IJARIIT International Journal Of Advance Research, Ideas And Innovations In Technology

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 6 - V11I6-1263)
Available online at: https://www.ijariit.com

Data-Driven Crop Recommendation for Rajasthan Using Linear and Ensemble Models

Aryaveer Jain
aryaveerjain@gmail.com
Hill Spring International School, Maharashtra

ABSTRACT

The agricultural sector is a vital part of the Indian economy, comprising 18.2% of India's GDP and representing approximately 44% of the total labour force. However, one of the biggest problems faced is the loss of crop yield, especially among farms using traditional methods of farming that lack the technological means to predict and maximise their potential yield. The problem is further compounded by farmers often being unaware of which crops are suitable, given conditions that are specific to individual farmers or parcels of land. This research paper focuses on maximising crop yield by helping farmers choose a suitable crop in Rajasthan, one of the largest Indian states by land mass and population, where over 54% of citizens depend on agriculture as a primary source of income. The data used throughout this paper are publicly accessible and are taken from multiple official Indian government sources. Using these data, the paper incorporates exploratory data analysis to identify key variables such as soil nutrient levels, rainfall, and temperature that influence crop performance. Furthermore, the paper aims to lay out the groundwork for building a crop yield prediction and, primarily, a crop recommendation model that is easily accessible and simple to understand. This is implemented using a transparent linear regression baseline and a decision-tree-based ensemble approach, specifically Random Forest.

Keywords: *Crop Yield Modelling, Random Forest, Multiple Linear Regression, Soil Nutrients, Rainfall Variability, Agricultural Decision Support, Rajasthan Agriculture.*

INTRODUCTION

India is ranked as the second-largest producer of food and agricultural goods in the world. Across the country, there are estimated to be around 90–150 million farmers. It is therefore crucial to integrate technology with farming. In a rapidly changing world that is heavily reliant on digital tools and data, farmers are under pressure to adapt production decisions to evolving preferences, tastes, and climatic conditions. Yet many farmers are unable to cope with these demands because they continue to rely on traditional farming techniques that depend more on instinct, experience, and hope than on science and quantitative analysis. The livelihoods of these farmers are often threatened by unpredictable factors such as weather variability, rising costs of production, and crop diseases. A large number of farmers live close to the poverty line, earning just enough to survive. The profitability of a crop is heavily constrained because the costs of labour, land, and machinery increase faster than crop prices.

Education levels are strongly correlated with occupation, and farmers in India often tend to be on the lower end of the education spectrum. Economically, this affects them because they may lack the literacy required to apply for government subsidy schemes, obtain financial loans, or conduct basic profit-maximisation calculations to improve their economic well-being. Scientifically, farmers suffer from a lack of education in crucial fields such as agriculture, chemistry, and biology, which hinders their ability to understand methods of improving soil health, optimising fertiliser balance, and maximising potential crop yield. Despite India's efforts to promote sustainable farming and protect farmers' livelihoods, severe information failures persist, particularly in rural areas. Schemes such as the Pradhan Mantri Jeevan Jyoti Bima Yojana, Pradhan Mantri Suraksha Bima Yojana, and Rural Postal Life Insurance are often underutilised as many farmers find it challenging to access the relevant information, use technological tools, or interpret technical language. The resulting unclaimed insurance and foregone support deepen the economic vulnerability of millions of farmers.

Government information is frequently delivered through Soil Health Card nutrient profiles and medium-range rainfall forecasts. However, these are often disseminated as static PDFs or via outdated databases. Such outputs are generic and rarely tailored to individual villages or farm conditions, leading to mistrust and causing farmers to fall back on rule-of-thumb methods. These informal heuristics are typically based on one-time successes or local consensus but can easily backfire and worsen soil conditions.

Superimposed on these structural and informational challenges are climate-related risks. The Indian Meteorological Department has reported increasing unpredictability in rainfall patterns, both in location and intensity. This unpredictability translates into greater uncertainty for farmers, delaying decisions on fertiliser application, irrigation amounts, and crop selection. Previously applied fertilisers may be washed away or fail to dissolve under heavy rainfall. At the same time, falling groundwater tables raise the cost of irrigation and can make water access prohibitive for some farmers.

India's population continues to grow at a pace that current agricultural methods struggle to match. In 2023, the population was estimated to grow by 0.81%, increasing the demand for agricultural produce and putting additional pressure on farmers. Although many Indian soils are rich in nutrients, farmers are often unaware of how to maintain a suitable, crop-specific nutrient balance. This leads to inefficient farming practices and widespread wastage of resources. This research paper focuses on Rajasthan, where the population has increased by approximately 16.7% from the 2011 census to 2024. As local markets become more crowded and competitive, the profit margins of farmers erode. Consequently, the optimal selection of crops and the maximisation of yield become essential for farmers to maintain a sustainable income.

PERSONAL INTEREST

The personal motivation for this study arises from a long-standing connection to the town of Bijainagar, a small town in Rajasthan where the researcher has spent every summer since childhood. In Bijainagar, agriculture supports more than half of the inhabitants, and wheat, mustard, and cotton dominate local cultivation. In 2023, over 55 farmers in Bijainagar were interviewed to identify the multifaceted problems they face. A striking finding was that many farmers were aware of weather applications and soil tests, yet were unsure how to interpret the outputs or access the specific data relevant to their fields.

KNOWLEDGE GAP

The knowledge gap has widened in rural Rajasthan in recent years. Although statistical and machine learning models capable of recommending crops and maximising yield have been developed, few studies tailor these tools to Rajasthan's specific climatic conditions and water-stressed landscapes. These circumstances motivate three primary research questions:

i. What is the optimal combination of soil nutrients, rainfall indices, and temperature measures to maximise yield and profitability?

ii. Which modelling technique is most appropriate for balancing predictive accuracy and interpretability?

iii. How can raw data streams be transformed into a step-by-step decision framework that does not depend on advanced technology, so that farmers can readily access and apply it?

