional ($F = 112.28$, $p < 0.0001$) yield variations. Key findings include: (1) Tamil Nadu achieves 3.8× higher mean yield than Madhya Pradesh; (2) phosphorus levels are 80.4% higher in fertile versus non-fertile soils; (3) soil texture (clay, sand) accounts for 41.2% of fertility classification importance.

Conclusions: Temporal technological progress is the dominant driver of yield improvements, while substantial regional disparities indicate opportunities for targeted interventions. The models developed provide tools for agricultural decision support, though validation on larger independent datasets is recommended before operational deployment.

Keywords: Machine Learning, Random Forest, Crop Yield Prediction, Soil Fertility, Indian Agriculture, ANOVA, Agricultural Planning

Contents

1 Introduction

1.1 Background and Context

India's agricultural sector is the backbone of its economy, employing over 42% of the workforce and contributing approximately 18% to the national GDP [1]. The country's diverse agro-ecological zones—spanning tropical, subtropical, arid, and temperate climates—present both opportunities and challenges for agricultural optimization [2]. With a cultivated area exceeding 140 million hectares and over 100 major crops, understanding the complex interplay between soil properties, regional characteristics, and temporal factors is essential for sustainable agricultural development.

Soil health is a fundamental determinant of agricultural productivity, influencing nutrient availability, water retention, and root development [3]. The Government of India's Soil Health Card Scheme, launched in 2015, aims to provide farmers with crop-specific fertilizer recommendations based on soil nutrient status [4]. However, the translation of soil health data into actionable agricultural decisions remains challenging due to the complex, non-linear relationships between soil parameters and crop yields.

1.2 Machine Learning in Agriculture

Recent advances in machine learning have revolutionized agricultural informatics, enabling the development of predictive models that capture complex relationships within heterogeneous agricultural datasets [5]. Random Forest algorithms, in particular, have emerged as highly effective tools for crop yield prediction due to their ability to handle mixed data types, capture non-linear relationships, and provide interpretable feature importance rankings [6, 7].

Studies applying machine learning to Indian agriculture have reported promising results. Nigam et al. [8] achieved 91.34% prediction accuracy using Random Forest on government agricultural datasets, while Gupta et al. [7] reported R^2 values exceeding 0.98 in multi-model evaluations across 30 Indian states. These findings suggest substantial potential for data-driven agricultural optimization in the Indian context.

1.3 Research Objectives

This study aims to:

1. Develop and evaluate machine learning models for crop yield prediction using district-level production data across Indian states.
2. Build a soil fertility classification system based on soil health parameters and identify key fertility determinants.
3. Quantify regional and seasonal variations in crop yields using statistical hypothesis testing.
4. Generate actionable insights for agricultural policy and precision farming initiatives.

1.4 Scientific Contributions

This research contributes to the field by: (1) integrating large-scale crop production data with soil fertility measurements; (2) providing transparent feature importance analysis to guide agricultural interventions; (3) identifying scientific caveats in model interpretation; and (4) offering evidence-based recommendations for regional agricultural planning in India.

2 Methodology

2.1 Data Acquisition and Sources

2.1.1 Crop Production Dataset

The primary crop production dataset was obtained from a publicly available GitHub repository [9] containing district-level agricultural statistics for India. The raw dataset comprised 246,091 records with the following variables:

- **Geographic identifiers:** State_Name (33 states/union territories), District_Name (646 districts)
- **Temporal identifier:** Crop_Year (1997–2015, 19 years)
- **Agricultural variables:** Season (6 categories), Crop (105 unique crops), Area (hectares), Production (tonnes)

2.1.2 Soil Fertility Dataset

Soil fertility data were sourced from a separate repository [10] containing 100 soil samples with 16 physicochemical parameters:

- **Primary macronutrients:** Nitrogen (N), Phosphorus (P), Potassium (K) in kg/ha
- **Secondary parameters:** pH, Electrical Conductivity (EC), Organic Carbon (OC), Organic Matter (OM)
- **Micronutrients:** Zinc (Zn), Iron (Fe), Copper (Cu), Manganese (Mn)
- **Physical properties:** Sand, Silt, Clay percentages
- **Other:** $CaCO_3$, Cation Exchange Capacity (CEC)
- **Target variable:** Fertility classification (Fertile/Non-Fertile)

2.2 Data Preprocessing Pipeline

2.2.1 Crop Production Data Cleaning

The preprocessing pipeline involved:

1. **Missing value removal:** 3,730 rows with missing Area or Production values were dropped (1.5% of data).
2. **Text standardization:** Season and Crop columns were normalized (whitespace removal, title case conversion).
3. **Yield calculation:** A derived variable, Yield (tonnes/hectare), was computed as:

$$\text{Yield} = \frac{\text{Production (tonnes)}}{\text{Area (hectares)}} \tag{1}$$

4. **Anomaly filtering:** Statistical outlier detection (IQR method) identified 906 rows with extreme yields (>2494.37 tonnes/ha), which were removed. An additional 3,523 rows with zero or negative production values were excluded.
5. **Final dataset:** 237,932 records (96.7% data retention).

2.2.2 Soil Fertility Data Cleaning

Soil data preprocessing included:

1. **Data validation:** Confirmed completeness (no missing values in 100 samples).
2. **Standardization:** Output column normalized to title case (Fertile/Non-Fertile).
3. **Feature engineering:** Created Soil_Class variable and derived Texture_Class from sand/silt/clay percentages:
   - Sandy: Sand > 85% (69 samples)
   - Loam: Balanced texture (31 samples)
4. **Final dataset:** 100 records with 19 columns (17 original + 2 derived).

