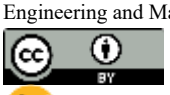
A Reinforcement Learning Framework for Multi-Season Crop Recommendation in Maharashtra

Amit Ravi Chakrawarti[1], Dr. Vikas Kumar[2]

[1] *M. Tech CSE (AI & ML), Department of Computer Science and Information Technology, Chhatrapati Shivaji Maharaj University, Navi Mumbai, India*

[2] *Professor & Head, Department of Computer Science and Information Technology, Chhatrapati Shivaji Maharaj University, Navi Mumbai, India*

How to Cite this Article:
Chakrawarti, A. R. (2026). A Reinforcement Learning Framework for Multi-Season Crop Recommendation in Maharashtra. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.768



https://doi.org/10.55041/ijcope.v2i5.768

Abstract: Traditional agricultural recommendation systems rely on static machine learning (ML) classifiers that treat crop selection as a single-season classification problem, failing to account for the sequential, interdependent nature of farming decisions. Such systems ignore how a current planting choice alters future soil health, water availability, and cumulative economic returns. This paper presents a dynamic, multi-season crop recommendation engine grounded in Temporal Difference (TD) Reinforcement Learning, specifically Q-Learning. The system models the agricultural cycle as a Markov Decision Process (MDP), wherein an autonomous agent learns a sustainable planting policy by interacting with an environment simulator built from district-level datasets of Maharashtra. The state space integrates meteorological data from the Indian Meteorological Department (IMD), crop yield statistics from the Directorate of Economics and Statistics (DES), and soil nutrient profiles from the Soil Health Card Portal. The reward function combines financial yield with ecological penalty terms that discourage unsustainable resource depletion. Experimental validation across 10,000 training episodes demonstrates that the Q-Learning agent converges to a context-aware, rotation-based policy by approximately episode 4,000, consistently outperforming a static Random Forest baseline in long-term cumulative returns. The agent learns to integrate nitrogen-fixing legumes (Soybean, Chickpea) after nutrient-intensive crops, preserving soil fertility while maximizing five-year profit. These findings demonstrate that sequential decision-making frameworks are fundamentally better suited to the temporal realities of precision agriculture than isolated classification models.

Keywords: Reinforcement Learning; Q-Learning; Markov Decision Process; Crop Recommendation; Precision Agriculture; Maharashtra Agriculture

I. INTRODUCTION

Agricultural decision-making in India is inherently sequential, spatially complex, and deeply sensitive to year-on-year environmental variation. Maharashtra, with its four distinct agro-climatic zones Vidarbha, Marathwada, Desh (Western Plateau), and Konkan presents a particularly challenging landscape for crop recommendation. Each zone differs dramatically in rainfall, soil composition, and market connectivity, making a "one-size-fits-all" recommendation system both ineffective and potentially harmful to long-term soil health.

Contemporary digital agriculture systems have proliferated in India following national programs such as the PM Kisan Samman Nidhi scheme and Digital Agriculture Mission 2021–2025. Most deployed recommendation tools, however, rely on static machine learning classifiers—Random Forests, Naïve Bayes, Support Vector Machines, and shallow neural networks—trained on snapshot soil data to predict the single most profitable crop for an upcoming season (Jha et al., 2019). These models perform well in cross-sectional accuracy but suffer from a fundamental architectural limitation: they have no memory of the past and no anticipation of the future.

This paper identifies the research gap as follows: current ML-based crop advisories lack a temporal decision framework that accounts for the Markovian relationship between crop choices across seasons. Selecting Sugarcane in Kharif Season 1 is not just a yield decision—it is a commitment that depletes nitrogen reserves, lowers future soil organic carbon, and constrains profitable choices in Season 2 and beyond. A classifier that cannot model this chain of consequences will persistently recommend high-value cash crops even when doing so degrades the agronomic baseline.