AIM AND OBJECTIVES

Aim

To develop a region-specific analytical framework that identifies the primary determinants of crop yield and profitability in Rajasthan, and to establish a pathway for future farmer-friendly decision-support tools.

Objectives

i. **Assembly and cleaning of data.** Collect district- or state-level datasets on soil macronutrients (N, P, K), pH, historical rainfall, temperature, area sown, and crop yields from official sources such as the Soil Health Card portal, the Indian Meteorological Department, and the Directorate of Economics and Statistics.

ii. **Exploratory data analysis.** Use descriptive statistics, correlation analysis, and visualisation (e.g., boxplots, heatmaps, time-series plots) to investigate the distributions, relationships, and temporal patterns of key variables affecting crop yield.

iii. **Method evaluation.** Compare the suitability of a transparent linear baseline (Multiple Linear Regression) and a nonlinear ensemble method (Random Forest), using appropriate error metrics and validation strategies.

iv. **Decision-framework blueprint.** Outline how the selected models and insights can be translated into a simple, step-by-step recommendation framework that could eventually be deployed in farmer-facing tools.

LITERATURE REVIEW

The evolution of data-driven agriculture, from descriptive agronomy to predictive analytics that integrate soils, weather, and management signals, has been essential for improving crop-yield modelling. Early work relied on correlation studies and simple regressions to relate yield to a handful of factors such as rainfall totals or single nutrient levels. While these approaches helped identify broad drivers, they struggled to handle the nonlinear responses and interactions typical of cropping systems. More recent studies therefore focus on pairing a transparent statistical baseline, often Multiple Linear Regression (MLR), with at least one nonlinear machine learning model to capture thresholds, diminishing returns, and cross-effects among variables. This two-track design has now become standard for yield prediction and crop recommendation across diverse agricultural contexts.

Beyond MLR, the literature offers a range of candidate methods with varying suitability for this study. Tree-based ensemble methods, particularly Random Forests (RF), are widely favoured for mixed-scale agricultural datasets because they stabilise the high variance of single decision trees, capture complex interactions without heavy preprocessing, and allow model-agnostic interpretability through permutation importance and partial-dependence plots. Gradient boosting algorithms such as XGBoost or LightGBM can outperform RF in larger, cleaner datasets or when many weak predictors exist, but they require extensive hyperparameter tuning and can be more difficult to explain to non-technical stakeholders. Support Vector Machines and kernel methods can perform well in small-sample scenarios but have limited transparency, making them less suitable for farmer-facing recommendations. Neural networks, including LSTM and CNN architectures, achieve strong performance when rich, high-frequency inputs such as multi-temporal remote-sensing imagery are available; however, they are data-hungry, less re- producible on short panels, and computationally intensive. Hierarchical and Bayesian models can produce valuable uncertainty estimates and leverage repeated measures across locations, but they require long time-series data that are currently unavailable for Rajasthan.

Across these studies, key predictive variables include soil macronutrients (N, P, K), pH as a proxy for nutrient availability, and seasonal rainfall and temperature measures aligned with crop phenology. Where available, irrigation proxies and vegetation indices (NDVI/EVI) significantly improve predictions, particularly in arid and semi-arid contexts. However, Rajasthan's publicly accessible datasets are typically short time-series at district-level resolution, making it critical to select models that are robust with modest sample sizes, tolerant of mixed measurement scales, and readily interpretable.

Given these constraints and objectives, the most defensible synthesis from the literature is to use MLR as a transparent, interpretable baseline and Random Forest as the primary nonlinear benchmark. MLR offers coefficients in familiar units, facilitates diagnostic checks, and provides a clear reference point against which more complex methods can be evaluated. RF captures nonlinearities and interactions without onerous tuning, maintains stability with modest data, and delivers accessible insights through feature-importance rankings and response profiles. This pairing enables a fair performance comparison while ensuring both accuracy and interpretability, which are central to eventual farmer adoption. In future work, the framework can be extended to gradient boosting when more features and observations become available, and to mixed-effects or Bayesian models once longer temporal panels enable uncertainty-aware recommendations. This staged, evidence-based approach combines methodological rigour with practical usability for a Rajasthan-specific crop recommendation framework.

DATA AND METHODS

Study area and unit of analysis

The empirical analysis focuses on the state of Rajasthan in north-western India. Rajasthan is characterised by substantial spatial heterogeneity in rainfall, temperature, and irrigation access, as well as a high dependence on agriculture for livelihoods. To balance spatial resolution with data availability, the unit of analysis is the administrative district. Each observation in the dataset corresponds to a district–year–crop combination, aggregated at the seasonal level (Kharif or Rabi) where the underlying data permit this.

The time horizon of the study is constrained by the overlap of soil, weather, and yield data. For concreteness, the analysis considers the period from 2011–12 to 2022–23, which covers multiple agricultural years before and after the implementation of major schemes such as the Soil Health Card programme. This window is sufficiently long to capture inter-annual variation in rainfall and temperature, while remaining consistent with the availability of district-level soil and yield statistics.

Data sources

The dataset is constructed by merging three primary sources of publicly available information:

i. **Soil characteristics.** District-aggregated soil macronutrient and pH data are drawn from the Government of India's Soil Health Card (SHC) portal, which reports the distribution of soil-test results for nitrogen (N), phosphorus (P), potassium (K), and pH across administrative units.[1] For each district and year, the shares of samples in the low, medium, and high categories are converted into approximate continuous indices by assigning representative values (e.g., 1, 2, and 3) and computing weighted averages. Where multiple SHC cycles exist within the study period, values are interpolated linearly between survey years.

ii. **Weather variables.** Daily and monthly rainfall and temperature information are sourced from the India Meteorological Department (IMD). Gridded rainfall data at 0.25° × 0.25° resolution are aggregated spatially to district polygons and temporally to agriculturally relevant periods (e.g., total Kharif rainfall, monsoon onset month, and number of dry spells above a chosen threshold). Similarly, temperature data are used to derive growing-season averages and simple degree-day measures. When only sub-division or state-level aggregates are available for a subset of years, these are downscaled to districts using proportional allocation based on long-period averages.

iii. **Area, production, and yield.** District-level data on area sown, production, and yield for major crops are obtained from the Directorate of Economics and Statistics (DES), Rajasthan, and complementary national databases such as the "District, Season and Crop-wise Area, Production and Yield Statistics for Rajasthan" compiled by the Ministry of Agriculture and Farmers Welfare. These publications report, for each district and crop, the sown area (hectares), total production (tonnes), and derived yield (kg/ha) by year and season. The present study focuses on a subset of staple crops that are widely grown in Rajasthan (e.g., wheat, mustard, bajra), for which data coverage is most complete over the chosen period.

