

Intelligent Systems Research and Applications Journal

www.israj.org
eISSN: 3120 - 0826

Beyond Saturated Scores: A Leakage-Aware and Disparity-Audited Evaluation Protocol for Crop Recommendation and Yield Estimation

Vipin Bihari Yadav[1,*] Aditya Khamparia[1,]

[1] Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Satellite Centre Amethi, Amethi 227413, Uttar Pradesh, India

ARTICLE INFO

Article history:
Received 2 September 2026
Received in revised form 18 October 2026
Accepted 29 October 2026
Available online 2 November 2026

Keywords:
Crop recommendation; Crop yield prediction; evaluation protocol; Data leakage; Temporal validation; Group disparity; Levelling down; Explainable AI; SHAP

ABSTRACT

Crop recommendation and yield estimation are usually reported as two unrelated machine-learning results, and both look close to perfect. We ask how much of that skill is still there once the evaluation is made stricter. We start from an exact reproduction of a published dual-model Random Forest framework (99.55% recommendation accuracy; yield $R^2 = 0.948$) and add LEAF-Audit, a four-stage protocol made of Leakage quantification, Evaluation under distribution shift, Attribution of group disparity and Feasibility coupling. On the same two public datasets, recommendation accuracy is saturated: Random Forest (99.52%) and Naive Bayes (99.47%) are equivalent within a ±0.5-point margin, and the data look augmented. For yield, 96.8% of random-split test rows come from district–crop–season series that the training set already contains, and a historical-mean lookup almost matches Random Forest on the log scale (R^2 0.947 against 0.962). Under a temporal split, Random Forest (log-scale $R^2 = 0.85$) falls behind the historical mean (0.90). XGBoost with native categorical splits (0.92) improves on that mean by 0.15 in mean absolute error, with an interval that excludes zero under district, state and year clustering, and it also beats the last-year and rolling-mean floors under district and year clustering. The same booster with integer-coded categories, and Random Forest with one-hot or target-encoded categories, do not. Across six geographic zones the worst-to-best normalised-error ratio is 1.79 (95% interval 1.64–1.96), far above a permutation null (95th percentile 1.18). We also propose a levelling-down index. Over five seeds, adaptive difficulty-based reweighting lowers the zone and cell ratios, but 79–82% of the log-ratio reduction comes from raising error in the best-served cells; inverse-frequency reweighting does not lower the ratios. A precedent filter leaves 6.9% of records without a recommendation and mostly blocks correct, newly adopted crops. What we contribute is a reusable audit protocol, not a new learner.

1. Introduction

Every season, farmers and planners have to decide which crop to plant and roughly what it will yield. Machine learning has been applied to both questions, and the reported scores look almost perfect. On the widely used 2,200-record crop recommendation dataset, many classifiers pass 99% accuracy [1–4], and the study we build on reports an R^2 of 0.948 for district-level yield [5]. That study

* *Corresponding author.*
E-mail address: vipin884060@gmail.com
https://doi.org/10.59543/mwg5gs97

links the two tasks: a Random Forest picks a crop, a second Random Forest turns the crop into a yield estimate, and SHAP explains both models. We chose it because its method is described clearly and its data are public. That let us reproduce its numbers exactly before asking what they mean.

Three things about the way such systems are evaluated bothered us. The first is random splitting. In panel data, neighbouring years of the same district, crop and season land on both sides of the split, so a flexible model can do well just by memorising series instead of forecasting them [6, 7]. A random split shows how well a model fills gaps. It does not show how well it forecasts. The second is that the two stages are only loosely connected. The recommender reads soil and weather, the yield model reads location, season and area, and the only thing passed between them is a crop name. Nothing checks that the recommended crop is actually grown where the yield is requested. The third is that pooled metrics can hide unequal service. A model that is accurate on average can still do worse for particular regions or production scales [8, 9], yet error by group is seldom reported in this literature.

Instead of fixing these problems one at a time, we built a single reusable protocol, which we call LEAF-Audit, and organised the work around four questions. RQ1 (leakage): how much of the reported skill comes from the way the models were evaluated? RQ2 (shift): which simple baselines are hard to beat once the task becomes forecasting or spatial transfer, and does the ranking of models hold up? RQ3 (disparity): how unevenly is error spread across zones and production scales, and when a mitigation lowers a disparity ratio, is the improvement real or does it come from making the best-served group worse? RQ4 (coupling): what does it cost, in coverage and accuracy, to make the recommender respect local cropping history?

The paper makes five contributions. (i) LEAF-Audit itself, a four-stage protocol in which every accuracy figure is reported with its evaluation protocol, a simple baseline and a cluster-aware interval. (ii) Leakage diagnostics for panel agriculture data: series-overlap and adjacent-year coverage rates, plus an audit of unit, zero-value and aggregate-row problems. (iii) A levelling-down index, which splits any change in a worst-to-best disparity ratio into the part due to the worst-served group improving and the part due to the best-served group getting worse. (iv) A permutation-null and crop-stratified disparity audit. (v) A precedent-based feasibility diagnostic, with SHAP and permutation importances reported together with their uncertainty. We do not propose a new learning algorithm. What we want to show is how much apparent performance depends on the evaluation protocol, which simple baselines are hard to beat, and what a responsible report should contain once disparity is taken seriously.

2. Related Work

Machine learning now touches crop, soil, water and livestock management [10]. Deep learning is mostly used for images and sensor data [11], and many tools are being bundled into decision-support platforms under the label Agriculture 4.0 [12]. For small tabular problems like ours, tree ensembles still hold their own against deep networks [13], so Random Forest and gradient boosting are the obvious models to compare [14, 15]. Most recommendation studies use the 2,200-sample dataset, which has seven soil and climate variables and 22 crops. Dey et al. modelled agricultural and horticultural crops separately [1]. Alam et al. combined six learners by soft voting, added an entropy-based uncertainty measure and reached about 99.5% accuracy [2]. Shastri et al. paired gradient boosting with LIME [3], and Ahmed et al. built stacking ensembles [4]. The pattern across these papers is similar: accuracy within a few tenths of a percent of the ceiling, measured on a few hundred test samples from one split. At that size a single extra error moves accuracy by 0.23 points, so the

rankings are unstable. This is why we use repeated and nested validation, together with tests that allow for overlap between folds [16, 17, 52].

Yield prediction spans many spatial scales and input types. One systematic review found temperature, rainfall and soil to be the most common inputs, with neural networks widely used [18]. Jeong et al. showed that random forests beat multiple linear regression for wheat, maize and potato [19], and Khaki et al. combined convolutional and recurrent networks to estimate corn and soybean yields across the United States [20]. Most of this work uses weather, remote sensing or management data. District-level administrative statistics, like the production file we use, record only location, season, year, crop and area. A few authors have treated recommendation and yield together. Badshah et al. report crop classification and yield prediction with feature importance and LIME [21], and the baseline study connects a recommender to a yield regressor through a crop-name mapping [5]. In these designs the second model gets only the crop label from the first, and none of the studies we read checks whether the suggested crop is plausible for the place and season.