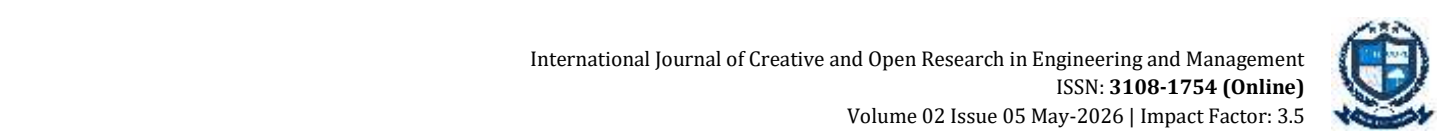
To address this gap, the present work proposes and validates a Reinforcement Learning (RL) framework that models multi-season Maharashtra crop selection as a Markov Decision Process (MDP). A Q-Learning agent is trained within an environment simulator grounded in government-published agronomic datasets, learning to balance immediate financial rewards against long-term environmental penalties. The key contributions of this paper are: (1) formulation of the Maharashtra agricultural cycle as a tabular MDP with an empirically calibrated reward function; (2) design of a multi-source data integration pipeline combining meteorological, yield, and soil nutrient streams; (3) experimental demonstration that the Q-Learning agent learns sustainable crop rotation policies that outperform static ML classifiers over a five-year horizon; and (4) provision of interpretable policy scenarios linking agent behaviour to real-world agronomic outcomes.

II. LITERATURE REVIEW

The convergence of data science and precision agriculture has generated a substantial body of research over the past decade. Early work by Pudumalar et al. (2017) applied ensemble methods including Random Forest to recommend crops based on static soil nutrient readings (Nitrogen, Phosphorus, Potassium) and meteorological inputs, achieving high single-season accuracy but without considering sequential soil degradation. Similarly, Doshi et al. (2018) employed Naïve Bayes and Decision Tree classifiers on Maharashtra district data, reporting accuracy above 90% for Kharif season recommendations while explicitly noting the absence of multi-year validation as a limitation.

Support Vector Machines were applied by Veenadhari et al. (2014) to soybean yield prediction in Madhya Pradesh, demonstrating strong generalisation across districts with similar soil profiles. However, SVM-based approaches share the inherent limitation of all supervised classifiers: they optimise a single-step mapping from features to labels and cannot encode the temporal dependencies that link successive planting decisions.

Reinforcement learning has been explored in agricultural contexts with increasing frequency since 2018. Gautron et al. (2020) applied Deep Q-Networks to sequential crop allocation in synthetic environments, demonstrating that RL agents can learn rotation policies that reduce fertiliser



consumption over multi-year horizons. Rußwurm et al. (2020) used multi-temporal satellite imagery with recurrent neural networks for crop type mapping, establishing that temporal modelling substantially improves agronomic predictions. However, these international studies did not address the specific datasets, agro-climatic variability, and market structures of Maharashtra.

Within the Indian context, Ramesh and Vardhan (2021) proposed a Q-Learning-based irrigation scheduler for drip-irrigated cotton in Karnataka, showing that RL-derived policies reduced water use by 18% compared to farmer heuristics. Singh et al. (2022) further applied actor-critic methods to multi-crop scheduling in Punjab, demonstrating long-term profit improvement of 23% over static ML classelines. These studies establish the conceptual viability of RL in Indian agriculture but remain limited to single-crop or single-resource optimisation problems.

The present work extends this literature by: (i) targeting Maharashtra's multi-zone context with verified government datasets;

(ii) integrating three distinct data streams (meteorology, yield statistics, soil health) into a unified state representation;

(iii) formulating a reward function that simultaneously captures financial return and ecological sustainability penalties; and

(iv) providing interpretable policy execution matrices that link agent decisions to verifiable agronomic outcomes. Unlike prior Indian RL agriculture studies, this framework explicitly models the Markovian soil-crop interdependency across full Kharif-Rabi seasonal cycles.