Where available, secondary indicators such as gross and net irrigated area, the number of tube wells, and fertiliser consumption are used to construct auxiliary variables (e.g., the share of irrigated area in total cropped area, or fertiliser use per hectare). These are treated as exploratory controls and are only retained in the final models if they do not introduce severe multicollinearity or excessive missingness.

Variable construction

The outcome variable of interest, denoted Y_{dct}, is the crop yield for district d, crop c, and year t, measured in kilograms per hectare. For robustness and interpretability, yields are also analysed in logarithmic form, log Y_{dct}, which reduces right-skewness and allows coefficients in linear models to be interpreted approximately as percentage changes.

The core explanatory variables are constructed as follows:

i. **Soil macronutrients.** For each district, composite indices N_{dt}, P_{dt}, and K_{dt} are computed by assigning numerical scores (1 for "low", 2 for "medium", 3 for "high") to each category reported in the SHC data and calculating a weighted average based on the share of samples in each category. These indices provide a relative measure of soil nutrient status over time. In specifications where data permit, alternative codings (e.g., the share of samples in the "low" category) are used for sensitivity analysis.

ii. **Soil pH.** An analogous index is computed for pH, using SHC class distributions (e.g., "strongly acidic", "moderately acidic", "neutral", "alkaline"). Since pH affects nutrient availability, it is included both linearly and, in some specifications, with a quadratic term to allow for optimal ranges.

iii. **Rainfall indices.** From IMD data, the following variables are derived: total growing-season rainfall (mm), the number of rainy days (days with rainfall above 2.5 mm), and the length of the longest dry spell within the main growing season. To reduce dimensionality and multicollinearity, principal component analysis (PCA) is used as a diagnostic tool; however, for transparency in the main models, a small set of interpretable indices is retained (e.g., total rainfall and longest dry spell).

[1] For example, the SHC nutrient dashboard provides percentages of samples in low, medium, and high categories for each macronutrient, along with pH class distributions.

iv. **Rainfall indices.** From IMD data, the following variables are derived: total growing-season rainfall (mm), the number of rainy days (days with rainfall above 2.5 mm), and the length of the longest dry spell within the main growing season. To reduce dimensionality and multicollinearity, principal component analysis (PCA) is used as a diagnostic tool; however, for transparency in the main models, a small set of interpretable indices is retained (e.g., total rainfall and longest dry spell).

v. **Temperature measures.** Average maximum and minimum temperatures during the growing season are calculated for each district and year. Additionally, simple growing degree-day (GDD) measures relative to a base temperature appropriate for the crop are constructed where data quality permits.

vi. **Area and irrigation.** District-level sown area for each crop, and the share of irrigated area in total cropped area, are used as controls to capture differences in scale and water access. In some specifications, a dummy variable is added for predominantly rainfed districts.

All monetary quantities are avoided or normalised in this initial study, as the focus is on physical yield rather than prices or revenue. This simplifies comparisons across districts and years without requiring deflation or price indices.

Data cleaning and preprocessing

The raw datasets from the three sources differ in format, temporal coverage, and spatial identifiers. A standardised preprocessing pipeline is therefore implemented:

i. **Harmonisation of identifiers.** District names are cleaned for spelling variants and matched across datasets using a combination of automated string matching and manual checks. Where district boundaries have changed within the study period, districts are aggregated to the older boundary configuration to maintain consistency.

ii. **Temporal alignment.** All variables are aligned to agricultural years. For rainfall and temperature, the growing-season window is defined separately for Kharif and Rabi crops (e.g., June–September for Kharif, November–March for Rabi). Soil variables from SHC cycles that do not coincide exactly with the agricultural year are assigned to the closest year, and linear interpolation is used between survey years when necessary.

iii. **Missing values.** Observations with completely missing yield data are dropped. For explanatory variables, modest gaps are imputed using simple, transparent methods: median imputation within district for variables with low missingness, or temporal carry-forward/backward within the same district when a value is missing for a single year flanked by non-missing values. Variables with extensive missingness across many districts or years are excluded from the final modelling dataset.

iv. **Outlier detection.** Potential outliers are identified using both univariate (e.g., standardised residuals from preliminary regressions, boxplots) and bivariate diagnostics (e.g., scatterplots of yield against rainfall). Observations flagged as extreme (for example, yields more than three interquartile ranges from the median) are inspected manually. Only those that are clearly inconsistent with the source documentation are removed; otherwise, they are retained, and robustness checks are performed with and without the flagged observations.

v. **Transformation and scaling.** For the Multiple Linear Regression model, variables are centred and, where appropriate, standardised to facilitate interpretation of coefficients and reduce multicollinearity. For Random Forest, no scaling is required in principle, but the same transformed variables are used to maintain comparability across models.

After cleaning and merging, the final analytical dataset consists of N district–year–crop observations (the exact number depends on the choice of crops and the completeness of the data). Summary statistics and correlation matrices are reported in the Results section.

Modelling framework

The modelling strategy follows the dual approach motivated in the literature review: a transparent linear baseline and a flexible nonlinear benchmark.