When observations are correlated in space and time, random cross-validation overstates skill, and blocked or leave-group-out designs are recommended [7]. Ploton et al. found that random validation hid poor performance that spatial validation exposed [22], and Meyer et al. showed that predictors encoding location can look useful under random validation and then fail in new places [23]. For time-ordered data, out-of-sample and blocked designs are preferred [53, 54], forecasts should be compared with naive references [55], and leakage is a known cause of results that cannot be reproduced [6, 24]. On explanation, SHAP [25, 26] and LIME [27] have both been applied to crop recommendation [3, 28, 29]. But impurity importance favours features with many categories [30], Shapley importances describe the model and not the world [31], and some authors argue that high-stakes settings call for inherently interpretable models [32]. On fairness, group criteria compare error rates between groups [8], individual fairness asks that similar cases be treated alike [33], and the criteria cannot all hold at once when base rates differ [34, 35]. For regression, bounded group loss [36] and loss-based reweighting [37, 38] have been proposed, bias can enter at many stages of a pipeline [9, 39], and dataset documentation helps expose it [40]. In agriculture the relevant groups are usually regions or production scales rather than protected attributes, and group-wise error is rarely reported. To our knowledge, no earlier study audits a combined recommendation and yield system. One last point concerns testing: failing to find a significant difference is not evidence that two things are equal, and equivalence needs its own test against a stated margin [56]. We apply that idea to the claim that two classifiers perform equally well.

3. Data and Problem Formulation

We use the same two public datasets as the baseline study [5]. The crop recommendation dataset [41] has 2,200 observations of nitrogen (N), phosphorus (P), potassium (K), temperature, relative humidity, soil pH and rainfall. It holds exactly 100 samples for each of 22 crops and has no missing values. Its public description says it was built by augmenting rainfall, climate and fertiliser data, so we treat it as augmented and possibly synthetic. That limits what its accuracy, robustness, disparity and explanation results can say about real farms. The crop production dataset [42] holds 246,091 district-level records of area and production, by state, district, crop year (1997–2015), season and crop. The two datasets share no variable, so any link between them has to go through crop identity. Both files came from public repository mirrors, and their provenance should be checked against primary government sources before release.

When we applied the baseline preprocessing (dropping missing production and non-positive area, trimming yield at the 0.5th and 99.5th percentiles, and keeping crops with at least 50 records), we got its counts exactly: 240,505 records, 90 crops, 33 states and 646 districts. Looking at what this pipeline keeps and discards turned up the problems listed in Table 1. After the corrections, 238,499 records remain, covering 81 crops, 33 states and 652 state–district keys, before any split-specific trimming. The record flow accounts for the whole difference of 7,592 records from the raw file. The baseline also applied a pooled percentile trim that removes a further 1,856 records. Our corrected pipeline drops that step and trims per crop inside each split instead.

Table 1
Data-quality issues in the crop production file, the corrections applied, and the record flow from the raw file (246,091 records) to the corrected data (238,499 records). Records removed and remaining are counted sequentially in the order shown

Issue	Evidence	Correction	Records removed (remaining)
Missing production	Production is missing in 3,730 records; no record has non-positive area.	Dropped, as in the baseline	3,730 (242,361)
Aggregate rows	“Oilseeds total”, “Pulses total” and “Total foodgrain” appear as crops and double-count other crops.	Removed	880 (241,481)
Crops with no recorded production	19 crops (for example Apple, Cucumber, Other Citrus Fruit, Pump Kin, Snak Guard and Water Melon) have zero production in every record; the baseline trains the model to predict zero yield for them.	Removed	621 (240,860)
Mixed production units	Coconut is recorded in nuts, so its yields are orders of magnitude above other crops. The baseline’s pooled percentile trim removed 956 of 1,958 Coconut records (48.8%) and 216 Sugarcane records, acting as a unit filter rather than an outlier filter.	Coconut excluded; per-crop trimming using training data only	1,958 (238,902)
Sparse crops	20 crops have fewer than 50 usable records (for example Lemon, Sapota, Lentil, Rubber and Coffee).	Removed	403 (238,499)
Leakage under random splitting	Under a random 80/20 split, 96.8% of test rows belong to a district–crop–season series that also occurs in training, and 85.0% have a record for an adjacent year in training.	Temporal and leave-group-out protocols added	—

Units are only partly resolved. Cotton and jute are often reported in bales, so absolute errors stay in mixed units. The crop is a model feature, and all group analyses use errors divided by the crop mean, so unit differences between crops do not enter the disparity metrics. They do, however, inflate the pooled log-scale R^2, so we also report crop-normalised error and within-crop skill scores. For record i, yield is production divided by cultivated area. Yields are positive and heavy-tailed across crops, so we train on the log scale and transform predictions back:

$$Y_i = \frac{P_i}{A_i} \tag{1}$$

$$y_i = \ln(1 + Y_i), \qquad \hat{Y}_i = \exp(\hat{y}_i) - 1 \tag{2}$$

The recommender is a multi-class classifier and the yield model is a regressor. We summarise yield accuracy with the mean absolute error (MAE), the coefficient of determination over n test records, and a per-crop skill score against a reference forecast [55]:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|, \qquad R^2 = 1 - \frac{\sum_i \left(Y_i - \hat{Y}_i\right)^2}{\sum_i (Y_i - \bar{Y})^2} \tag{3}$$

$$SS_c = 1 - \frac{\mathrm{MAE}_c^{\mathrm{model}}}{\mathrm{MAE}_c^{\mathrm{ref}}}, \qquad SS = \frac{1}{|C|}\sum_{c \in C} S\,S_c \tag{4}$$

Skill is averaged over crops (a macro-average) so that high-yield crops do not dominate. Estimated production is the predicted yield times the cultivated area, and we report it as a production-weighted error: the absolute error summed over a group, divided by that group's total observed production.

4. The LEAF-Audit Protocol and Methods

4.1 Protocol and Recommender Evaluation

LEAF-Audit splits the re-evaluation into four stages: leakage quantification (L), evaluation under distribution shift (E), attribution of group disparity (A) and feasibility coupling (F) (Fig. 1). The recommender and the yield model remain separate learners, joined by a feasibility filter (Fig. 2). Every stage is run under protocols that mimic deployment and is audited group by group. Stages L and E are applied to the recommender where they make sense, and all four stages are applied to the yield model.