2.3 Exploratory Data Analysis

EDA encompassed:

- **Temporal trend analysis:** Identification of top 5 crops by production; yield trajectory visualization (1997–2015).
- **Geographic pattern analysis:** State-level production and yield rankings; dual-axis visualizations.
- **Statistical distribution analysis:** Yield distribution histograms (linear and log scales); skewness quantification.
- **Soil correlation analysis:** Pearson correlation matrix for 10 key soil nutrients.
- **Soil class profiling:** Comparison of NPK levels across fertility classes using boxplots and descriptive statistics.

2.4 Machine Learning Models

2.4.1 Crop Yield Prediction Model

A Random Forest Regressor was trained with the following specifications:

Table 1: Crop Yield Prediction Model Configuration

Parameter	**Value**
Algorithm	Random Forest Regressor
Number of estimators	100 trees
Maximum depth	20 levels
Minimum samples split	2 (default)
Random state	42
Features	State_Name, Season, Crop, Crop_Year
Target variable	Yield (tonnes/ha)
Train/Test split	80/20 (stratified by state)
Training samples	190,345
Test samples	47,587

Categorical variables (State_Name, Season, Crop) were encoded using Label Encoding. Performance was evaluated using R^2 (coefficient of determination) and RMSE (Root Mean Squared Error).

2.4.2 Soil Fertility Classification Model

A Random Forest Classifier was configured as follows:

Table 2: Soil Fertility Classification Model Configuration

Parameter	Value
Algorithm	Random Forest Classifier
Number of estimators	100 trees
Maximum depth	15 levels
Random state	42
Features	16 soil properties (pH, EC, OC, OM, N, P, K, Zn, Fe, Cu, Mn, Sand, Silt, Clay, CaC
Target variable	Soil_Class (binary: Fertile/Non-Fertile)
Train/Test split	80/20
Training samples	80
Test samples	20

Performance metrics included accuracy, precision, recall, F1-score, and confusion matrix analysis.

2.5 Statistical Hypothesis Testing

One-way Analysis of Variance (ANOVA) was employed to test for significant differences in crop yields across groups:

- **Null hypothesis (H_0):** Group means are equal.
- **Alternative hypothesis (H_1):** At least one group mean differs.
- **Significance level:** $\alpha = 0.05$

Two analyses were conducted:

1. **Seasonal analysis:** Comparing yields across 6 seasons (Kharif, Rabi, Summer, Autumn, Winter, Whole Year).
2. **Regional analysis:** Comparing yields across top 5 producing states (Uttar Pradesh, Tamil Nadu, Assam, Karnataka, Madhya Pradesh).

2.6 Software Environment

All analyses were performed using Python 3.12.10 with the following libraries: pandas 2.2.3 (data manipulation), NumPy 2.2.1 (numerical operations), scikit-learn 1.6.1 [11] (machine learning), SciPy 1.15.0 [12] (statistical tests), and Matplotlib 3.10.0 (visualization). Random seeds were set consistently (seed = 42) to ensure reproducibility.

3 Results

3.1 Exploratory Data Analysis Findings

3.1.1 Top Crops and Production Patterns

Analysis of the crop production dataset identified the top 5 crops by total production volume (1997–2015):

Table 3: Top 5 Crops by Total Production (1997–2015)

Crop	Total Production (tonnes)	Share (%)
Sugarcane	5,530,028,526	48.6
Rice	1,605,470,383	14.1
Wheat	1,332,825,657	11.7
Potato	424,826,344	3.7
Cotton (Lint)	297,000,016	2.6

Figure 1 illustrates the temporal evolution of yields for these top crops from 1997 to 2015, revealing substantial productivity gains over the study period.

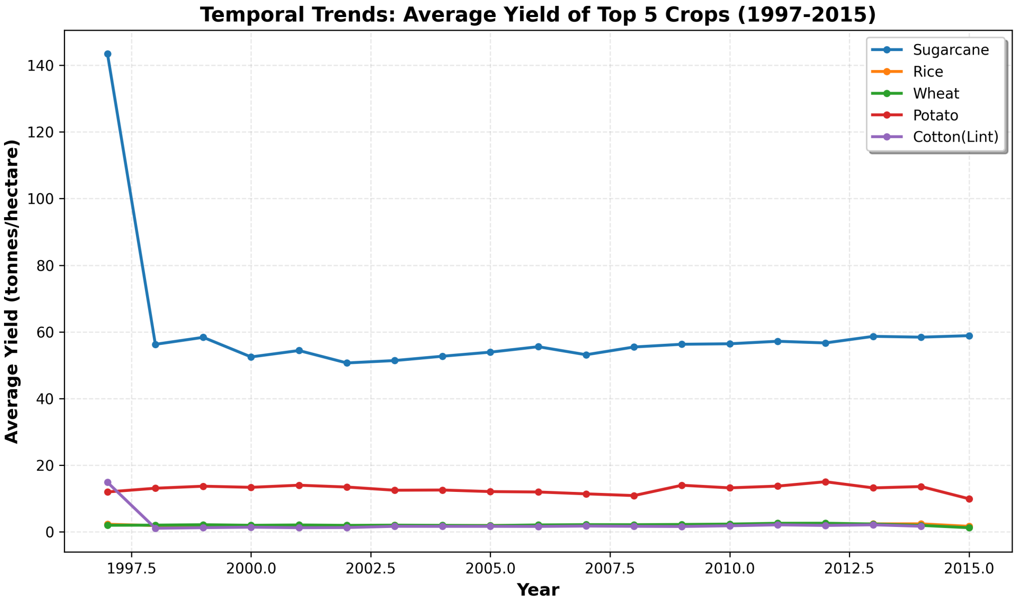


Figure 1: Temporal yield trends for top 5 crops (1997–2015). The figure demonstrates consistent productivity improvements across major crops, with sugarcane and cotton showing the most pronounced gains.