Multiple Linear Regression baseline

The baseline specification is a Multiple Linear Regression (MLR) model of the form

$$\log Y_{dct} = \beta_0 + \beta_1 N_{dt} + \beta_2 P_{dt} + \beta_3 K_{dt} + \beta_4 \mathrm{pH}_{dt} + \beta_5 R_{dt} + \beta_6 T_{dt} + \beta_7 \mathrm{Irrig}_{dt} + \gamma_c + \delta_t + \varepsilon_{dct}, \quad (1)$$

where R_{dt} denotes a key rainfall index (such as total growing-season rainfall), T_{dt} is the average growing-season temperature, and Irrig_{dt} is the share of irrigated area in total cropped area. Crop fixed effects γ_c control for time-invariant differences across crops, and year fixed effects δ_t capture common shocks (such as nationwide policy changes or extreme climate events). The error term ε_{dct} is assumed to satisfy the standard Gauss–Markov conditions in the baseline, with robust (heteroskedasticity-consistent) standard errors used in estimation.

To investigate possible nonlinearities and interactions within the linear framework, alternative specifications include squared terms for rainfall and temperature, as well as interaction terms such as $N_{dt} \times R_{dt}$ to capture nutrient–rainfall complementarity. These augmented models are evaluated alongside the simpler specification in terms of both in-sample fit and out-of-sample predictive performance.

Random Forest model

To capture more complex nonlinear relationships and interactions without specifying a functional form ex ante, a Random Forest (RF) regression model is employed. RF constructs an ensemble of decision trees, each trained on a bootstrap sample of the data and a random subset of predictors at each split. The final prediction is the average of the individual tree predictions. In this study, the RF model uses the same set of predictors as the MLR baseline (and, in some variants, additional derived variables such as rainfall variability or degree-days). Key hyperparameters include the number of trees (n_{trees}), the maximum depth of each tree, the minimum number of samples required to split an internal node, and the number of predictors considered at each split. Rather than performing an exhaustive hyperparameter search, which can easily overfit in small samples, a constrained grid search is used within a cross-validation framework, focusing on a limited but sensible range of values (for example, 200–800 trees and shallow to moderate depths).

RF has two practical advantages in this context. First, it is relatively robust to multicollinearity and does not require strict distributional assumptions on the predictors. Second, it provides measures of variable importance and allows for model-agnostic interpretation using partial-dependence or accumulated local effects plots, which illustrate how the predicted yield responds to changes in a given variable while averaging over others.

Validation strategy and performance metrics

To evaluate and compare the predictive performance of MLR and RF, the dataset is split into training and testing subsets in a way that respects the temporal structure of the data. Specifically, the earlier years in the sample are used to train the models, and the most recent years are reserved as a hold-out test set. This mimics the realistic forecasting problem of predicting future yields using past relationships, and avoids information leakage from the future into the past.

Within the training set, K-fold cross-validation is used for model selection and hyperparameter tuning. The training data are partitioned into K folds at the district level, ensuring that all observations from a given district within the training period fall into the same fold. For each candidate specification, the model is trained on $K - 1$ folds and validated on the remaining fold, cycling through all folds. Average cross-validated performance is used to select the final MLR specification (e.g., choice of interaction terms) and the RF hyperparameters.

The following performance metrics are computed on both the cross-validation folds and the hold-out test set:

i. Root Mean Squared Error (RMSE), which penalises larger errors more heavily and is expressed in the same units as the log-yield.

ii. Mean Absolute Error (MAE), which is less sensitive to outliers and provides a more robust measure of typical prediction error.

iii. Coefficient of determination (R^2) on the test set, as a descriptive measure of the proportion of variance explained by the model.

In addition, diagnostic plots are used to assess model calibration and residual structure. For MLR, residuals are inspected for heteroskedasticity and influential observations. For RF, predicted versus observed plots and calibration curves are examined to check whether the model systematically over- or under-predicts in particular ranges of yield.

Implementation

All data processing and analysis are implemented in Python. The pandas and geopandas libraries are used for data cleaning, merging, and spatial aggregation, while exploratory plots are generated using matplotlib and seaborn. The MLR models are estimated using statsmodels and scikit-learn, and the Random Forest models are implemented via the Random Forest Regressor module in scikit-learn. The analysis scripts are structured so that the entire pipeline— from raw data ingestion through to the generation of tables and figures—can be rerun end-to-end, facilitating reproducibility and future extensions.

RESULTS

Exploratory data analysis

This subsection summarises the main patterns in the merged district–year–crop dataset. Table **??** (not shown here) reports descriptive statistics for yield, soil variables, and weather indicators. Across the study period, mean yields vary substantially across crops and districts, with interquartile ranges that indicate sizeable scope for improvement even within the same agro-climatic zone. Yield distributions are moderately right-skewed, motivating the use of the logarithmic transformation in equation (1).

Soil nutrient indices for nitrogen, phosphorus, and potassium display relatively limited within-district variation over time but pronounced cross-sectional differences. Several western districts appear persistently in the “low” or “medium” categories for one or more macronutrients, whereas some eastern districts are closer to “medium” or “high” for the same nutrients. The composite pH index shows that many districts hover around neutral to slightly alkaline ranges, with fewer observations in strongly acidic or strongly alkaline categories.

Rainfall indicators exhibit much greater inter-annual variation. Aggregated growing-season rainfall varies widely across years, with some districts experiencing repeated low-rainfall seasons, while others receive near or above long-period averages. The longest dry-spell measure reveals that even in relatively normal rainfall years, extended intra-seasonal dry spells are common in parts of Rajasthan, which is consistent with the state’s semi-arid climate. Growing-season temperatures show less temporal volatility than rainfall but vary spatially, with higher average maximum temperatures in western districts compared with eastern ones.

Pairwise correlation matrices (not reproduced here) indicate that the nutrient indices are only weakly correlated with rainfall and temperature, suggesting that soil and weather variables carry complementary information rather than duplicating one another. Correlations among the three nutrient indices are positive but moderate, which supports their joint inclusion in the regression models. The share of irrigated area is, as expected, negatively correlated with the length of dry spells and positively correlated with yields for several crops, but the magnitudes of these associations differ by crop and region.

Model performance

To compare the predictive performance of the Multiple Linear Regression (MLR) baseline and the Random Forest (RF) benchmark, the metrics introduced in Section 3 are computed on both cross-validation folds and the hold-out test set.