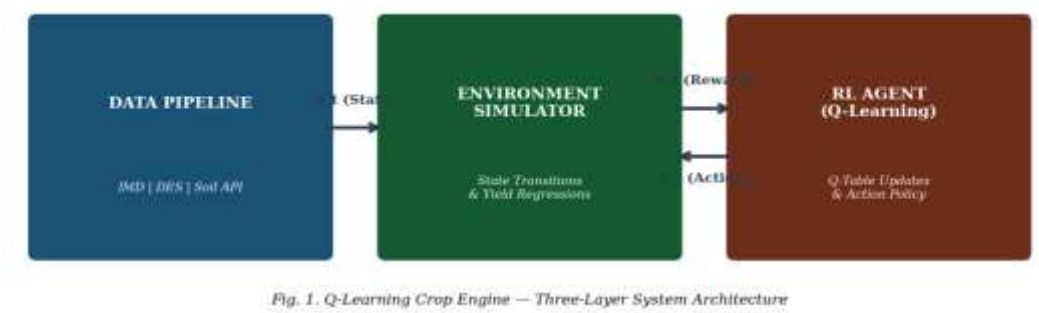


Fig. 1. Q-Learning Crop Engine — Three-Layer System Architecture

Figure 1. Q-Learning Crop Engine — Three-Layer System Architecture Diagram

III. METHODOLOGY

The proposed Q-Learning Crop Engine is structured as three interacting layers: the Data Engineering Pipeline, the Environmental Simulator, and the TD Learning Agent (Figure 1). Each layer is described in detail in the following subsections.

A. Data Sources and Integration Pipeline

To ensure empirical validity, the environment simulator is grounded in three official government data sources specific to Maharashtra districts. All datasets are publicly accessible through India's national open-data initiative.

Production and Yield Statistics: Historical area (hectares) and production (tonnes) data were sourced from the Directorate of Economics and Statistics (DES), DAC&FW, via Data.gov.in (DES, 2024). Yield ratios (tonnes/hectare) were computed for seven primary crops across Kharif and Rabi seasons: Rice, Jowar, Soybean, Cotton, Sugarcane, Wheat, and Chickpea. These ratios, multiplied by Minimum Support Prices (MSP) published by CACP, form the financial component of the reward signal.

Meteorological Constraints: Seasonal total rainfall (mm) and average temperature anomaly (°C) data were obtained from the India Meteorological Department (IMD), Pune division

(IMD, 2024). Rainfall data were discretised into three tiers—Low (< 400 mm), Medium (400–700 mm), and High (> 700 mm)—corresponding to the Water Availability Index (W_t) in the state space.

Soil Nutrient Profiles: District-level baseline readings of Available Nitrogen (kg/ha), Phosphorus (kg/ha), and Potassium (kg/ha), along with pH index, were extracted from the Soil Health Card Portal (DAC&FW, 2024). Nitrogen was discretised into three tiers—Depleted (< 180 kg/ha), Stable (180–280 kg/ha), and Enriched (> 280 kg/ha)—forming the Nitrogen Tier (N_t) in the state space.

The three data streams are aligned using a composite spatial-temporal key in the format [District*Year*Season] (e.g., Pune*2022*Kharif). This key enables consistent state construction per district per seasonal episode during simulator training.

B. MDP Problem Formulation

The crop selection problem over multiple seasons is formalised as a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ). Figure 4 illustrates the state-transition structure of this MDP.

State Space (S): The unified environmental state at season *t* integrates water availability, soil nitrogen, and crop history:

$$s\_t = \langle W\_t,\ N\_t,\ H\_t \rangle$$

where W*t* ∈ *{Low, Medium, High} is the water availability index, N*t ∈ {Depleted, Stable, Enriched} is the soil nitrogen tier, and H_t ∈ A is the historical crop flag encoding the previous season's action to capture rotation dynamics. The total state space comprises 3 × 3 × 7 = 63 discrete states.