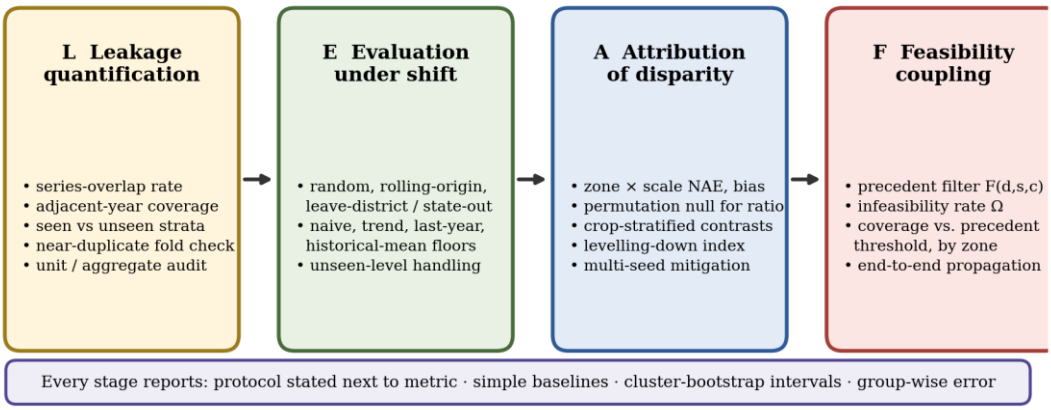


Fig. 1. The four LEAF-Audit stages and the outputs each reports

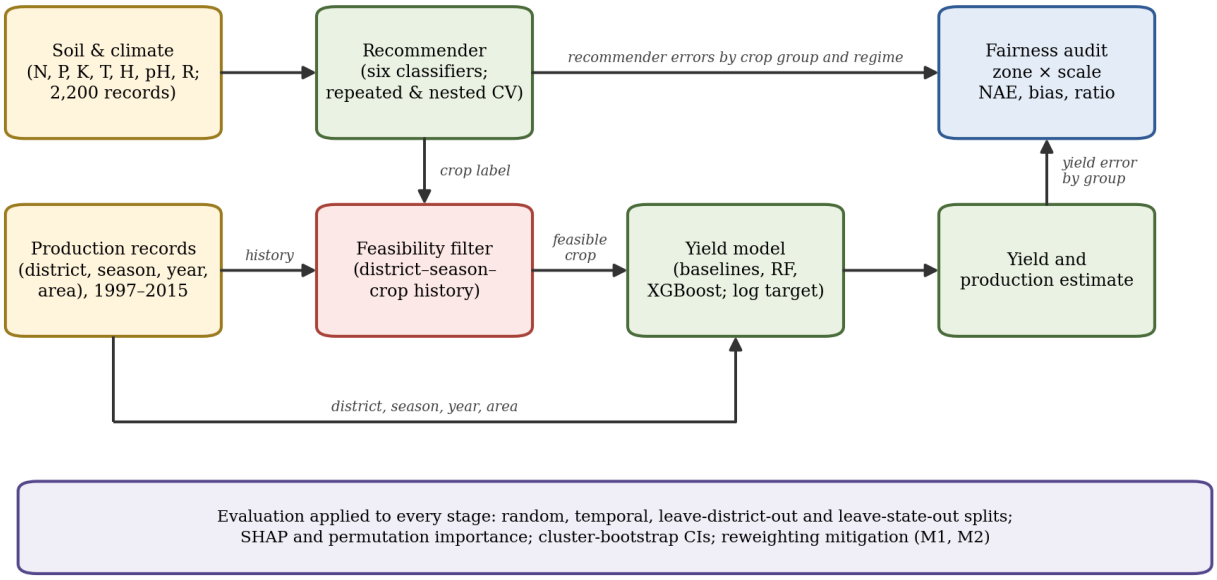


Fig. 2. System overview: recommender, feasibility filter, yield model, disparity audit and the protocols used to evaluate them

We compare six classifiers: Decision Tree, Gaussian Naive Bayes, SVM with an RBF kernel (C = 10), Logistic Regression, Random Forest (200 trees) and XGBoost [14, 15, 43]. Results are given in three ways: (a) the baseline's single stratified 80/20 split (seed 42); (b) 5×5 repeated stratified cross-validation; and (c) nested cross-validation (5 outer folds × 2 repeats) with an inner 3-fold grid search over a small hyper-parameter grid for each model. We test differences under repeated cross-validation with the variance-corrected resampled t statistic [16]. On pooled out-of-fold predictions we apply Cochran's Q test [44] across the classifiers, then exact pairwise McNemar tests [17, 45] with Holm's step-down correction [46]. We also report Wilson score intervals [47] and macro-averaged scores, since the 22 classes are balanced. A non-significant difference does not show that two models are equal, so we also test the equivalence of Random Forest and Naive Bayes with two one-sided tests (TOST) [56] on the fold-wise accuracy differences, using a margin of ±0.5 percentage points (margins of 0.25 and 1.0 points are reported too). Cross-validation folds overlap, so these p-values are only indicative.

Three checks ask whether the data themselves inflate accuracy. First, we look for exact duplicates and compute nearest-neighbour distances in standardised feature space. Second, we count same-class nearest-neighbour pairs that fall on opposite sides of the cross-validation folds and compare the count with a permuted-fold reference. Third, we inspect within-class correlation and normality, because independently sampled Gaussian features within each crop would suggest generated data. Robustness is tested in two ways: a leave-one-regime-out test, where each tercile of rainfall, temperature or humidity is held out in turn, and a noise test that adds Gaussian noise of 0.1–1.0 times each feature's own standard deviation to the test inputs.

4.2 Yield Models and Feasibility Integration

The yield features are state, district, crop year, season, crop and cultivated area. We use five baselines: (i) hierarchical historical means of log yield (crop × state × season and crop × district × season, each falling back to a coarser level when a combination is unseen); (ii) a ridge regression on one-hot encoded categories; and three forecasting floors that use the time dimension, namely (iii) the last-year value, (iv) a 3-year rolling mean and (v) the historical mean plus a per-crop linear trend. The tree models follow the baseline study. The Decision Tree has depth 20, and the Random Forest has maximum depth 20 (60 trees and a minimum leaf size of 2 in multi-fold protocols to save computation, 200 trees in the exact replication). XGBoost [15] uses 400 trees of depth 8 with native categorical splits. A "no-district" variant tests whether the district code simply enables memorisation, and a detrended-target variant fits each model to the residual from a per-crop linear trend, since trees cannot extrapolate a numeric crop year. We also tune hyper-parameters inside each protocol with time-blocked inner validation [53, 54]. Per-crop trimming thresholds (the 0.5th and 99.5th percentiles) come from the training portion only, and we also report results on untrimmed test sets.