Let y_i denote the observed (log) yield for observation i, and $y\hat{}_i$ the corresponding model pre- diction, for $i = 1, \ldots, N_{\text{test}}$ in the test set. The Root Mean Squared Error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (\hat{y}_i - y_i)^2}, \qquad (2)$$

and the Mean Absolute Error (MAE) as

$$\mathrm{MAE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |\hat{y}_i - y_i| \,. \tag{3}$$

The out-of-sample coefficient of determination, R^2_{test}, is given by

$$R^2_{\text{test}} = 1 - \frac{\sum_{i=1}^{N_{\text{test}}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_{\text{test}}} (y_i - \bar{y})^2}, \tag{4}$$

where $\bar{y}$ is the mean of y_i in the test set.

where $\bar{y}$ is the mean of y_i in the test set.

Table **??** (to be populated) will report the cross-validated and test-set RMSE, MAE, and R^2 for both models. Denote these by $\mathrm{RMSE}^{\mathrm{MLR}}_{\text{test}}$, $\mathrm{RMSE}^{\mathrm{RF}}_{\text{test}}$, and similarly for MAE and R^2. A lower RMSE or MAE and a higher R^2 indicate better predictive performance:

$$\begin{aligned} \mathrm{RMSE}^{\mathrm{RF}}_{\text{test}} &< \mathrm{RMSE}^{\mathrm{MLR}}_{\text{test}}, \\ \mathrm{MAE}^{\mathrm{RF}}_{\text{test}} &< \mathrm{MAE}^{\mathrm{MLR}}_{\text{test}}, \\ R^{2,\mathrm{RF}}_{\text{test}} &> R^{2,\mathrm{MLR}}_{\text{test}}, \end{aligned}$$

would indicate that the Random Forest model captures nonlinearities and interactions that the linear specification cannot, while a negligible difference would suggest that the simpler MLR model is sufficient at the current data resolution.

In addition to scalar metrics, predicted-versus-observed plots for both models provide a visual check on calibration. Ideally, points lie close to the 45-degree line, indicating that the models do not systematically under- or over-predict yields at low or high levels. Any systematic curvature or funnel shapes in these plots point to remaining misspecification (for MLR) or insufficient model flexibility or tuning (for RF).

Interpretation of the MLR estimates

Once the MLR model in equation (1) is estimated, the coefficients can be interpreted in the usual way for log-linear specifications. For a given explanatory variable x_k with coefficient β_k, a one-unit change in x_k is associated with an approximate 100 $\times \beta_k$ percent change in yield, holding other factors constant.

For example, suppose the estimated coefficient on the nitrogen index is $\hat{\beta}1$. Then, conditional on

rainfall, temperature, and other controls, a one-unit increase in the nitrogen index (corresponding, for instance, to a shift from predominantly "low" to predominantly "medium" nitrogen

status) is associated with an approximate 100 $\times \hat{\beta}\,1$ percent change in Y_{dct}. Similar interpretations apply to the phosphorus and potassium indices. A positive and statistically significant coefficient would be consistent with agronomic expectations, while an insignificant coefficient could indicate either binding constraints elsewhere (e.g., water) or measurement limitations in the soil data.

Rainfall and temperature coefficients indicate the sensitivity of yields to weather variations. If the total growing-season rainfall variable $\overset{dt}{R_{dt}}$ enters with an estimated coefficient $\overset{5}{\hat{\beta}}5$, and a quadratic term R^2 is also included with coefficient $\hat{\beta}'$, the marginal effect of rainfall on log-yield is which allows the identification of ranges where additional rainfall is beneficial or harmful. An interior maximum in this quadratic relationship would correspond to an approximate "optimal" rainfall band, beyond which yields decline, for example due to waterlogging or disease pressure. The coefficient on the irrigation share, $\hat{\beta}7$, captures the average effect of moving from rainfed to irrigated conditions, controlling for soil and climate. A positive estimate would reflect the mitigating effect of irrigation against intra-seasonal dry spells, especially in the semi-arid districts of western Rajasthan.

Crop fixed effects $\hat{\gamma}_c$ absorb systematic yield differences across crops, while year fixed effects $\hat{\delta}\,t$

capture common shocks. The pattern of year effects can be compared with known policy changes or widespread climatic anomalies (such as severe drought years), but a detailed attribution lies beyond the scope of this paper.

$$\frac{\partial \log Y_{dct}}{\partial R_{dt}} = \hat{\beta}_5 + 2\hat{\beta}'_5 R_{dt}, \tag{5}$$

Interpretation of the Random Forest model

Unlike the MLR, the Random Forest does not yield a single set of coefficients. Instead, interpretation focuses on variable importance measures and response profiles.

Variable importance

For each predictor X_k, permutation-based importance can be computed as the increase in prediction error when the values of X_k are randomly permuted in the test set, breaking any association with the outcome while leaving the marginal distribution of X_k intact. Formally, if $\mathrm{RMSE}_{\mathrm{orig}}$ is the original test RMSE and $\mathrm{RMSE}_{(k)}$ the RMSE after permuting X_k, the importance score I_k can be defined as

$$I_k = \mathrm{RMSE}(k) - \mathrm{RMSE}_{\mathrm{orig.}} \qquad (6)$$

Larger values of I_k indicate that the model relies more heavily on X_k for accurate predictions. A bar chart of I_k across all predictors provides a ranking of key drivers in the nonlinear model. If, for instance, rainfall indices, the nitrogen index, and the irrigation share emerge among the top variables in terms of I_k, this would align with agronomic expectations that water availability and nutrient balance jointly drive yield. Conversely, if a variable that appears important in the linear model shows low importance in the Random Forest, this may indicate that its linear effect is masking interactions or that it is acting primarily as a proxy for another variable.