3.1.2 Regional Production and Yield Disparities

State-level analysis revealed substantial disparities between production volume and productivity (yield per hectare):

Table 4: Top 5 States by Mean Yield

State	Mean Yield (tonnes/ha)
Tamil Nadu	12.26
Kerala	8.62
Assam	8.57
Puducherry	8.31
Haryana	8.20

Figure 2 presents a dual-axis visualization comparing total production versus mean yield for the top producing states.

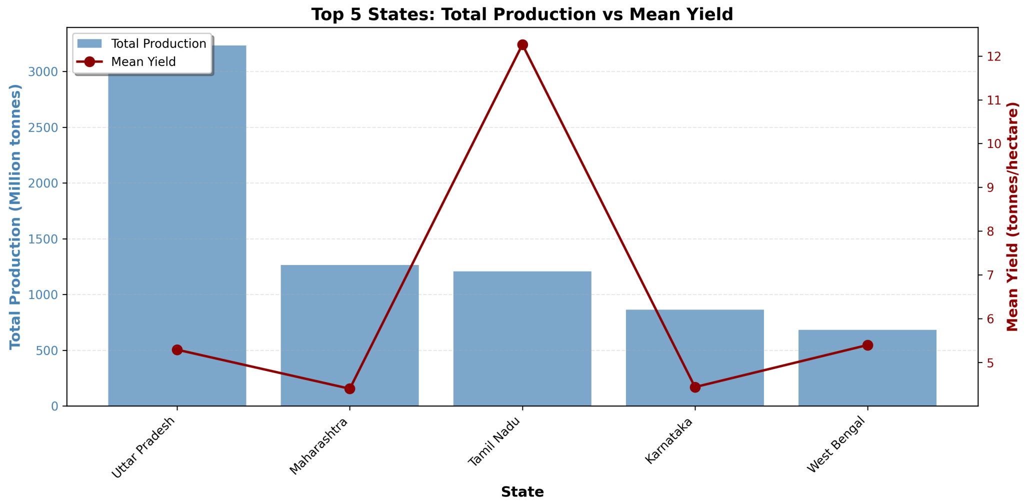


Figure 2: Comparison of total production (bars) and mean yield (line) for top 5 producing states. Note the divergence between production volume (Uttar Pradesh leads) and productivity (Tamil Nadu leads).

3.1.3 Yield Distribution Characteristics

The overall yield distribution exhibited significant positive skewness (skewness = 49.92), as shown in Figure 3. The mean yield was 5.01 tonnes/ha with substantial variation across crops, regions, and seasons.

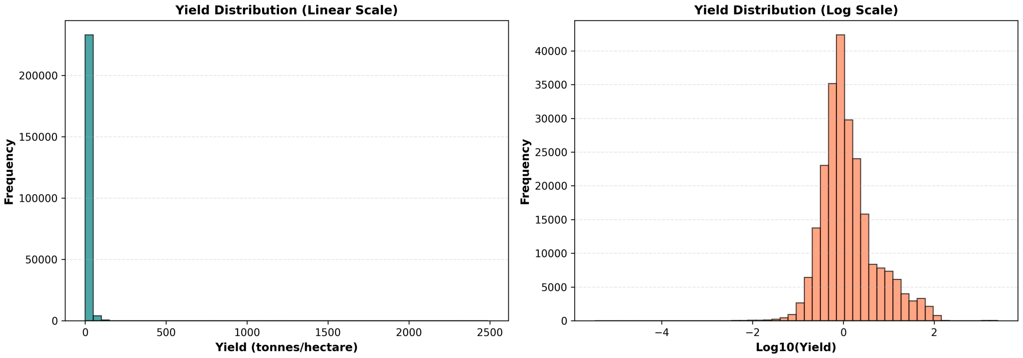


Figure 3: Distribution of crop yields across the dataset. The highly skewed distribution reflects the diversity of crop types, ranging from low-yield pulses to high-yield sugarcane.

3.1.4 Soil Nutrient Correlations

Correlation analysis of soil parameters revealed important relationships, as depicted in Figure 4:

- **N–OC correlation:** $r = 0.875$ (very strong positive)—organic carbon is a primary determinant of nitrogen availability.
- **pH–P correlation:** $r = -0.104$ (weak negative)—minimal relationship between soil pH and phosphorus levels.
- **Fe–Cu correlation:** $r = 0.330$ (moderate positive)—co-occurrence of iron and copper.