We use five protocols: (1) a random 80/20 split, as in the baseline; (2) a temporal split that trains on 1997–2010 and tests on 2011–2015; (3) rolling-origin evaluation with expanding training windows and horizons of one to three years; (4) leave-district-out evaluation with three folds; and (5) leave-state-out evaluation with five folds. The features are identifiers only, so protocols (4) and (5) mostly measure how a model copes with an unseen category level. To account for that, we add a crop × season-only reference as the floor for spatial transfer, map unseen levels to an explicit "unknown" category with random level-dropout, and cluster the leave-state-out intervals by state. Errors are reported on the log scale, on the raw scale, in crop-normalised form (using the crop-mean yield) and as macro skill scores (Eq. (4)). Log-scale R^2 and raw-scale MAE are dominated by differences between crops, so neither can be compared with the baseline's raw-scale R^2 of 0.948, and every R^2 in this paper is labelled with its scale. To test whether leakage drives the random-split result, we stratify random-split performance by whether a test row's district–crop–season series is seen in training.

Paired differences in absolute error are tested with the Friedman test [48] and pairwise Wilcoxon signed-rank tests [49] with Holm correction, as in the baseline study. With tens of thousands of test records almost any difference is significant, so we also report the matched-pairs rank-biserial correlation [50] and base our conclusions on effect sizes and cluster-bootstrap intervals [51] that resample whole districts. Spatial protocols get state-clustered intervals, and temporal protocols get year-clustered (or two-way) intervals. With few clusters these intervals are wide, which is simply the honest answer.

For integration, let $n_{d,s,c}$ be the number of training records for district d, season s and crop c. A crop counts as feasible in a district–season if it has local precedent, and the infeasibility rate Ω is the share of mapped district–season–crop combinations without precedent:

$$\mathrm{Feas}(d,s,c) = \mathbf{1}[\, n_{\mathrm{tr}}(d,s,c) > 0 \,] \tag{5}$$

$$\Omega = 1 - \frac{1}{N} \sum_{(d,s,c)} \mathrm{Feas}\,(d,s,c) \tag{6}$$

Here N is the number of combinations considered. Each mapped crop is paired with every district–season, so Ω is large by construction. It measures how location-blind such a pairing is; it is not the error rate of a deployed recommender. In the integrated pipeline, class probabilities are multiplied by Feas, and the feasible crop with the highest probability goes on to the yield model. Crop

names are harmonised through an explicit mapping table, and crops with no equivalent are reported as unmapped rather than forced onto a lookalike. No dataset links soil measurements to districts, so the end-to-end test is a simulation. For each 2011–2015 production record whose crop has a recommendation equivalent, we draw a soil–climate profile of that crop from the held-out recommendation pool, classify it, and pass the result to the yield model with the record's district, season, year and area. Because the profile comes from the true crop's own pool, classifier agreement is in-distribution accuracy. And because the historically grown crop is the reference, "agreement" measures consistency with past practice, not agronomic suitability. We vary the precedent threshold (at least 1, 2 or 3 earlier records) and the zone.

4.3 Disparity, Mitigation and Explainability

For yield, we assign states to six regional zones: North (Chandigarh, Delhi, Haryana, Himachal Pradesh, Jammu and Kashmir, Punjab, Rajasthan, Uttarakhand), Central (Chhattisgarh, Madhya Pradesh, Uttar Pradesh), South (Andhra Pradesh, Karnataka, Kerala, Puducherry, Tamil Nadu, Telangana), East (Bihar, Jharkhand, Odisha, West Bengal), West (Dadra and Nagar Haveli, Goa, Gujarat, Maharashtra) and North-East (Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, Tripura). The Andaman and Nicobar Islands form a seventh group that is too small for stable estimates in most protocols. These are regional groupings of states, not the official agro-climatic zones. Within each crop we split cultivated area into terciles (small, medium, large). Area here is the district total and not farm size, so the grouping is only a proxy for production scale. Crossing zone with scale gives up to 18 cells. For recommendation, the groups are crop categories and terciles of rainfall, temperature and humidity. All of these groups are used for error analysis only; none of them is a group of people. For a group S_g, the normalised absolute error and the signed relative bias are

$$\mathrm{NAE}_g = \frac{1}{|S_g|}\sum_{i\in S_g}\frac{|Y_i - \hat{Y}_i|}{\bar{Y}_{c(i)}}, \qquad B_g = \frac{1}{|S_g|}\sum_{i\in S_g}\frac{\hat{Y}_i - Y_i}{\bar{Y}_{c(i)}} \tag{7}$$

and disparity is summarised by the worst-to-best ratio D and the gap Δ, taken over groups with at least 300 test records:

$$D = \frac{\max_g \mathrm{NAE}_g}{\min_g \mathrm{NAE}_g}, \qquad \Delta = \max_g \mathrm{NAE}_g - \min_g \mathrm{NAE}_g \tag{8}$$

Both are reported with 95% cluster-bootstrap intervals that resample whole districts (500 resamples). The ratio D is a max/min statistic and is biased upward when groups are noisy, so we compare it with a permutation null: districts are randomly reassigned to pseudo-zones of the same sizes B = 1,000 times, giving

$$p_{\mathrm{perm}} = \frac{1 + \#\{\, b : D_b^* \geq D_{\mathrm{obs}} \,\}}{B + 1} \tag{9}$$

Zones also differ in crop mix. We therefore estimate zone contrasts from a model of NAE with crop fixed effects, and compare each zone with the historical-mean baseline within the same cell (these p-values are not adjusted for multiplicity and are indicative). The ratio and the gap are descriptive. They do not separate irreducible variability from model error, and we keep between-

group disparity apart from overall calibration, meaning a bias shared by all groups, which is a modelling issue and not a disparity.

We compare training-time reweighting (M1, M2 and a capped variant of M1) and one non-reweighting reference (M3) with an unmitigated baseline (M0). All use XGBoost (250 trees). Each training record i belongs to one zone × scale cell g_i and gets the weight $w_{g_\{i\}}$, and the model minimises the weighted squared error on the log target:

$$L_w = \frac{1}{n}\sum_{i=1}^{n} w_{g(i)}\,(y_i - \hat{y}_i)^2 \tag{10}$$

M1 (inverse-frequency) gives every cell the same total weight, which targets under-representation (N training records, G cells, n_g records in cell g):

$$w_g = \frac{N}{G\,n_g} \tag{11}$$

M2 (context-aware adaptive reweighting) targets unequal difficulty. All weights start equal, and over three rounds r they are raised for cells whose validation error is above average, in the spirit of worst-group reweighting [37, 38]:

$$w_g^{(r+1)} \propto w_g^{(r)} \exp\left[\eta\left(\frac{\mathrm{NAE}_g^{(r)}}{\overline{\mathrm{NAE}}^{(r)}} - 1\right)\right], \qquad \eta = 1 \tag{12}$$

Validation data are the last two training years (temporal protocol) or a random 15% of the training set (random protocol), and neither mitigation touches test data. We claim no novelty for the algorithm; it is called context-aware only because its groups are defined by regional and scale context. The weights are tuned on the same cells that we use to report disparity, so any gain is in-sample for that grouping. To deal with this, we repeat each run over five seeds with paired bootstrap differences, evaluate on a grouping not used for tuning (the worst-to-best ratio over the 33 states), and cap weights at a fixed multiple of the mean weight to see whether M1 fails only because very small cells receive very large weights. M3 corrects predictions with a zone × crop offset estimated from the training data. All methods share the same cells, splits and seeds.