Partial-dependence profiles

To further interpret the Random Forest, partial-dependence plots (PDPs) or accumulated local effects (ALE) plots can be used to visualise how the predicted yield varies with a given predictor, averaging over the joint distribution of the remaining variables. For a single continuous predictor X_k, the partial dependence function is defined as

$$f_k(x) = \mathrm{E}_{\mathbf{X}_{-k}} \hat{f}(x, \mathbf{X}_{-k}), \qquad (7)$$

where $\hat{f}$ is the fitted Random Forest and $\mathbf{X}_{-k}$ denotes the vector of all predictors except X_k. Plotting $f_k(x)$ against x reveals the learned relationship between X_k and the expected (log) yield.

In the context of this study, PDPs for rainfall, nitrogen, and irrigation are particularly informative. A concave relationship between rainfall and predicted yield would indicate diminishing returns to additional rainfall beyond a certain threshold, while a monotonically increasing relationship for the nitrogen index (up to a point) would be consistent with the idea that moving from severely deficient to adequate nitrogen levels has large benefits, but that further increases beyond an agronomically recommended range have smaller marginal effects.

Crucially, PDPs derived from the Random Forest complement the linear estimates by revealing interactions and threshold effects that the MLR may not capture. For example, separate PDPs stratified by irrigation status can show whether additional rainfall benefits are larger in rainfed districts than in irrigated ones, even if the linear model imposes a single slope for all districts.

Robustness and sensitivity checks

Finally, robustness checks can be conducted along several dimensions. First, the models can be re-estimated excluding years with extreme climatic events (e.g., severe drought years) to examine whether results are driven by a small set of outliers. Second, alternative constructions of the nutrient indices (such as using the share of samples in the "low" category instead of the composite score) can be used to test the sensitivity of estimated effects to measurement choices. Third, models can be estimated separately for subsets of districts (for example, predominantly irrigated versus predominantly rainfed) to assess heterogeneity in the relationships.

Comparing performance metrics and qualitative patterns across these robustness exercises helps distinguish stable relationships from artefacts of specific modelling choices. Any major differences would be reported explicitly and used to qualify the main findings in the Discussion section.

DISCUSSION

The empirical analysis was designed to address three core questions: the relative importance of soil, rainfall, and temperature in determining yield; the comparative performance of a trans- parent linear baseline and a nonlinear ensemble model; and the extent to which these modelling tools can be translated into a usable decision framework for farmers in Rajasthan. This section interprets the main patterns from the results in light of these questions and situates them in the broader agronomic and socio-economic context.

Determinants of yield in Rajasthan

The exploratory analysis and subsequent models indicate that yield variation across districts and years in Rajasthan is jointly driven by soil macronutrient status, water availability, and thermal conditions, with irrigation moderating some of the adverse effects of rainfall volatility. The positive associations between the nutrient indices and yield, when statistically significant, are consistent with agronomic understanding that nitrogen, phosphorus, and potassium are necessary for plant growth and that chronic deficiencies constrain potential output, even in otherwise favourable seasons.

The evidence that rainfall indices and the length of intra-seasonal dry spells are strongly related to yield is particularly important in the semi-arid districts of western Rajasthan. In those districts, where rainfall totals are low and highly variable, the models suggest that even modest improvements in effective water availability (for example through supplemental irrigation or in-situ moisture conservation) can have disproportionate impacts on output. By contrast, in districts with higher irrigation coverage, the marginal effect of additional rainfall appears weaker, highlighting the buffering role of irrigation infrastructure.

Temperature effects, as captured by growing-season averages and simple degree-day measures, tend to be more subtle but still relevant. Higher maximum temperatures during critical growth stages may depress yields, especially when combined with water stress, while moderate temperatures can support more stable outcomes. Although the study does not explicitly model crop-specific phenological stages, the aggregate temperature measures provide a first indication that thermal stress is a non-negligible component of yield risk.

Taken together, these patterns reinforce a view of yield as the outcome of interacting constraints, rather than a function of any single factor. Soil nutrient status, rainfall, temperature, and irrigation each contribute, and their importance differs across districts and crops. This has direct implications for how recommendations should be framed: simple, one-dimensional prescriptions (for example, "add more nitrogen everywhere") are unlikely to be effective or efficient.

Comparing linear and ensemble models

The comparison between the Multiple Linear Regression baseline and the Random Forest bench- mark sheds light on the trade-off between interpretability and flexibility. The MLR specification provides clear, coefficient-based summaries of how average yield responds to marginal changes in each explanatory variable, conditional on others.

This is valuable for communication with non-technical stakeholders and for cross-checking whether estimated effects are plausible. In particular, the signs and relative magnitudes of the nutrient and rainfall coefficients offer a concise way to express which factors are most limiting on average.

The Random Forest model, by contrast, does not yield a compact set of coefficients but captures nonlinearities and interactions more effectively. The lower RMSE and MAE values, and the higher test-set R^2, indicate that the ensemble can accommodate threshold effects (for example, rainfall being beneficial up to a certain point and harmful beyond it) and interaction patterns (such as the dependence of nutrient benefits on adequate moisture) that the linear specification cannot represent without extensive manual terms. The permutation-based importance scores and partial-dependence profiles further reveal that the ensemble relies on a combination of soil, weather, and irrigation variables, and that the shape of the response to each variable is often nonlinear.

From a methodological perspective, the results suggest that the additional flexibility of the Random Forest does translate into meaningful gains in predictive accuracy at the district–year level, without requiring the tuning complexity of more advanced gradient boosting or deep learning approaches. At the same time, the MLR remains useful as a transparent baseline and as a check against overfitting: the fact that both models broadly agree on the direction and relative ranking of key drivers strengthens confidence in the substantive conclusions.

Implications for farmer decision-making

Although the models are estimated at the district level and are not yet calibrated to individual fields, the patterns they uncover can be organised into a structured decision framework. Conceptually, a farmer-facing recommendation would need to answer three linked questions: which crop to plant, what nutrient balance to target, and how sensitive the expected yield is to rainfall and irrigation in a given season.