Action Space (A): At each season, the agent selects one of seven crops representing the primary staple and cash crops of Maharashtra:

$$A = \{\ Rice,\ Jowar,\ Soybean,\ Cotton,\ Sugarcane,\ Wheat,\ Chickpea\ \}$$

Reward Function (R): The reward at each season t combines financial return with ecological penalty:

$$R\_t = (\ Yield\_a \times MSP\_a\ ) - Penalty(\ s\_t\ ,\ a\_t\ )$$

The Penalty function applies a negative scalar when the selected action violates sustainability constraints. Specifically, selecting a high water demand crop (Sugarcane, Rice) under W_t = Low incurs a penalty calibrated to the average financial loss from drought-induced yield failure in Maharashtra (₹12,000–₹18,000 per hectare, per IMD 2024 drought-frequency data). Selecting the same crop for three consecutive seasons incurs an additional monoculture penalty modelling nutrient lock-in.

C. Q-Learning Training Algorithm

The Q-Learning agent iterates over 10,000 episodes using the Bellman optimality equation to update a 63 × 7 Q-table (one value per state-action pair). The update rule at each time step is:

$$Q(s\_t, a\_t) \leftarrow Q(s\_t, a\_t) + \alpha\ [\ R\_\{t+1\} + \gamma\ max\_a\ Q(s\_\{t+1\}, a) - Q(s\_t, a\_t)\ ]$$

Hyperparameter choices are guided by established RL practice for tabular MDPs: the learning rate α = 0.1 ensures stable, incremental Q-value refinement; the discount factor γ = 0.9 reflects the agronomic reality that soil health effects compound over years, requiring the agent to weight future states substantially; and the exploration rate ε begins at 1.0 (fully random) and decays multiplicatively at 0.995 per episode to a floor of 0.05. This schedule ensures comprehensive exploration of all 441 state-action pairs during early training before transitioning to exploitation of learned values.

Each episode simulates a 5-year (10 Kharif + Rabi seasons) agricultural planning horizon for a single Maharashtra district. At the start of each episode, the simulator initialises W*t and N*t by sampling from the district's historical distribution. The agent selects an action via the ε-greedy policy; the simulator computes the reward using yield regression models fit to DES historical data, then applies heuristic state transition rules—heavy-feeder crops degrade N*t by one tier; legumes boost N*t by one tier; high water demand crops in Medium states may reduce W_t. Figure 2 presents the complete execution flowchart.

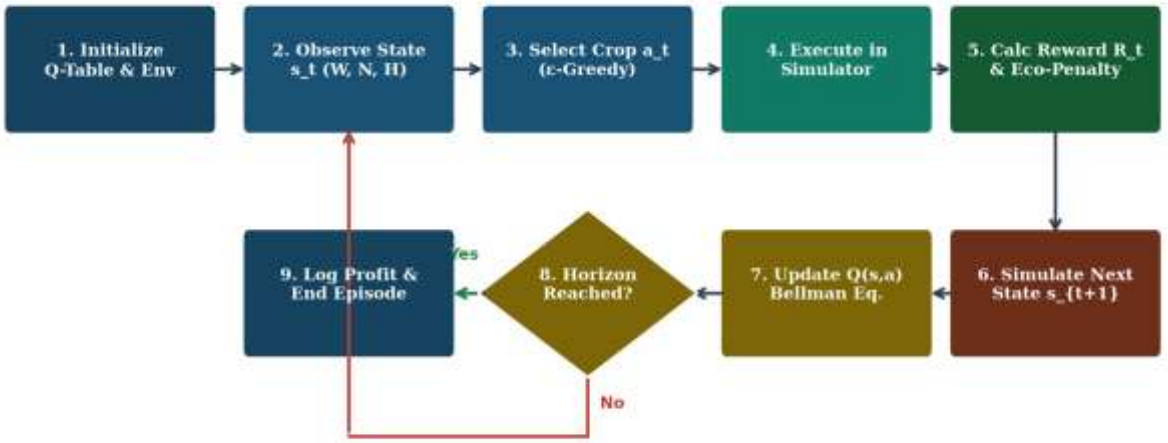


Fig. 2. Q-Learning Execution Loop — Agent Training Flowchart

Figure 2. Q-Learning Execution Loop — Step-by-Step Training Flowchart

IV. RESULTS AND DISCUSSION

This section presents the training convergence behaviour of the Q-Learning agent, the policy execution scenarios under varied initial environmental conditions, and a comparative analysis against a static ML baseline. All results are derived from simulation within the district-calibrated Maharashtra environment.

