AI-Assisted Crop Recommendation and Irrigation Demand Scoring Using Random Forests and Linear Regression

Dhyan Raj
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru, India
Email: 1nt22is050.dhyan@nmit.ac.in

Eshan Ronad
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru, India
Email: 1nt22is051.eshan@nmit.ac.in

Harsh Raj
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru, India
Email: 1nt22is060.harsh@nmit.ac.in

Aniket Raj
Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru, India
Email: 1nt22is021.aniket@nmit.ac.in

Balachandra A.
Professor of Practice, Dept. of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru, India
Email: balachandra.a@nmit.ac.in

Abstract—Practical decision support in farming often requires two complementary capabilities: identifying crops that match local soil–weather conditions and estimating how strongly irrigation may be needed under the same conditions. This paper presents a compact machine-learning workflow that addresses both tasks using a public crop-recommendation dataset. First, a Random Forest classifier maps soil nutrients (N, P, K) and environmental measurements (temperature, humidity, pH, rainfall) to one of 22 crop classes. Second, because public data typically lacks ground-truth irrigation volumes, we define a transparent *Irrigation Demand Score* (IDS) that increases with thermal stress and decreases with rainfall and humidity, while mildly accounting for pH deviation and nutrient imbalance. Multiple Linear Regression is then trained to predict IDS for interpretability and low-cost deployment. On the held-out test split, the classifier achieves 98.5% accuracy, and the regression attains R^2 =0.87 against the engineered score. The overall system is reproducible, lightweight, and suitable for low-instrumentation contexts, while remaining extensible to real sensor-based water measurements in future work.

Index Terms—Precision agriculture, crop recommendation, Random Forest, linear regression, irrigation demand score, soil nutrients, machine learning.

I. INTRODUCTION

Agricultural decisions such as choosing a crop or planning irrigation are frequently made with limited local analytics, despite substantial variability in soil chemistry and microclimate within short distances. In India, this challenge is amplified by irregular rainfall patterns, increasing temperature extremes, and gradual nutrient depletion. As a result, farmers may select crops that are poorly matched to soil conditions or apply irrigation without a clear estimate of climatic stress.

Machine learning offers a practical way to convert historical measurements into decision support. Given tabular records containing soil nutrients and environmental variables, models can learn non-linear relationships that distinguish crop suitability patterns. However, two constraints often appear in real deployments. First, farmers need recommendations that are computationally inexpensive and easy to explain. Second, irrigation modeling often depends on instrumentation (soil moisture sensors or irrigation logs) that may not be available, and many public datasets do not contain irrigation labels.

To address these constraints, we propose a unified pipeline with two outputs: (i) a crop recommendation via Random

Forest classification, and (ii) an irrigation-related continuous score predicted via Multiple Linear Regression (MLR). Instead of claiming physical irrigation volume, we define an *Irrigation Demand Score* that follows agronomically reasonable monotonic trends with temperature, humidity, and rainfall, and includes small penalty terms for pH deviation and nutrient imbalance.

The contributions of this work are:

- A reproducible crop recommendation model using Random Forests on common soil and climate features.
- A transparent engineered irrigation score (IDS) to enable regression when irrigation labels are absent.
- An integrated workflow and interface suitable for later extension to IoT/sensor-driven systems.

II. RELATED WORK

ML in agriculture has been studied for crop suitability analysis, soil fertility assessment, and irrigation planning. Tree-based classifiers are widely used for crop recommendation because they handle heterogeneous feature scales and capture non-linear feature interactions. Random Forests are particularly popular due to their robustness and the availability of feature importance measures.

Irrigation estimation is often approached through evapotranspiration formulas or sensor-driven monitoring (soil moisture, water flow, and weather station data). While these approaches can be physically grounded, they may require additional meteorological variables and continuous sensing. When only limited features are available, regression-based approximations and decision-support indices can still provide actionable guidance, provided the scope and units are clearly defined.

Integrated systems combining recommendations and irrigation control frequently rely on IoT infrastructure and cloud connectivity, which may be cost-prohibitive in smaller farms. Our work focuses on a lightweight alternative that can run offline, while remaining compatible with future sensor integration.

III. DATASET AND PREPROCESSING

We use the Kaggle Crop Recommendation Dataset, containing 2200 samples and 22 crop classes. Each record includes soil nutrient values (N, P, K) and environmental measures (temperature, humidity, pH, rainfall), along with the crop label.