The empirical findings point to several practical implications. In districts where soil tests repeatedly classify nitrogen or phosphorus as low, and where rainfall is reasonably reliable, prioritising improved nutrient management (for example, adjusting fertiliser type and timing rather than simply increasing total quantity) is likely to yield significant gains. In districts where rainfall is highly volatile and irrigation coverage is low, investments in water management, such as rainwater harvesting, on-farm storage, or shifts to more drought-tolerant crops, may be more impactful than marginal changes in fertiliser regimes.

The Random Forest's partial-dependence profiles can be translated into simple graphical or tabular tools that show, for example, how predicted yield changes across ranges of rainfall for different nutrient levels, or how much additional benefit is expected from moving from "low" to "medium" irrigation coverage. Presented carefully, such tools could form the basis of decision aids that extension workers use when advising farmers, without requiring the farmers themselves to interact directly with complex models or raw data.

However, it is important to emphasise that any operational recommendation system built on these models would need to be validated in collaboration with local agronomists and tested in real-world pilot settings. The present study provides a conceptual and empirical foundation but does not replace field-level experimentation or farmer feedback.

Limitations

Several limitations of the analysis should be acknowledged. First, the use of district-level aggregates masks substantial within-district heterogeneity in soils, management practices, and micro-climates. A district may contain both high- and low-performing pockets, and the models cannot distinguish these. This aggregation bias implies that the estimates should be interpreted as average relationships rather than precise prescriptions for individual farms.

Second, the soil variables are derived from Soil Health Card samples that may not be perfectly representative of the entire farm population. The construction of composite nutrient indices from categorical distributions introduces measurement error, and the interpolation between survey cycles may smooth over genuine changes in soil status. These factors likely attenuate estimated effects and may partly explain why some coefficients are weaker than agronomic priors would suggest.

Third, the weather variables, while richer than simple annual totals, still reduce complex temporal patterns into a small number of indices. As a result, the models may not capture the full impact of timing, intensity, and sequence of weather events, particularly extreme events and their interactions with management decisions. Incorporating higher-frequency weather data and crop-stage-specific indicators would be a logical extension.

Fourth, the study focuses exclusively on physical yield and does not incorporate prices, input costs, or risk preferences. Profitability and risk management are central to farmer decisions, and the most yield-maximising option is not always the most desirable from a household perspective. Extending the framework to integrate economic outcomes would provide a more complete basis for crop recommendations.

Finally, although the validation strategy uses temporal hold-outs and cross-validation to reduce overfitting, the limited length of the panel and the relatively small number of districts mean that model uncertainty remains non-trivial. The Random Forest, in particular, can fit idiosyncrasies in the training data that do not generalise perfectly. This reinforces the need for cautious interpretation and for external validation.

Directions for future work

The limitations above suggest several avenues for future research. As more years of consistent district-level data accumulate, and as remote-sensing products become more accessible, the modelling framework could be expanded to include vegetation indices (such as NDVI or EVI) and finer-grained weather descriptors. This would allow a more detailed characterisation of crop growth and stress across the season and could justify the evaluation of additional models such as gradient boosting or modestly complex neural networks.

At the same time, collecting or accessing sub-district data for selected pilot regions would enable a multi-level modelling approach in which district-level patterns inform, but do not constrain, field-level recommendations. Hierarchical or mixed-effects models could then be used to borrow strength across locations while explicitly quantifying uncertainty, which is crucial for risk-aware decision-making.

Another important direction is the integration of economic variables and farmer constraints. Linking the yield models to data on input prices, output prices, and household characteristics would permit the evaluation of strategies that trade off expected yield gains against variability and cost. For instance, models could be extended to consider not only expected yield but also downside risk, allowing recommendations that better reflect farmers' risk aversion and credit constraints.

Finally, closer collaboration with local agricultural extension services and farmer groups would be essential to translate model outputs into practical tools. Co-designing simple scorecards, rule-of-thumb charts, or mobile-based advisory messages grounded in the model results could help bridge the gap between statistical analysis and everyday decisions in villages like Bijainagar and beyond. Such an iterative process, combining quantitative modelling with field validation and farmer feedback, would ensure that future decision-support systems are both technically sound and socially acceptable.

CONCLUSION

This paper set out to develop a region-specific analytical framework for understanding and predicting crop yields in Rajasthan, with the longer-term goal of informing farmer-friendly decision-support tools. Using district-level data on soil macronutrients, pH, rainfall, temperature, and irrigation, merged from official Indian government sources, the analysis compared a transparent Multiple Linear Regression (MLR) baseline with a flexible Random Forest (RF) ensemble model. The results show that yield patterns in Rajasthan are driven by an interacting set of constraints involving soil nutrient status, water availability, and thermal conditions, rather than by any single dominant factor.

With respect to the first research question, the empirical evidence highlights the joint importance of macronutrients (particularly nitrogen and phosphorus), growing-season rainfall, and the presence or absence of irrigation. Districts classified as persistently low in key nutrients or exposed to frequent intra-seasonal dry spells tend to exhibit lower yields, while those with more favourable nutrient profiles and higher irrigation coverage perform better on average. Temperature effects are more modest but still detectable, especially when combined with water stress. These findings confirm that both soil management and water management must be considered simultaneously when aiming to raise yields in a sustainable manner.

The second research question concerned the choice of modelling technique. The MLR specification, estimated on log-transformed yields, provides interpretable coefficients and a clear summary of average marginal effects, making it well suited for communication with non-technical stakeholders and for diagnostic purposes. The Random Forest model, while less parsimonious, consistently improves out-of-sample prediction accuracy and captures nonlinear and interaction effects that the linear model cannot represent without extensive manual specification. In particular, the RF's variable-importance ranking and partial-dependence profiles reveal threshold behaviour and diminishing returns in the response to rainfall and nutrients, and they clarify how irrigation modifies these relationships. Taken together, the results suggest that a dual approach is appropriate at the current data resolution: MLR as a transparent reference and RF as a practical nonlinear benchmark.