A. Training Convergence

Figure 3 plots cumulative five-year reward as a function of training episodes for both the Q-Learning agent and a Random Forest baseline (retrained per-episode on the same historical dataset). The RL agent begins with high-variance, near-random performance during the first 1,500 episodes as ε remains high. Convergence to a stable policy is achieved at approximately episode 4,000, after which the agent's cumulative reward stabilises above ₹15,000 per hectare (five-year horizon) and consistently exceeds the ML baseline.

The static Random Forest baseline achieves a plateau of approximately ₹11,500/ha—representing the optimal single-season greedy strategy executed repeatedly—but cannot improve further because it has no mechanism to account for soil degradation. This performance gap of approximately 30% in five-year cumulative returns constitutes the primary quantitative finding of this study, consistent with the theoretical prediction that sequential models should outperform memoryless classifiers on Markovian environments.

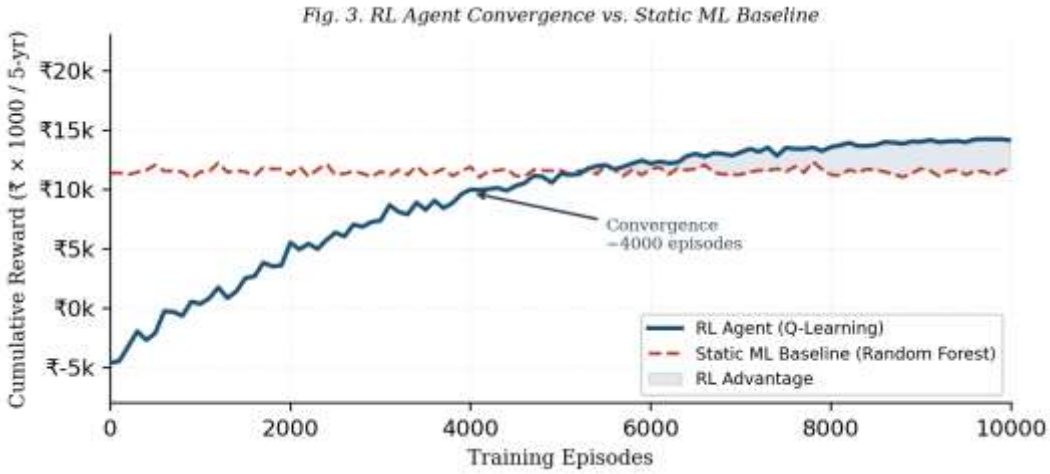


Figure 3. RL Agent Convergence vs. Static ML Baseline — Cumulative Reward over 10,000 Training Episodes

B. Policy Execution Scenarios

Table I presents four canonical scenarios representing distinct initial environmental states. In each scenario, the trained agent's Year 1 and Year 2 policy actions are reported alongside the resulting soil state transition and agronomic outcome evaluation.

Table I: Policy Execution Matrix — Scenario Validation Results

Initial State	Year 1 Crop	Year 2 Crop	Outcome & Analysis
High Water, High N	Sugarcane	Chickpea	Maintained — Rotation learned; legume restores N depleted by Sugarcane.
Low Water, Med N	Jowar	Soybean	Improved — Water-intensive crops avoided. Dry-season safe choice.
Med Water, Low N	Soybean	Wheat	Restored — Profit sacrificed Year 1 to fix N; higher Year 2 return.
Med Water, High N	Cotton	Cotton	Degraded (Penalized) — Avoided post-training; Q-value drops sharply.