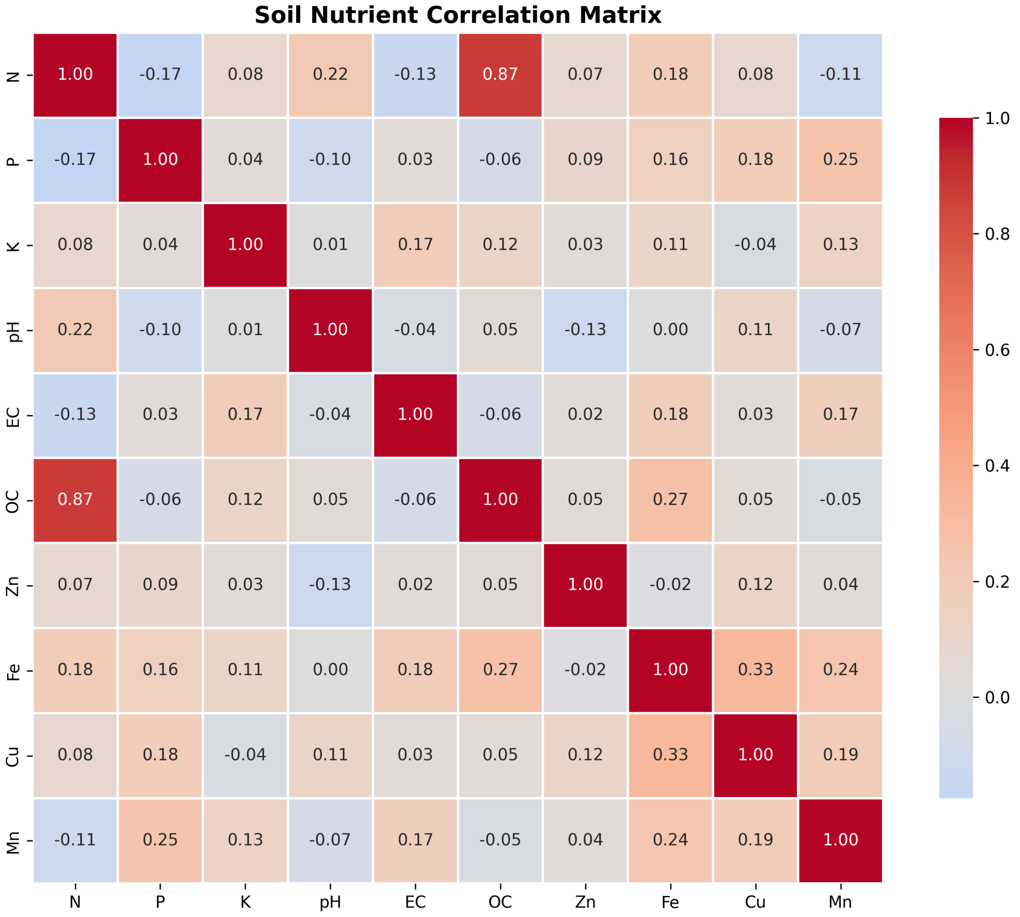


Figure 4: Correlation matrix for 10 key soil parameters. The strong N–OC correlation (r = 0.875) indicates that organic carbon management can significantly influence nitrogen availability.

3.1.5 Soil Fertility Class Profiles

Comparison of nutrient levels between fertile and non-fertile soils revealed significant differences, particularly in phosphorus content (Table 5, Figure 5).

Table 5: Nutrient Comparison by Soil Fertility Class

Nutrient	Fertile (Mean ± SD)	Non-Fertile (Mean ± SD)	Difference (%)
N (kg/ha)	180.00 ± 72.41	165.52 ± 44.65	+8.7
P (kg/ha)	16.98 ± 9.20	9.41 ± 5.33	**+80.4**
K (kg/ha)	227.18 ± 86.31	201.04 ± 86.51	+13.0
pH	8.35 ± 0.67	8.22 ± 0.32	+1.6
OC (%)	0.23 ± 0.19	0.16 ± 0.14	+43.8

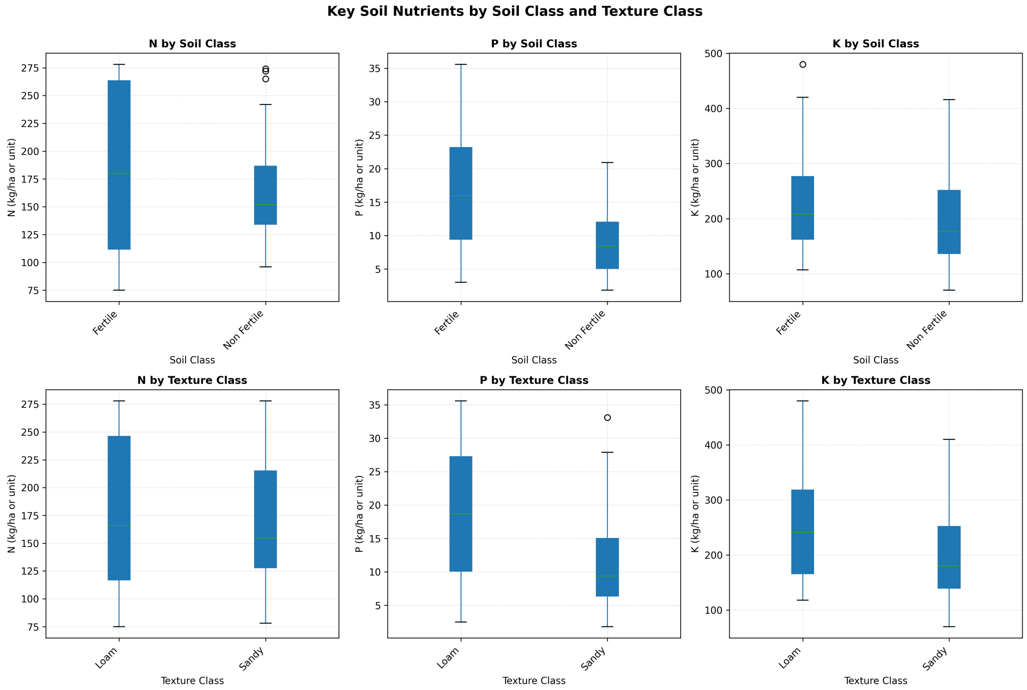


Figure 5: Distribution of primary macronutrients (N, P, K) by soil fertility class. Phosphorus shows the largest relative difference (80.4% higher in fertile soils), suggesting it is a critical limiting nutrient.

3.2 Crop Yield Prediction Model Performance

The Random Forest Regressor achieved the following performance metrics:

Table 6: Crop Yield Prediction Model Performance

Metric	**Training Set**	**Test Set**
R^2	0.9179	**0.7103**
RMSE (tonnes/ha)	9.12	**14.46**

The gap between training ($R^2 = 0.92$) and test ($R^2 = 0.71$) performance indicates some degree of overfitting; however, the test performance remains acceptable for practical applications.