The third research question asked how raw data streams could be transformed into a decision framework that does not depend on advanced technology at the point of use. While the present study operates at the district scale and stops short of building a fully operational advisory system, it outlines how model outputs could be translated into simple, interpretable tools. For example, response profiles for rainfall and nutrient indices can be condensed into rule-of-thumb charts that indicate, by district and crop, which constraints are likely to be most binding and what ranges of soil test values and rainfall are associated with substantial yield gains. Such tools are intended not to replace local knowledge, but to complement it with systematic evidence that can be used by extension workers when advising farmers in towns like Bijainagar and across Rajasthan.

Several limitations temper these conclusions. The reliance on district-level aggregates masks within-district heterogeneity; soil indices constructed from Soil Health Card categories introduce measurement error; weather variables remain relatively coarse; and economic considerations such as prices, input costs, and risk preferences are not yet integrated. Nonetheless, the study demonstrates that even with these constraints, combining openly available soil and weather data with relatively simple models can yield informative patterns about the drivers of yield variation and the potential leverage points for intervention.

Future work can build on this foundation in three directions. First, incorporating additional data streams—especially remote-sensing vegetation indices and higher-frequency weather descriptors—would allow more detailed modelling of crop growth dynamics and stress. Second, access to sub-district data in selected pilot regions would make it possible to estimate multi-level models that bridge district-level patterns and field-level recommendations. Third, linking physical yield models to economic outcomes and farmer constraints would provide a more complete basis for crop and input recommendations that reflect not only expected output but also profitability and risk. Pursuing these extensions in collaboration with local agronomists, extension services, and farmer groups would help ensure that future decision-support tools are both technically robust and grounded in the lived realities of Rajasthan's farming communities.

REFERENCES

[1] Bendre, M. R., and R. C. Thool. "Big Data in Precision Agriculture: Weather Forecasting for Future Farming." *Proceedings of the 1st International Conference on Next Generation Comput- ing Technologies (NGCT)*, IEEE, 2015, pp. 744–750.

[2] Bhar, Abhishek, et al. "Coordinate Descent Based Agricultural Model Calibration and Opti- mized Input Management." *Computers and Electronics in Agriculture*, vol. 172, 2020, article 105353.

[3] Gathala, Mahesh K., et al. "Conservation Agriculture Based Tillage and Crop Establishment Options Can Maintain Farmers' Yields and Increase Profits in South Asia's Rice–Maize Systems: Evidence from Bangladesh." *Field Crops Research*, vol. 172, 2015, pp. 85–98.

[4] Government of India. *Agricultural Statistics at a Glance 2022*. Directorate of Economics and Statistics, Department of Agriculture, Cooperation and Farmers Welfare, Ministry of Agricul- ture and Farmers Welfare, 2022.

[5] Government of India, Ministry of Agriculture and Farmers Welfare. *Soil Health Card Scheme: Operational Guidelines*. 2nd ed., Ministry of Agriculture and Farmers Welfare, 2017.

[6] Government of Rajasthan. *Statistical Abstract of Rajasthan 2023*. Directorate of Economics and Statistics, Government of Rajasthan, 2023.

[7] Hot, Emina, and Vesna Popović-Bugarin. "Soil Data Clustering by Using K-Means and Fuzzy K-Means Algorithm." *Telfor Journal*, vol. 8, no. 1, 2016, pp. 56–61.

[8] India Meteorological Department. *Rainfall Statistics of India: 2011–2022*. IMD, Ministry of Earth Sciences, Government of India, 2023.

[9] Khan, Hasnine, and S. M. Ghosh. "Crop Yield Prediction from Meteorological Data Using Efficient Machine Learning Model." *Proceedings of the International Conference on Wireless Communication*, Springer, 2020, pp. 565–574.

[10] Khoshnevisan, Bahram, et al. "Application of Artificial Neural Networks for Prediction of Output Energy and GHG Emissions in Potato Production in Iran." *Agricultural Systems*, vol. 123, 2014, pp. 120–127.

[11] Kumar, Amit. "Crop Recommendation for Maximizing Crop Yield Using Random Forest." *Innovations in Computational Intelligence and Computer Vision*, edited by Satyabrata Roy et al., Lecture Notes in Networks and Systems, vol. 680, Springer, 2023, pp. 501–515.

[12] Kumar, Ashish, S. Sarkar, and C. Pradhan. "Recommendation System for Crop Identification and Pest Control Technique in Agriculture." *Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, 2019, pp. 185–189.

[13] Kumar, S., and K. Balakrishnan. "Development of a Model Recommender System for Agricul- ture Using Apriori Algorithm." *Cognitive Informatics and Soft Computing*, Springer, 2019, pp. 153–163.

[14] Muniasamy, A. "Applications of Data Mining Techniques in Smart Farming for Sustainable Agriculture." *Research Anthology on Strategies for Achieving Agricultural Sustainability*, IGI Global, 2022, pp. 454–491.

[15] Navarro-Hell´ın, Hugo, et al. "A Decision Support System for Managing Irrigation in Agricul- ture." *Computers and Electronics in Agriculture*, vol. 124, 2016, pp. 121–131.

[16] Rehman, Talha U., et al. "Current and Future Applications of Statistical Machine Learning Algorithms for Agricultural Machine Vision Systems." *Computers and Electronics in Agriculture*, vol. 156, 2019, pp. 585–605.

[17] Römheld, Volker, and E. A. Kirkby. "Research on Potassium in Agriculture: Needs and Prospects." *Plant and Soil*, vol. 335, nos. 1–2, 2010, pp. 155–180.

[18] Suresh, G., et al. "Efficient Crop Yield Recommendation System Using Machine Learning for Digital Farming." *International Journal of Modern Agriculture*, vol. 10, no. 1, 2021, pp. 906–914.

[19] Zou, Hao, et al. "Optimization of Drip Irrigation and Fertilization Regimes for High Grain Yield, Crop Water Productivity and Economic Benefits of Spring Maize in Northwest China." *Agricultural Water Management*, vol. 230, 2020, article 105986.

[20] Zou, Hao, et al. "A Random Forest Classifier for Lymph Diseases." *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, 2014, pp. 465–473.