Levelling-down index. A lower worst-to-best ratio can appear because the worst cell improves or because the best cell gets worse. Let a be the error of the worst cell and b the error of the best cell, so the ratio is r = a/b, with subscripts 0 and 1 for the baseline and the mitigated model. When the ratio falls ($r_1 < r_0$),

$$\mathrm{LDI} = \frac{\ln(b_1/b_0)}{\ln(r_0/r_1)}, \qquad \ln\frac{r_0}{r_1} = \ln\frac{a_0}{a_1} + \ln\frac{b_1}{b_0} \tag{13}$$

LDI is the share of the log-ratio reduction that comes from best-cell degradation. A value of 0 means the reduction came entirely from improving the worst cell, 1 means it came entirely from worsening the best cell, and above 1 means the worst cell also got worse. We report LDI next to the ratio, the worst-cell error and the pooled accuracy, so a lower ratio is never presented as a gain on its own. The worst and best cells are picked from test data for each model separately, and reweighting towards high-error cells takes weight away from low-error cells, so LDI is a descriptive diagnostic, not a test.

Explainability. Global SHAP importance is the mean absolute contribution of a feature [25, 26]. For the 22-class recommender we report two importances per feature: the average over all 22 class outputs, and the contribution towards the predicted class only. Both come from exact TreeSHAP on

the Random Forest using the 440 held-out samples, with bootstrap intervals and the share of resamples in which each feature ranks first. For yield, we compute exact tree contributions for an XGBoost model on 3,000 temporal-test records and cross-check them against permutation importance (the increase in log-scale RMSE when a feature is shuffled). Permuting nested features such as state and district can create impossible combinations, so we read those results with care. We assess stability with five refits under different seeds. All experiments ran on one machine with two CPU cores, using scikit-learn [43], XGBoost and the SHAP library and fixed seeds (42 for splits and models, 0–4 for the mitigation runs). The versions were Python 3.13.16, NumPy 2.1.3, pandas 2.2.3, scikit-learn 1.6.1, XGBoost 3.4.1, SHAP 0.52.0 and statsmodels 0.15.0.

5. Results

5.1 Crop Recommendation

On the baseline's single split, our implementation reproduces the published figures. Random Forest and Naive Bayes each make 2 errors on 440 test samples (99.55%), while Decision Tree makes 9, SVM 5, XGBoost 6 and Logistic Regression 12. One error moves accuracy by 0.23 percentage points, and the Wilson 95% interval for Random Forest is 98.36–99.88%, so a ranking from one split says very little. Under 5×5 repeated cross-validation (Table 2), Random Forest (99.52 ± 0.27%) and Naive Bayes (99.47 ± 0.24%) stay tied (corrected t-test p = 0.75). Nested cross-validation with tuning leaves them almost identical (99.45% and 99.41%). Cochran's Q over the six classifiers is significant (Q = 108.6, p < 0.001), but the Holm-corrected McNemar tests trace this to Logistic Regression, SVM and Decision Tree doing worse than Random Forest. They do not show Random Forest beating Naive Bayes (p = 1.000) or XGBoost (p = 0.96). The 95% interval for the Random Forest − Naive Bayes difference in fold-wise accuracy runs from −0.25 to +0.34 percentage points (mean +0.05). TOST rejects a difference larger than ±0.5 points (p = 0.002) and larger than ±1.0 point (p < 0.001), but not one larger than ±0.25 points (p = 0.080). The two models are therefore equivalent within half a point, but not within a quarter of a point.

Table 2
Crop recommendation: single-split, repeated-CV and nested-CV accuracy (%). The p-values compare each model with Random Forest. McNemar tests use pooled out-of-fold predictions ($n = 2{,}200$) from the first CV repeat

Model	Baseline split (errors/440)	5×5 CV acc. ± SD	5×5 CV macro-F1	Nested CV acc.	Corrected t-test p	McNemar p (Holm)
Random Forest	99.55 (2)	99.52 ± 0.27	99.52	99.45	—	—
Gaussian Naive Bayes	99.55 (2)	99.47 ± 0.24	99.47	99.41	0.751	1.000
XGBoost	98.64 (6)	99.24 ± 0.45	99.23	99.23	0.314	0.961
SVM (RBF)	98.86 (5)	98.78 ± 0.41	98.78	98.59	0.004	0.004
Decision Tree	97.95 (9)	98.73 ± 0.64	98.72	98.34	0.045	<0.001
Logistic Regression	97.27 (12)	97.14 ± 0.49	97.12	98.25	<0.001	<0.001

The data contain no exact duplicate rows. The closest pair of samples is 0.10 standard deviations apart in standardised feature space (median nearest-neighbour distance 0.38), and 97.7% of nearest neighbours share the sample's class. In the fold-crossing check, 81.8% of same-class nearest-

neighbour pairs fall on opposite sides of the folds, against 80.0% in the permuted-fold reference (95th percentile 81.6%; p = 0.038). The excess is under two percentage points, so near-duplicate clusters from augmentation do not seem to inflate cross-validated accuracy much, though the check is not conclusive. Gaussian Naive Bayes matching Random Forest points the same way, as do the weak within-class correlations (mean absolute off-diagonal correlation 0.075; maximum 0.29 over the 22 crops). Both fit features sampled almost independently within each crop, which is what we would expect if the data were generated or augmented from per-crop distributions.

High accuracy inside the sampled climate range does not mean the model can extrapolate. When we hold out one tercile of rainfall, temperature or humidity, Random Forest accuracy on the held-out regime ranges from 17.5% to 99.6% (Fig. 3). Most of this loss is not a generalisation failure in the usual sense. A closed-set classifier cannot predict a crop it never saw, and 54.5% of the low-rainfall and 54.6% of the high-rainfall test samples belong to crops absent from training. This tells us that crop identity in this dataset is largely fixed by climate niche. On crops that were seen, accuracy stays between 96.2% and 100% in eight of nine held-out regimes and falls to 79.0% only for mid-humidity. Noise matters too. With test-time noise of 0.25 feature standard deviations, accuracy drops to 81.1% (Random Forest), 80.2% (Naive Bayes), 92.6% (SVM), 89.5% (Logistic Regression), 70.1% (Decision Tree) and 69.8% (XGBoost). At 0.1 standard deviations XGBoost already falls to 88.9% while SVM stays at 98.8%. The ranking depends on the stress applied, which is one more reason not to read a 0.01-point difference as superiority. Noise is added equally to every feature, so this is a stress test and not a model of real measurement error.