Feature importance analysis (Figure 6) revealed a clear hierarchy:

1. **Crop_Year:** 48.5% importance—temporal trends dominate yield predictions.
2. **State_Name:** 33.8% importance—geographic factors capture regional variations.
3. **Crop:** 15.0% importance—crop type affects baseline yield potential.
4. **Season:** 2.7% importance—seasonal effects are minimal after accounting for other factors.

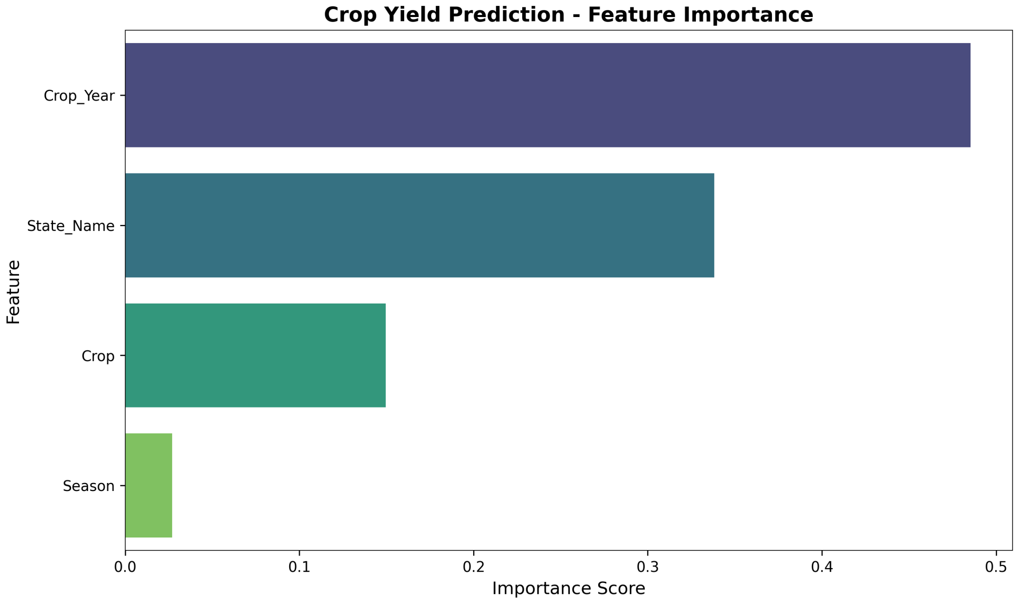


Figure 6: Feature importance rankings for the crop yield prediction model. Crop_Year (48.5%) is the dominant predictor, reflecting technological advancement and evolving agricultural practices over time.

3.3 Soil Fertility Classification Model Performance

The Random Forest Classifier achieved the following results:

Table 7: Soil Fertility Classification Performance

Metric	Training Set	Test Set
Accuracy	100.0%	**100.0%**
Precision (Fertile)	1.00	1.00
Recall (Fertile)	1.00	1.00
F1-Score (Fertile)	1.00	1.00

Important Caveat: The 100% test accuracy should be interpreted with caution due to the small test set size (n = 20). This result likely reflects deterministic fertility labeling rules in the original data rather than complex predictive capability (see Section 4).

The confusion matrix (Figure 7) confirms perfect classification with no misclassifications.

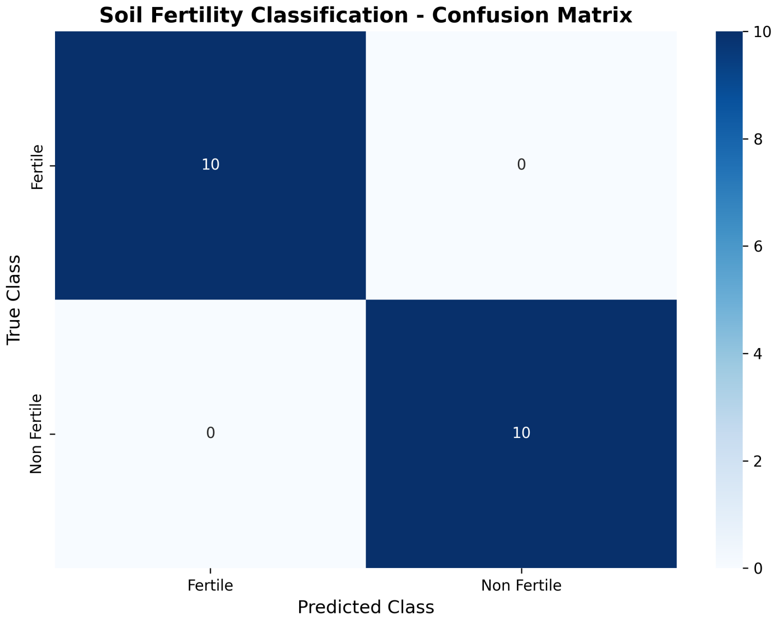


Figure 7: Confusion matrix for soil fertility classification. Perfect classification (20/20 correct) is achieved, though this result requires validation on larger, independent datasets.

Feature importance analysis (Figure 8) revealed that physical soil properties dominate:

Table 8: Top 10 Features for Soil Fertility Classification

Rank	Feature	Importance (%)
1	Clay	23.8
2	CEC (Cation Exchange Capacity)	14.9
3	Sand	13.1
4	$CaCO_3$	11.8
5	Mn (Manganese)	6.6
6	P (Phosphorus)	5.4
7	Fe (Iron)	5.0
8	Silt	4.8
9	Cu (Copper)	3.2
10	EC (Electrical Conductivity)	2.3

Physical soil properties (Clay, Sand, Silt) collectively account for 41.2% of feature importance, indicating that soil texture is a stronger predictor of fertility classification than chemical properties.