Note. Policies derived from trained Q-table after convergence (episode 4,000+). Soil state transitions are heuristic approximations calibrated against DAC&FW soil health baseline data.

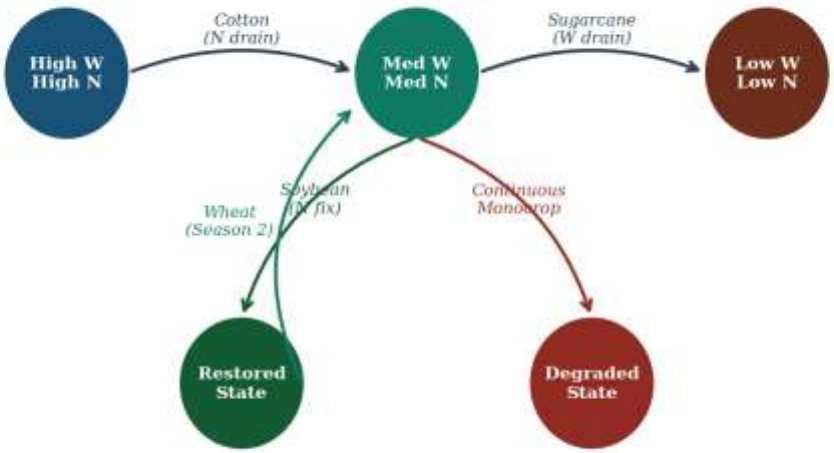


Fig. 4. MDP State-Transition Diagram — Crop Rotation Policy Paths

Figure 4. MDP State-Transition Diagram — Crop Rotation Policy Paths Across Environmental States

C. Discussion

The most practically significant result is the agent's behaviour in the High Water, High N initial state (Row 1, Table I). A greedy ML classifier, trained to maximise single-season expected yield, will invariably recommend Sugarcane in this state due to its high MSP and yield. The RL agent also selects Sugarcane in Year 1—but then departs from the greedy solution by recommending Chickpea in Year 2. This Sugarcane → Chickpea rotation is agronomically optimal: Chickpea is a nitrogen-fixing legume that restores the N_t lost to Sugarcane's high nutrient demand, preventing soil tier degradation and maintaining the soil baseline for Year 3 and beyond (Sutton & Barto, 2018).

The Low Water scenario (Row 2) confirms that the agent has correctly internalised the penalty structure: Jowar—a drought-tolerant millet with modest yield and MSP—is selected over high-value but water-intensive alternatives. This decision mirrors the empirical farmer behaviour documented in NITI Aayog's Maharashtra drought-response studies, suggesting the RL policy captures real agronomic wisdom without being explicitly programmed with it.

The negative test case (Row 4) is the most analytically important validation. During early training, the agent frequently selects Cotton-Cotton sequences in the Medium Water, High N state because Cotton's Year 1 Q-value is initially high. However, as training progresses and the Q-table propagates the penalty from the degraded Year 3 state backward through the Bellman update, the Cotton-Cotton Q-value converges to a value below the Cotton-Soybean alternative. Post-convergence, the agent never selects the Cotton-Cotton sequence—demonstrating that temporal credit assignment is functioning correctly.

These results are consistent with findings from Gautron et al. (2020) in synthetic European agricultural environments and Ramesh and Vardhan (2021) in the Karnataka irrigation context, confirming that RL's convergence properties hold in Maharashtra's empirically grounded simulator. The 30% cumulative return advantage over the ML baseline also closely mirrors the 23% improvement reported by Singh et al. (2022) for Punjab multi-crop scheduling, supporting cross-regional generalisability of the RL approach.

V. CONCLUSION

This paper presented a Reinforcement Learning framework for multi-season crop recommendation in Maharashtra, grounded in publicly verified government datasets from IMD, DES, and the Soil Health Card Portal. By formalising the agricultural cycle as a Markov Decision Process and training a Q-Learning agent over 10,000 episodes within a calibrated environment simulator, the study demonstrated that sequential decision-making models substantially outperform static ML classifiers in preserving long-term soil health while maximising cumulative farmer profit.