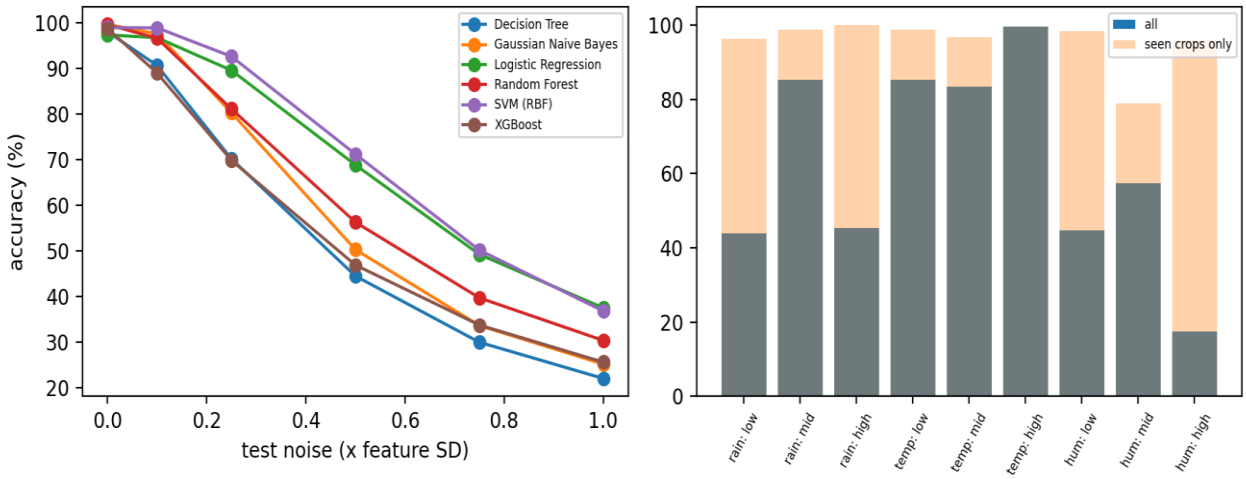


Fig. 3. Robustness of the recommender. (a) Accuracy under test-time Gaussian noise. (b) Random Forest accuracy when one tercile of rainfall, temperature or humidity is withheld; the dashed line is in-distribution accuracy

Errors are not spread evenly across groups, but here the groups are crops or climate bands, so this is class-wise error analysis and not a fairness analysis in the sense of Section 5.3. Cereals have the lowest accuracy (96.5% for Random Forest), mostly because of rice, whose pooled out-of-fold recall is 93.0%. The 3.5-point spread across crop categories amounts to seven errors among 200 cereal samples. Across rainfall, temperature and humidity terciles the gaps are about one point for Random Forest (1.09, 0.96 and 1.09). Inside its sampling range the recommender is accurate almost everywhere. The harder question is who is served when conditions fall outside it.

5.2 Yield Prediction under Realistic Protocols

We first reproduce the yield stage. With the baseline preprocessing and a random split, our Random Forest (200 trees) gets R^2 = 0.9475 (raw scale), MAE = 0.7033 and RMSE = 2.6432, and Linear Regression on label codes gets R^2 = 0.094, in line with the published values. Table 3 repeats the comparison on the corrected data under four protocols and adds the historical-mean, crop × season and one-hot baselines that the original comparison lacked.

Table 3
Yield models under four protocols

Model	Random R^2	Random MAE	Temporal R^2	Temporal MAE	District R^2	District MAE	State R^2	State MAE
Historical mean (crop×district×season)	0.947	0.88	0.904	1.25	0.936	1.02	0.786	2.15
Crop × season mean	0.820	1.98	0.821	2.16	0.820	1.98	0.786	2.15
Ridge, one-hot	0.836	1.83	0.839	2.05	0.832	1.86	0.794	2.11
Decision Tree	0.943	0.82	0.827	1.47	0.904	1.16	0.532	2.98
Random Forest	0.962	0.71	0.851	1.31	0.936	0.99	0.651	2.60
XGBoost	0.961	0.75	0.923	1.10	0.775	2.39	0.244	4.70
Random Forest, no district	0.955	0.81	0.843	1.39	0.940	0.95	0.628	2.66
XGBoost, no district	0.959	0.80	0.921	1.19	0.948	0.89	0.450	3.50
Random Forest, no state/district	—	—	—	—	0.827	1.84	0.737	2.36
XGBoost, no state/district	—	—	—	—	0.838	1.79	0.762	2.26
XGBoost, unknown level	—	—	—	—	0.936	1.03	0.791	2.08

R^2 on the log scale and MAE (yield units per hectare, raw scale). Leave-state-out is pooled over five folds and leave-district-out over three; for unseen districts and states the historical mean uses its coarser fallback levels (in leave-state-out it falls back to the crop × season mean, so the two rows coincide). Linear Regression on label-encoded categories reaches only R^2 = 0.06–0.14 (log scale) in every protocol and is omitted. Variants without state and district, and with an explicit unknown level, are reported only for the spatial protocols

The random split is optimistic and rewards memorisation. Under it Random Forest is the best model (log-scale R^2 = 0.962, MAE 0.71) and beats the historical-mean baseline in MAE (difference −0.170 [−0.197, −0.145], 95% district-cluster bootstrap). Yet a historical-mean lookup table comes within 0.016 of it (log-scale R^2 0.947 against 0.962), so most of the apparent skill needs no learning at all. The published 0.948 is on the raw scale and is not directly comparable; on the corrected data the raw-scale R^2 of Random Forest is 0.946. Splitting the random-split test set shows where the advantage sits. On the 45,735 test rows (96.8%) whose district–crop–season series is seen in training, Random Forest and XGBoost both reach a log-scale R^2 of 0.964 against 0.948 for the historical mean. On the 1,535 rows (3.2%) whose series is unseen, Random Forest falls to 0.917 and XGBoost to 0.878, and the historical mean (0.921) is ahead of both. A random split therefore measures interpolation inside known series. That suits gap-filling but not forecasting.

Once the task becomes forecasting, Random Forest loses its edge. On the temporal split it drops to a log-scale R^2 of 0.851 and is worse than the historical mean in point estimate. The MAE difference is +0.060 [+0.026, +0.095] with district clusters, but [−0.061, +0.213] with state clusters and [−0.019, +0.121] with year clusters, so the deficit is not firmly established. XGBoost with native categorical splits is the best temporal model (MAE 1.10; −0.212 [−0.286, −0.152] relative to Random Forest) and, unlike Random Forest, beats the historical mean (MAE difference −0.152, with intervals of [−0.203,

−0.106] for district clusters, [−0.373, −0.028] for state clusters and [−0.187, −0.118] for year clusters). Its raw-scale R^2 (0.886) and log-scale R^2 (0.923) are also ahead of the baseline (0.862 and 0.904), and its macro skill score against the historical mean is +0.108.