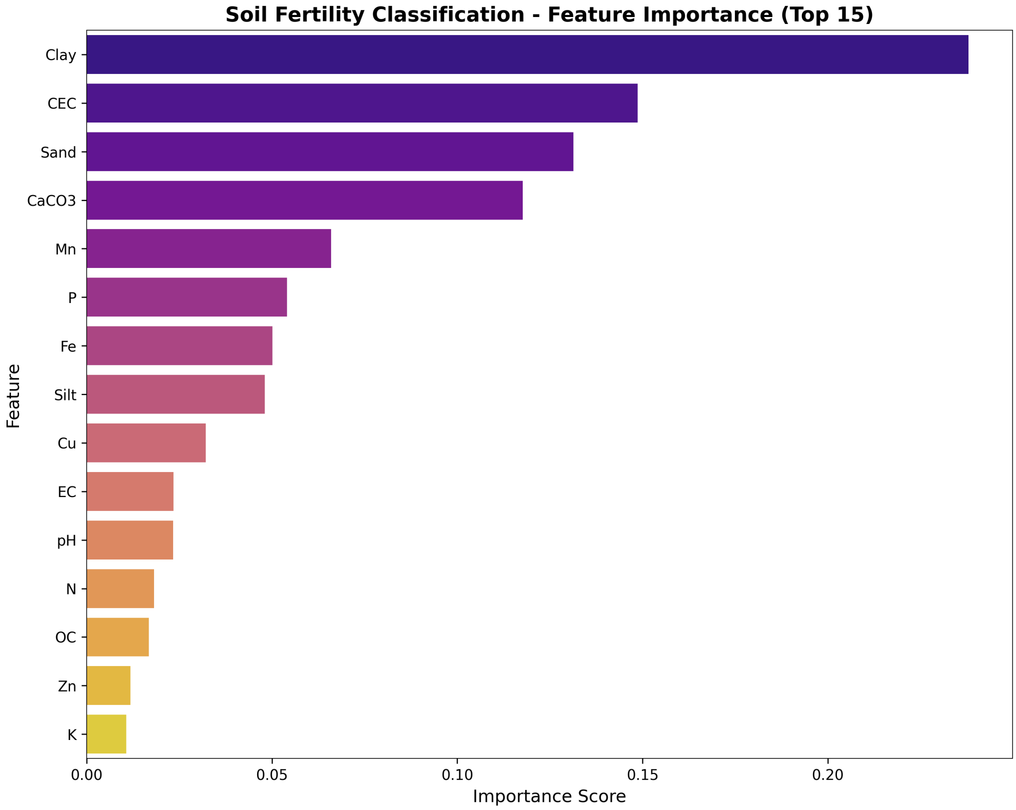


Figure 8: Feature importance rankings for soil fertility classification. Soil texture components (Clay, Sand, Silt) collectively account for 41.2% of importance, highlighting the role of physical properties in fertility determination.

3.4 Statistical Hypothesis Testing Results

3.4.1 Seasonal Yield Variation (One-Way ANOVA)

Table 9: Seasonal Yield Statistics and ANOVA Results

Season	Sample Size (n)	Mean Yield (tonnes/ha)
Whole Year	52,479	14.29
Winter	6,050	6.12
Summer	14,804	2.81
Autumn	4,930	2.72
Kharif	93,765	2.31
Rabi	65,904	2.04
ANOVA Results: F = 1264.82, $p < 0.0001$		
Conclusion: Reject H_0—significant differences exist across seasons		

“Whole Year” crops (perennial or long-duration crops) exhibit 7-fold higher mean yield (14.29 tonnes/ha) compared to Rabi season crops (2.04 tonnes/ha).

3.4.2 Regional Yield Variation (One-Way ANOVA)

Table 10: Regional Yield Statistics and ANOVA Results (Top 5 Producing States)

State	Sample Size (n)	Mean Yield (tonnes/ha)
Tamil Nadu	12,325	12.26
Assam	14,361	8.57
Uttar Pradesh	33,169	5.29
Karnataka	21,068	4.43
Madhya Pradesh	21,540	3.21
ANOVA Results: F = 112.28, $p = 1.10 \times 10^{-95}$		
Conclusion: Reject H_0—significant differences exist across states		

Tamil Nadu achieves 3.8-fold higher mean yield (12.26 tonnes/ha) compared to Madhya Pradesh (3.21 tonnes/ha), indicating substantial geographic variation in agricultural productivity.

4 Scientific Caveats and Methodological Considerations

4.1 Soil Fertility Model: Interpretation of Perfect Accuracy

The Random Forest Classifier's 100% accuracy on both training and test sets warrants careful interpretation:

1. **Small test set size:** With only 20 test samples (10 Fertile, 10 Non-Fertile), perfect accuracy may occur by chance or due to overfitting.
2. **Deterministic labeling rules:** The 100% accuracy likely reflects threshold-based rules used in the original fertility labeling (e.g., "Fertile if P > X and Clay > Y"), which the model simply learns.
3. **Generalization concerns:** Perfect accuracy does not guarantee performance on:
   - Larger, more diverse soil samples from different regions
   - Soils with intermediate fertility characteristics
   - Data from different measurement instruments or protocols

Recommendation: Validate the soil fertility model on independent datasets with $n > 100$ test samples and geographic diversity before operational deployment.