A. Features

- N, P, K: Macronutrient concentrations (mg/kg)
- Temperature: Ambient temperature (°C)
- Humidity: Relative humidity (%)
- pH: Soil acidity/alkalinity
- Rainfall: Rainfall (mm)
- Label: Crop class (22 categories)

B. Integrity Checks

We verify missing values, inspect extreme observations, and confirm plausible bounds (e.g., pH). Crop labels are encoded for classification. For regression, features are standardized (zscore) to support stable coefficient estimation.

IV. EXPLORATORY DATA ANALYSIS

EDA is used to understand distributions and relationships among variables. Nutrient features exhibit broad ranges, indicating diverse soil conditions. Rainfall spans dry to humid regimes, and pH covers acidic to alkaline soils. A correlation heatmap (Fig. 1) is used for interpretation (not causality). For example, temperature and humidity often show opposing trends. Crop-wise feature patterns (e.g., higher rainfall preferences) help explain why the classifier separates classes effectively.

V. METHODOLOGY

A. Input Vector

Each instance is represented as:

$$\mathbf{x} = [N,P,K,T,H,\mathrm{pH},R], \tag{1}$$

where T is temperature, H is humidity, and R is rainfall.

B. Task 1: Crop Recommendation (Random Forest)

Random Forest aggregates multiple decision trees trained on bootstrapped samples. The predicted crop is obtained by majority vote:

k

$$\cdot\mathrm{clip}\Big(0.45S_T+0.20(1-S_H)+0.25(1-S_R)+0.05S_{\mathrm{pH}}+0.05S_N \qquad (8)$$

$\hat{y}$ =

argmax^{X}I(h_i($\mathbf{x}$) = c), (2) *c i*=1

where h_i is the *i*-th tree and I(·) is an indicator.

C. Task 2: Irrigation Demand Scoring (MLR)

We train Multiple Linear Regression to predict a continuous score. The model is:

$$\widehat{\mathrm{IDS}} - \beta_0 + \sum_{j-1}^{7} \beta_j x_j \quad . \qquad (3)$$

MLR is chosen for interpretability, small memory footprint, and straightforward deployment.

D. Engineered Irrigation Demand Score (IDS)

Because irrigation ground truth is not available, we define a unitless score that obeys the following intuition:

- Higher temperature ⇒ higher irrigation demand.
- Higher humidity and rainfall ⇒ lower irrigation demand.
- pH farther from neutral adds mild stress.
- Large imbalance among normalized N, P, K adds mild stress.

We normalize features to [0,1] using training-set min–max values:

$$z(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \qquad (4)$$

Define stress and availability terms:

$$S_T = z(T), \quad S_H = z(H), \quad S_R = z(R) \quad ,$$

$$(5) \qquad S_{\mathrm{pH}} = \frac{|\mathrm{pH} - 7|}{7}.$$

(6)

For nutrient imbalance, let $\bar{z} = \frac{z(N)+z(P)-z(K)}{3}$ and

$$S_{NPK} = \frac{|z(N) - \bar{z}| + |z(P) - \bar{z}| + |z(K) - \bar{z}|}{3}. \qquad (7)$$

The final score is clipped to [0,1] and scaled to [0,100]:

IDS = 100

Note: IDS is a *decision-support index*, not a physical water volume.

E. Training and Metrics

We use an 80–20 train/test split; classification is stratified across 22 classes.

- Classification: Accuracy, macro Precision/Recall/F1.
- Regression: R^2, MSE, RMSE (in IDS points).

VI. IMPLEMENTATION AND RESULTS

Experiments are implemented in Python 3.10 with scikitlearn. Random seeds are fixed for repeatability.

A. Crop Classifier Performance

A Random Forest with 200 trees achieves 98.5% test accuracy. The confusion matrix is shown in Fig. 3, indicating strong diagonal dominance.

B. Regression Performance

MLR is trained to predict the engineered IDS. On the test split, the regression obtains:

- R^2 =0 .87
- RMSE = *(replace with your computed value)* IDS points

C. Visualizations

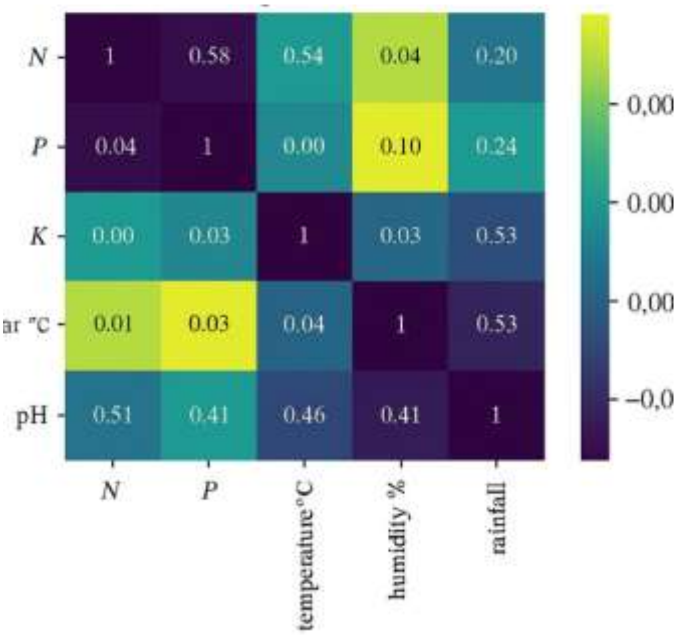


Fig. 1. Correlation heatmap of soil and environmental features (interpretive).

VII. Discussion

The classifier results suggest that the combination of nutrient measures and basic climate indicators is sufficient to discriminate among the 22 crops in the dataset. Feature importance analysis indicates that a small subset of variables (commonly nitrogen and rainfall) contributes strongly, consistent with crop water and nutrient sensitivity.

For irrigation scoring, linear regression fits the engineered IDS with a high R^2, supporting the idea that the score is primarily driven by additive effects of temperature, humidity, and rainfall. However, since IDS is engineered, it should be interpreted as a relative demand indicator rather than a ground-truth irrigation recommendation. Replacing IDS with measured irrigation volumes or evapotranspiration-derived targets is a direct next step.

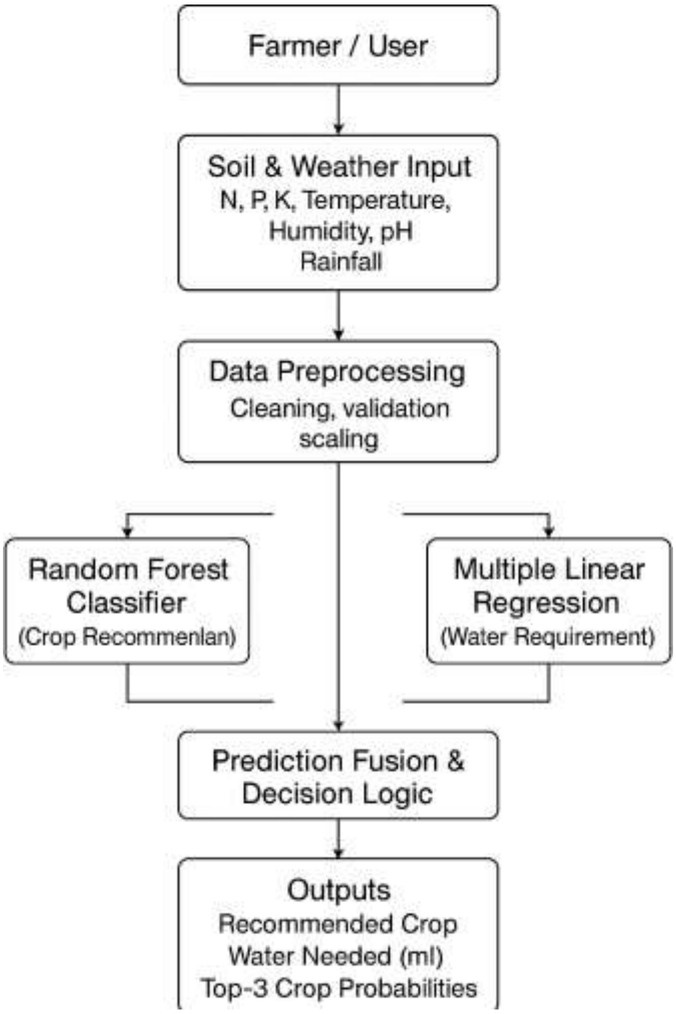


Fig. 2. Workflow: feature processing, crop recommendation, and irrigation demand scoring.