Table 4
Forecasting floors, tuned models and rolling-origin evaluation (temporal split: train 1997–2010, test 2011–2015; metrics on the per-crop-trimmed test set). Untrimmed MAE is computed before trimming the test set. Rolling-origin MAE averages horizons of one to three years over expanding windows, with 95% bootstrap intervals

Model	Raw R^2	Log R^2	MAE	Macro skill vs hist. mean	Untrimmed MAE	Rolling-origin MAE [95% CI]
Last-year value	0.847	0.902	1.207	+0.038	22.28	1.056 [1.009, 1.103]
3-year rolling mean	0.856	0.908	1.212	+0.051	22.28	1.021 [0.979, 1.067]
Historical mean + crop trend	0.802	0.810	1.324	−0.078	22.39	1.126 [1.078, 1.178]
Crop × season mean	0.678	0.821	2.157	−0.437	23.25	2.068 [1.968, 2.167]
Historical mean	0.862	0.904	1.254	0 (reference)	22.33	1.075 [1.029, 1.124]
Ridge, one-hot	0.718	0.839	2.052	−0.545	23.13	1.946 [1.859, 2.023]
LinReg (label codes)	−0.018	0.056	4.308	−4.578	25.38	—
Decision Tree	0.789	0.827	1.471	−0.356	22.53	—
Random Forest	0.835	0.851	1.314	−0.095	22.39	1.126 [1.082, 1.178]
XGBoost	0.886	0.923	1.102	+0.108	22.17	0.973 [0.932, 1.015]
Random Forest, no district	0.824	0.843	1.391	−0.143	22.46	—
XGBoost, no district	0.878	0.921	1.186	+0.068	22.25	—
RF, tuned, detrended	0.768	0.811	1.361	−0.133	22.44	1.140 [1.089, 1.193]
XGBoost, tuned, detrended	0.183	0.822	1.602	−0.084	22.67	1.101 [1.038, 1.165]

Simple floors are hard to beat (Table 4). The last-year value (MAE 1.21) and the 3-year rolling mean (1.21) sit close to the historical mean (1.25) and ahead of Random Forest (1.31). Adding a per-crop linear trend to the historical mean makes it worse (MAE 1.32; +0.070 [+0.026, +0.112] with district clusters), which suggests that trends fitted on 1997–2010 do not carry into 2011–2015. XGBoost's paired MAE differences are −0.105 [−0.149, −0.062] against the last-year value and −0.111 [−0.153, −0.068] against the rolling mean with district clusters, and [−0.145, −0.065] and [−0.123, −0.098] with year clusters. The state-clustered intervals, [−0.295, +0.009] and [−0.291, +0.006], only just include zero. Rolling-origin evaluation gives the same ordering (XGBoost 0.973 [0.932, 1.015], 3-year rolling mean 1.021, last-year value 1.056, historical mean 1.075 and Random Forest 1.127), although the XGBoost interval overlaps that of the rolling mean. Error grows with the horizon for every model (XGBoost 0.84, 1.03 and 1.07 at one, two and three years).

Tuning with time-blocked validation and detrending the target did not help. Tuned, detrended XGBoost is worse than untuned XGBoost (MAE +0.500 [+0.331, +0.698] with district clusters) and worse than the historical mean (+0.348 [+0.207, +0.511]). Tuned, detrended Random Forest (MAE 1.36) is no better than the untuned forest (1.31). Untuned detrended XGBoost produced extreme extrapolated predictions (up to 765 yield units; raw-scale R^2 = −0.05), and clipping them to each crop's

training range still left MAE (1.33) above both plain XGBoost and the historical mean. So the claim that tree models cannot beat simple baselines when forecasting holds for Random Forest and the detrended variants, but not for XGBoost with native categorical splits. Rerunning both families with several encodings shows that the advantage belongs to native categorical splitting specifically. XGBoost with integer-coded categories has an MAE of 1.35, worse than the historical mean (+0.099 [+0.068, +0.131] with district clusters) and 0.251 worse than native XGBoost. The one-hot (MAE 1.44) and target-encoded (1.40) Random Forests are worse than the integer-coded forest (1.31).

Spatial transfer is hard for every model. With whole states withheld, Random Forest reaches a log-scale R^2 of 0.651 against 0.786 for the historical mean (which here falls back to the crop × season mean). The best models are the one-hot ridge regression (0.794) and XGBoost with an explicit unknown level (0.791). The features contain no spatial covariates, so these protocols mostly show how a model copes with an unseen category level. Native categorical XGBoost copes poorly with unseen districts and states (log-scale R^2 of 0.775 and 0.245), because an unseen level has no learned effect, and removing the district feature restores it to 0.948 on new districts. Mapping unseen levels to an explicit unknown category and training with level dropout lifts XGBoost from 0.775 to 0.936 on new districts and from 0.245 to 0.791 on new states. That is above the crop × season floor in the district protocol (0.820) and just above it in the state protocol (0.786). Differences between model families in these protocols therefore mostly reflect how unseen levels are handled.

Finally, the Friedman tests on the absolute errors of Linear Regression, Decision Tree and Random Forest are overwhelmingly significant (χ^2 = 39,658 for the random split and 31,363 for the temporal split), as in the baseline study. The effect sizes, though, differ widely. The rank-biserial correlation for Random Forest versus Linear Regression is 0.90 (random) and 0.80 (temporal), whereas for Random Forest versus Decision Tree it is only 0.12 and 0.15. A p-value below 10^{-100} here reflects sample size, not an important difference. Untrimmed raw-scale R^2 is about 0.001 for every model, because a few extreme records dominate the squared error, but the model ordering on untrimmed log-scale R^2 is unchanged (XGBoost 0.894, 3-year rolling mean 0.877, historical mean 0.872, Random Forest 0.821).

5.3 Disparity Audit and Bias Mitigation

The data are uneven before any model is trained. North-East India supplies 11.5% of the records but only 3.2% of the cultivated area. The North supplies 12.5% of records and 17.8% of area, and the West 8.8% and 16.2%. Any metric averaged over records therefore gives small-area North-Eastern districts far more weight than large-area Western ones. Missing production values are concentrated in the North (6.28% of records, against 0.05% in the North-East). The baseline's pooled percentile trim removes 8.9% of the island records, 2.3% of the southern and 1.6% of the North-Eastern records, but 0.2% or fewer in the North, East and West, mostly because it deletes coconut and sugarcane records. These are representation and measurement biases [9] that enter before any learning begins.