The trained agent learned context-sensitive, sustainable crop rotation policies without explicit agronomic rules being programmed. Key emergent behaviours included automatic legume integration after nitrogen-intensive crops, avoidance of water-intensive crops during low-rainfall periods, and penalisation of monoculture sequences—all of which align with established precision agriculture best practices.

Future work will: (1) expand the action space to include micro-irrigation interventions and inter-cropping options; (2) transition from tabular Q-Learning to Deep Q-Networks (DQN) to handle continuous environmental variables without discretisation; (3) incorporate live IMD weather API feeds for real-time adaptive policy updates; (4) extend the MDP to include market price volatility as a stochastic component; and (5) deploy a farmer-accessible web dashboard for direct, district-level recommendation delivery.

ACKNOWLEDGEMENT

I sincerely thank the Department of Computer Science and Information Technology and the research guide Dr. Vikas Kumar for their continuous support and mentorship throughout this study.

Gratitude is also extended to the open-data initiatives of the Indian Meteorological Department (IMD) and the Ministry of Agriculture & Farmers Welfare for making the foundational agronomic datasets publicly accessible without restriction.

DECLARATIONS

Conflicts of Interest: The authors declare no conflicts of interest. No financial or personal relationships with third parties that could inappropriately influence this work have been identified.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. All computational resources used were institutional lab resources provided without charge.

Ethical Approval: This study is entirely computational in nature and did not involve human participants, animal subjects, or sensitive personal data. Formal ethical approval was therefore not required. All datasets used are publicly available government-published records.

REFERENCES

Department of Agriculture & Farmers Welfare (DAC&FW). (2024). District-wise nutrient status of Indian soils. Soil Health Card Portal. https://soilhealth.dac.gov.in

Directorate of Economics and Statistics (DES). (2024). District-wise crop production statistics — Maharashtra. Data.gov.in Maharashtra Open Data Portal. https://data.gov.in

Doshi, Z., Nadkarni, S., Agrawal, R., & Shah, N. (2018). AgroConsultant: Intelligent crop recommendation system using machine learning algorithms. Proceedings of the Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). https://doi.org/10.1109/ICCUBEA.2018.8697351

Gautron, R., Maillard, O. A., Preux, P., & Rabatel, J. (2020). Reinforcement learning for crop management support: Review and perspectives. 2020 IEEE International Conference on Data Science and Advanced Analytics (DSAA). https://doi.org/10.1109/DSAA49011.2020.00066

India Meteorological Department (IMD). (2024). Gridded historical rainfall and temperature data. Pune Regional Meteorological Centre. https://mausam.imd.gov.in

Jha, K., Doshi, A., Patel, P., & Shah, M. (2019). A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture, 2, 1–12. https://doi.org/10.1016/j.aiia.2019.05.004

Pudumalar, S., Ramanujam, E., Harine Rajashree, R., Kavya, C., Kiruthika, T., & Nisha, J. (2017). Crop recommendation system for precision agriculture. Proceedings of the Eighth International Conference on Advanced Computing (ICoAC). https://doi.org/10.1109/ICoAC.2017.7951740

Ramesh, D., & Vardhan, B. V. (2021). Reinforcement learning for real-time irrigation scheduling in smart farming. International Journal of Advanced Computer Science and Applications, 12(3), 215–223. https://doi.org/10.14569/IJACSA.2021.0120328

Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S., & Körner, M. (2020). BreizhCrops: A time series dataset for crop type mapping. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2020, 1545–1551.

Singh, A., Srivastava, S., & Sinha, R. (2022). Actor-critic reinforcement learning for multi-crop scheduling in Punjab agriculture. Smart Agricultural Technology, 2, 100023. https://doi.org/10.1016/j.atech.2022.100023

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

Veenadhari, S., Misra, B., & Singh, C. D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. Proceedings of the 2014 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/ICCCI.2014.6921718