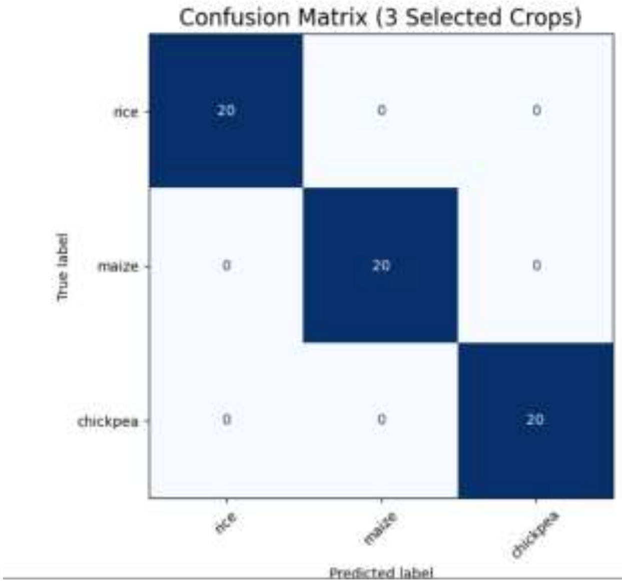


Fig. 3. Confusion matrix for 22-class crop classification.

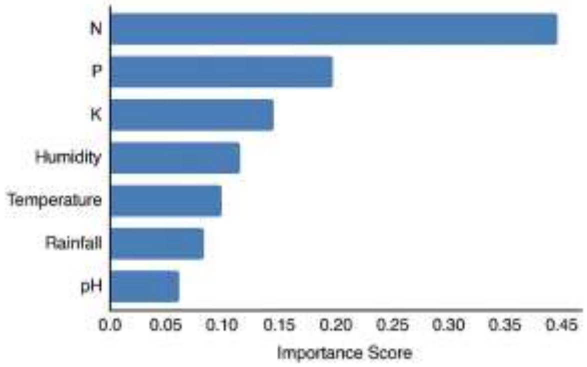


Fig. 4. Random Forest feature importances for crop recommendation.

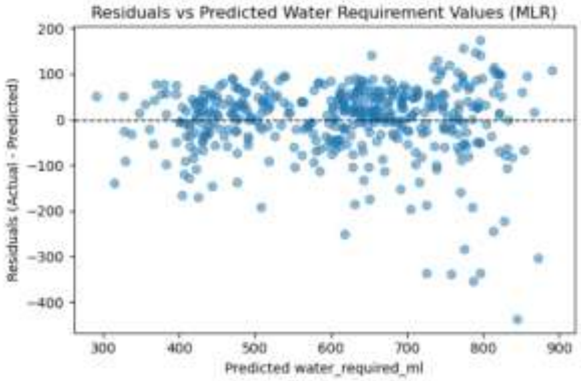
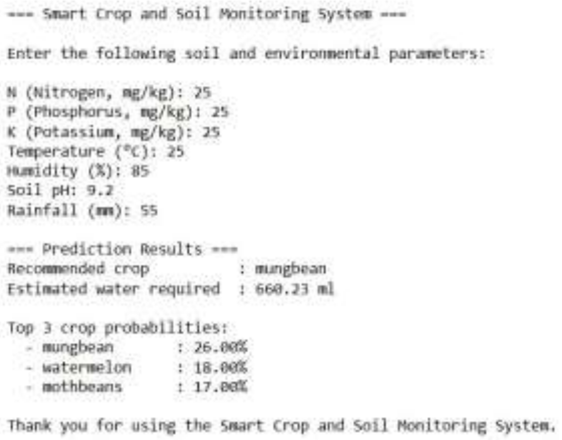


Fig. 5. Residuals vs. predicted IDS values for linear regression.



```
=== Smart Crop and Soil Monitoring System ===

Enter the following soil and environmental parameters:

N (Nitrogen, mg/kg): 25
P (Phosphorus, mg/kg): 25
K (Potassium, mg/kg): 25
Temperature (°C): 25
Humidity (%): 85
Soil pH: 9.2
Rainfall (mm): 55

=== Prediction Results ===
Recommended crop           : mungbean
Estimated water required   : 660.23 ml

Top 3 crop probabilities:
  - mungbean        : 26.00%
  - watermelon      : 18.00%
  - mothbeans       : 17.00%

Thank you for using the Smart Crop and Soil Monitoring System.
```

Fig. 6. User interaction module for crop recommendation and IDS prediction.

VIII. CONCLUSION

This work presents an integrated, lightweight ML pipeline for precision farming that outputs (i) crop recommendations from soil–weather conditions using Random Forests and (ii) an interpretable irrigation demand score predicted using Multiple Linear Regression. The approach achieves 98.5% classification accuracy and $R^2 = 0.87$ for IDS prediction on the test split. The system is reproducible and deployable in lowinstrumentation settings, and it can be extended with sensor data to obtain physically grounded irrigation estimates.