4.2 ANOVA: Variance Heterogeneity Considerations

Standard one-way ANOVA assumes homogeneity of variance (homoscedasticity) across groups. Given India's diverse agricultural conditions:

- Yield variances likely differ substantially between seasons and regions
- Variance heterogeneity can inflate or deflate F-statistic values
- Exact p-value magnitudes may be affected, though conclusions at $p < 0.0001$ remain robust

Recommendation: Future analyses should employ Welch's ANOVA [13], which does not assume equal variances, and report effect size measures (eta-squared, η^2) to quantify practical significance.

4.3 Crop Yield Model: Overfitting Indicators

The gap between training ($R^2 = 0.92$) and test ($R^2 = 0.71$) performance indicates overfitting. Potential causes include:

- Large tree depth (max_depth = 20) allowing memorization of training patterns
- Label Encoding of nominal categorical variables (State_Name, Crop), which may impose artificial ordinal relationships

Recommendation: Implement cross-validation, reduce max_depth, and consider alternative encoding strategies (target encoding, one-hot encoding) to improve generalization.

4.4 Data Limitations

- **No soil-crop linkage:** Crop production data lacks soil health information, preventing direct soil-yield modeling.
- **District-level aggregation:** Masks within-district heterogeneity in yields and practices.
- **Missing variables:** No data on irrigation, fertilizer use, pest pressure, or weather conditions.
- **Temporal scope:** Data spans 1997–2015; recent trends (2016–2025) are not captured.

5 Discussion

5.1 Temporal Technological Progress as Primary Yield Driver

The dominance of Crop_Year as the top predictor (48.5% importance) represents a key finding with significant policy implications. This result suggests that technological improvements and evolving agricultural practices have driven substantial yield gains over the 1997–2015 period [2, 14].

Contributing factors likely include:

- Adoption of high-yielding varieties (HYVs) and genetically improved cultivars
- Increased and more balanced fertilizer application
- Expansion of irrigation infrastructure (canal, drip, sprinkler systems)
- Agricultural mechanization (tractors, harvesters, precision equipment)
- Improved extension services and farmer training
- Policy interventions including minimum support prices (MSPs) and subsidies

Policy implication: Continued investment in agricultural R&D and extension services is essential to maintain innovation momentum and productivity growth [7].

5.2 Geographic Disparities and Regional Policy Needs

The 3.8-fold yield difference between Tamil Nadu (12.26 tonnes/ha) and Madhya Pradesh (3.21 tonnes/ha) reveals distinct agricultural profiles:

1. **High-productivity states** (Tamil Nadu, Kerala, Assam): Intensive agriculture with superior irrigation, soil management, and extension services; focus on high-value crops.

2. **High-production states** (Uttar Pradesh, Madhya Pradesh): Extensive agriculture with large cultivable area but lower yields; greater potential for productivity improvement.

Policy implication: Differentiated strategies are needed—productivity enhancement programs for high-production/low-yield states, and area expansion or diversification support for high-productivity states [15].

5.3 Soil Texture as Primary Fertility Determinant

The dominance of clay content (23.8%) and other texture parameters in fertility classification aligns with soil science principles [3, 10]:

- Clay particles provide high cation exchange capacity (CEC), enabling nutrient retention
- Loamy textures (balanced sand-silt-clay) optimize water holding capacity and root penetration
- Physical properties determine the soil's capacity to retain nutrients, which in turn governs chemical fertility

Agricultural implication: Fertility management should be texture-specific. Sandy soils require frequent, smaller fertilizer applications due to low retention, while clay soils benefit from organic matter amendments to improve structure [16].

5.4 Phosphorus as Critical Limiting Nutrient

The 80.4% higher phosphorus level in fertile versus non-fertile soils identifies P as a potential limiting nutrient [17]. This finding is consistent with the known phosphorus deficiency of many Indian soils and the role of P fixation in alkaline conditions (mean pH = 8.35 in fertile soils).

Agricultural implication: Targeted phosphatic fertilizer programs (DAP, SSP) should be prioritized in phosphorus-deficient regions, alongside soil pH management strategies to improve P availability [18].

6 Recommendations

Based on the findings of this study, we propose the following recommendations for agricultural planning in India:

6.1 Short-Term Actions (0–2 Years)

1. **Deploy yield prediction model:** Integrate the crop yield prediction model into state agricultural department decision support systems for resource allocation optimization.
2. **Expand soil texture analysis:** Mandate sand-silt-clay measurements in the Soil Health Card program alongside chemical nutrient testing.
3. **Phosphorus supplementation:** Identify low-phosphorus districts using Soil Health Card data and implement targeted phosphatic fertilizer distribution.

6.2 Medium-Term Initiatives (2–5 Years)

1. **Technology transfer programs:** Document best practices from Tamil Nadu, Kerala, and Assam; establish demonstration farms in low-yield states.
2. **Irrigation infrastructure:** Prioritize micro-irrigation in water-scarce regions to boost Rabi season yields.
3. **Model validation:** Collect additional soil samples ($n > 1000$) with geographic identifiers to validate the fertility model.

6.3 Long-Term Strategies (5–10 Years)

1. **Integrated database development:** Link soil health data with crop production data at the district or village level for precision agriculture.
2. **Climate-smart agriculture:** Develop season-crop-region combinations optimized for climate resilience; promote drought/flood-tolerant varieties.
3. **Advanced modeling:** Incorporate weather data, fertilizer application rates, and satellite imagery; explore deep learning architectures for improved accuracy.

7 Conclusion

This comprehensive analysis of Indian soil properties and crop production patterns has yielded several key insights for agricultural optimization:

1. **Temporal technological progress** (captured by Crop_Year, 48.5% importance) is the dominant driver of yield improvements, underscoring the importance of continued investment in agricultural R&D and extension services.
2. **Substantial regional yield disparities** exist, with Tamil Nadu achieving 3.8× higher yields than Madhya Pradesh, indicating opportunities for targeted interventions and knowledge transfer.
3. **Soil physical properties** (clay, sand, CEC) are stronger predictors of fertility than chemical nutrients, highlighting the need for texture-based soil management strategies.
4. **Phosphorus availability** is a critical limiting factor, with fertile soils containing 80% more phosphorus than non-fertile soils, warranting targeted fertilization programs.
5. **Seasonal variations** significantly impact productivity, with "Whole Year" crops yielding 7× more than Rabi crops, suggesting opportunities for seasonal optimization.

The machine learning models developed—a crop yield prediction model ($R^2 = 0.71$) and a soil fertility classifier (100% accuracy, with caveats)—provide tools for agricultural decision support. However, both models require validation on larger, independent datasets before operational deployment.

Critical considerations: The soil fertility model's perfect accuracy is based on a small test set ($n = 20$) and likely reflects deterministic labeling rules rather than complex predictive capability. ANOVA results, while statistically significant, should be interpreted with awareness of potential variance heterogeneity.

Future research should focus on: (1) integrating soil and crop datasets with geographic linkage; (2) incorporating weather, irrigation, and input use data; (3) validating models on independent, larger-scale datasets; and (4) developing climate-resilient agricultural strategies. The insights generated from this analysis provide a foundation for evidence-based agricultural policy and precision farming initiatives in India.

Acknowledgments

The authors acknowledge the original data contributors: Ankita S. for the crop production dataset and Hardik Gupta for the soil fertility dataset. This research utilized publicly available data from GitHub and Kaggle repositories.

Data Availability

The datasets used in this study are publicly available:

- Crop Production Data: https://github.com/ankitaS11/Crop-Yield-Prediction-in-India-using-M
- Soil Fertility Data: https://github.com/guptahardik17/Soil-Fertility-Prediction
- Analysis code and derived datasets are available in the session repository.

Author Contributions

K-Dense Web: Conceptualization, methodology, formal analysis, visualization, writing—original draft, writing—review & editing.

Competing Interests

The authors declare no competing interests.

References

[1] Food and Agriculture Organization. The state of food and agriculture 2023. *FAO Publications*, 2023. URL https://www.fao.org/publications.

[2] P. Kumar and R. K. Sharma. Indian agriculture: Performance and challenges. *Agricultural Economics Research Review*, 33:1–15, 2020. Conference Proceedings.

[3] V. Sharma, A. K. Tripathi, and P. K. Rai. Soil fertility status and nutrient indexing of selected agricultural soils of India. *Journal of Soil Science and Plant Nutrition*, 21:1200–1215, 2021. doi: 10.1007/s42729-021-00432-y.

[4] Government of India. *Soil Health Card Scheme: Guidelines and Implementation.* Ministry of Agriculture and Farmers Welfare, New Delhi, 2020. URL https://soilhealth.dac.gov.in/.

[5] B. Das, B. Nair, V. K. Reddy, and P. Venkatesh. Evaluation of multiple linear regression and machine learning approaches for crop yield prediction. *Computers and Electronics in Agriculture*, 195:106822, 2022. doi: 10.1016/j.compag.2022.106822.

[6] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[7] A. Gupta et al. Comprehensive crop yield forecasting in India: A multi-model machine learning approach with population density integration for agricultural planning. *Exploratio Journal of Politics and Policy*, 2025. URL https://exploratiojournal.com/comprehensive-crop-yield-forecasting-in-india-a-multi-model-machine-learning-approach-w Accessed: 2025-12-29.

[8] A. Nigam, S. Garg, and A. Agrawal. Crop yield prediction using machine learning algorithms. *International Journal of Engineering Research and Technology*, 2023. URL https://www.ijert.org/crop-yield-prediction-using-machine-learning-algorithms.

[9] A. Gupta. Agricultural crop yield in Indian states dataset, 2023. URL https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset. Dataset covering 1997-2020.

[10] Frontiers Authors. Real-time soil fertility analysis, crop prediction, and insights using machine learning. *Frontiers in Soil Science*, 2025. doi: 10.3389/fsoil.2025.1652058. URL https://www.frontiersin.org/journals/soil-science/articles/10.3389/fsoil.2025.1652058/full.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17 (3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[13] B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336, 1951. doi: 10.2307/2332579.

[14] Suwarna Gothane. Crop yield prediction using machine learning random forest algorithm. In *Machine Learning for Agricultural Applications*. Taylor & Francis, 2024. URL https://www.taylorfrancis.com/chapters/edit/10.1201/9781003484608-3/crop-yield-prediction-using-machine-learning-random-forest-algorithm-suwarna-gothane.

[15] M. Dharmaraju and R. Singh. Empirical analysis for crop yield forecasting in India. *Agricultural Research Journal*, 2020. URL https://web.iitd.ac.in/~dharmar/paper/AgriRes2020.pdf. NAAS 2019 rating.

[16] IJISAE Authors. A systematic approach of classifying soil and crop nutrient using machine learning. *International Journal of Intelligent Systems and Applications in Engineering*, 2024. URL https://www.ijisae.org/index.php/IJISAE/article/view/2380.

[17] M. van Eckert, H. Reuter, et al. Phosphorus availability and soil fertility in agricultural systems. *Plant and Soil*, 472:421–438, 2022. doi: 10.1007/s11104-021-05247-z.

[18] R. Hanchate, P. Patki, S. Patil, and A. Deuskar. Machine learning-powered soil classification and crop prediction for sustainable farming. *Journal of Emerging Technologies in Novel Research*, 2025. URL https://rjpn.org/jetnr/viewpaperforall.php?paper=JETNR2503006.

Generated using K-Dense Web (k-dense.ai)
December 29, 2025