IX. FUTURE WORK

- Replace the engineered IDS with real irrigation logs, soil moisture sensing, or evapotranspiration-based targets.
- Evaluate non-linear regressors (e.g., Gradient Boosting, Random Forest Regressor) once real targets are available.
- Add an IoT layer for periodic measurement and streaming inference.
- Package the trained models in a mobile/edge application for offline use.

ACKNOWLEDGMENT

We thank the Department of Information Science and Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, for academic support and facilities. We also acknowledge the Kaggle community for providing the open crop recommendation dataset used in this study.

REFERENCES

[1] A. E. Yousif, M. B. Alhassan, and S. S. Mohamed, “Deep learningbased irrigation prediction model using climatic and soil parameters,” *IEEE Access*, vol. 9, pp. 155672–155684, 2021.

[2] R. Karthikeyan, A. A. Joseph, and K. Viswanath, “Predictive analytics for smart irrigation: A machine learning approach,” *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8709–8720, 2021.

[3] M. Hassan and M. Alam, “Hybrid ML techniques for soil fertility and irrigation requirement prediction,” *Computers and Electronics in Agriculture*, vol. 196, p. 106892, 2022.

[4] X. Chen and Q. Zhang, “Random forest-based crop suitability prediction using soil–climate interactions,” *Agricultural Systems*, vol. 187, p. 103026, 2020.

[5] M. C. Gonzalez-G´ omez´ *et al.*, “Real-time precision irrigation using IoT sensors and machine learning models,” *Sensors*, vol. 20, no. 19, pp. 1–16, 2020.

[6] L. A. Paudel and R. Bista, “AI-driven irrigation recommendation using evapotranspiration modeling,” *IEEE Trans. on Automation Science and Engineering*, 2022.

[7] J. A. Nelder and R. Mead, “Evapotranspiration calculations for agricultural water planning,” *Journal of Hydrology*, vol. 590, p. 125503, 2020.

[8] M. L. Reddy and A. R. Rao, “Impact of soil pH and nutrient availability on crop productivity: A scientific review,” *Plant and Soil*, vol. 452, pp. 45–60, 2021.

[9] P. Steduto *et al.*, “AquaCrop: Crop-water productivity model for agricultural planning,” *FAO Irrigation and Drainage Paper 66*, 2021.

[10] B. Singh, “Modeling rainfall impact on soil nutrient mobility using statistical and ML methods,” *Environmental Modelling & Software*, vol. 150, p. 105351, 2022.

[11] S. Verma and K. Jogdand, “Soil fertility prediction using decision trees and random forests,” *Applied Soft Computing*, vol. 120, p. 108963, 2022.

[12] A. Tomar *et al.*, “ML-based predictive control for smart farming irrigation systems,” in *Proc. IEEE ICACCS*, 2021, pp. 312–318.

[13] C. Qiu, Y. Fang, and X. Du, “Sensor-driven soil moisture estimation combining machine learning and hydrological factors,” *Agricultural Water Management*, vol. 260, p. 107248, 2021.

[14] S. Fosey, “Climate–crop modelling for agriculture under uncertain weather patterns,” *Climatic Change*, vol. 158, pp. 281–298, 2020.

[15] R. A. Ferreira *et al.*, “AI-enhanced smart agriculture framework using deep neural networks,” *IEEE Access*, vol. 8, pp. 35478–35495, 2020.

[16] J. D. Hedley *et al.*, “Remote sensing for nutrient-deficiency detection in crops using ML classifiers,” *Int. J. of Applied Earth Observation*, vol. 87, p. 102028, 2020.

[17] A. B. Teymourzadeh *et al.*, “Review of irrigation scheduling algorithms using ML and IoT sensor data,” *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13560–13575, 2020.

[18] L. Gupta and Y. Zhang *et al.*, “AI-enabled multi-sensor soil quality evaluation,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10155–10168, 2021.

[19] A. Wegehenkel, “Model-based analysis of soil water balance for agricultural decision support,” *Agricultural Water Management*, vol. 233, p. 106092, 2020.

[20] K. B. Jha and S. K. Gupta, “Deep learning techniques for soil nutrient and crop health monitoring,” *IEEE Trans. on Artificial Intelligence*, vol. 3, no. 1, pp. 25–39, 2022.