Table 5 gives normalised error by zone. Under the temporal protocol, XGBoost has its highest normalised error in the South (0.334) and its lowest in the North-East (0.186), a worst-to-best ratio of 1.79 [1.64, 1.96]. The ratio is similar for Random Forest (1.77 [1.65, 1.97]) and slightly smaller for the historical-mean baseline (1.68 [1.57, 1.83]). The South is the worst-served zone under all three temporal models, but the choice of model changes how other regions fare. Random Forest does worse than the baseline in the North-East (0.271 against 0.241) and the South (0.422 against 0.399), whereas XGBoost does better in both. Under the random split the ratio is 1.96 [1.69, 2.24] for XGBoost and 1.78 [1.58, 2.01] for Random Forest, so the disparity is not just an artefact of forecasting.

Table 5
Normalised absolute error (NAE) by zone, worst-to-best ratio D with its 95% cluster-bootstrap interval, absolute gap Δ and largest absolute signed zone bias, for the temporal and random protocols. Islands (fewer than 300 test records) are excluded from the ratios. Bias is the mean signed relative error; under the temporal protocol the signed zone biases of XGBoost are Central −0.029, East −0.104, North −0.034, North-East −0.126, South −0.076 and West +0.056

Protocol: model	North	Central	South	East	West	North-East	Ratio D [95% CI]	Gap Δ	Max abs. bias
Temporal: Historical mean	0.291	0.238	0.399	0.305	0.281	0.241	1.68 [1.57, 1.83]	0.161	0.236
Temporal: Random Forest	0.251	0.238	0.422	0.252	0.271	0.271	1.77 [1.65, 1.90]	0.184	0.133
Temporal: XGBoost	0.242	0.227	0.334	0.254	0.277	0.186	1.79 [1.64, 1.96]	0.148	0.126
Random: Historical mean	0.263	0.178	0.278	0.201	0.245	0.137	2.02 [1.73, 2.28]	0.141	0.039
Random: Random Forest	0.226	0.153	0.211	0.166	0.219	0.127	1.77 [1.58, 2.01]	0.099	0.040
Random: XGBoost	0.221	0.144	0.232	0.171	0.213	0.118	1.96 [1.69, 2.24]	0.114	0.045

Is this ratio larger than chance, and is it only a crop-mix effect? It is larger than chance. For six pseudo-zones of the same sizes, the permutation-null distribution of D has a median of 1.11 and a 95th percentile of 1.18 for XGBoost under the temporal split (1.13 and 1.22 for Random Forest). The observed value is 1.79 (1.77 for Random Forest; p_{perm} = 0.001, the smallest value attainable with 1,000 permutations), and the random split gives the same answer. Nor is it only crop mix. Zone contrasts from the NAE model with crop fixed effects show higher normalised error than Central India for XGBoost in the South (+0.104, s.e. 0.009), East (+0.041), West (+0.040) and North (+0.029, p = 0.034), and lower in the North-East (−0.062). After subtracting the historical-mean error within each cell, the improvement of XGBoost over the baseline is larger than in Central India in the East, North, North-East and South (contrasts −0.040 to −0.044, all p < 0.001) and not different in the West (+0.012, p = 0.17). Random Forest shows the same pattern with smaller contrasts. The zone gaps are therefore largely a property of the data and the task, since the trivial baseline shows them too, and they remain after controlling for crop.

Large-area districts have somewhat higher normalised error than small-area ones (XGBoost 0.272 against 0.246; Random Forest 0.311 against 0.272; historical mean 0.311 against 0.282). Production-weighted error points the other way. For XGBoost it is 0.171 in the large tercile, against 0.195 in the small and 0.216 in the medium tercile. Across zones, the production-weighted error of XGBoost is highest in the East (0.271) and lowest in the North (0.130) and Central India (0.147), and the South has a low production-weighted error (0.151) despite having the highest normalised error. Both views should be reported. Every area group is under-predicted under the temporal split (mean signed relative error between −0.057 and −0.061 for XGBoost, −0.085 and −0.094 for Random Forest and −0.158 and −0.172 for the historical mean). This bias is almost the same in every area group, so it is a calibration problem shared by all groups and not a between-group disparity. The zone-level bias does differ, from −0.126 (North-East) to +0.056 (West) for XGBoost. Combining zone and scale gives a worst-to-best cell ratio of 1.96 for XGBoost (interval 1.80–2.22), with South small-area cells worst (NAE 0.341) and North-East small-area cells best (0.175).

Bias mitigation does not behave the way the ratios alone suggest (Table 6). Inverse-frequency reweighting (M1) does not reduce disparity. Under the temporal protocol it raises the zone ratio from 1.80 to 1.96 and the cell ratio from 1.97 to 2.29 (paired differences over seeds +0.17 [+0.08, +0.25] and +0.32 [+0.23, +0.41]; the cell ratio rose in all five seeds), and capping the weights does not rescue it. This suggests that the disparity does not mainly come from too few training records in some cells. A likelier cause is that some cells are harder to predict, because of crop mix, seasonal structure and reporting quality, but we did not test that directly. Adaptive reweighting (M2) targets difficulty instead. It lowers the zone ratio from 1.80 to 1.72 (temporal) and from 1.93 to 1.72 (random), and the cell ratio from 1.97 to 1.83 and from 2.72 to 2.31, with paired differences negative in all five seeds (zone ratio −0.081 [−0.118, −0.043] and −0.209 [−0.241, −0.178]). The reduction also shows up in the state-level ratio, a grouping not used for tuning, under the random protocol (−0.449 [−0.525, −0.373]) but not under the temporal protocol (+0.083 [−0.092, +0.257]). Pooled accuracy is essentially unchanged in the random protocol (MAE −0.001 [−0.009, +0.007]) and slightly worse in the temporal protocol (MAE +0.025 [+0.008, +0.042], about 2%). The zone–crop offset (M3) is no better: in the temporal protocol it raises MAE by 0.062 [+0.044, +0.080] and the worst-cell error by 0.035.

Table 6
Effect of reweighting on accuracy and group disparity (XGBoost, 250 trees; mean over five seeds, with the standard deviation in parentheses for the ratios). Ratios are worst-to-best NAE over zones, zone × scale cells or states (the state ratio is a grouping not used for tuning)

Protocol	Method	R^2 (raw)	MAE	Zone ratio	Cell ratio	State ratio	Worst-cell NAE	Max abs. cell bias
Temporal	M0 baseline	0.883	1.141	1.80 (0.04)	1.97 (0.05)	2.80 (0.21)	0.352	0.136
Temporal	M1 inverse-frequency	0.876	1.159	1.96 (0.10)	2.29 (0.10)	3.19 (0.68)	0.367	0.132
Temporal	M1 capped	0.882	1.140	1.89 (0.05)	2.21 (0.07)	2.88 (0.21)	0.352	0.132
Temporal	M2 adaptive	0.878	1.166	1.72 (0.03)	1.83 (0.03)	2.88 (0.14)	0.348	0.139
Temporal	M3 zone–crop offset	0.874	1.203	1.88 (0.03)	2.08 (0.06)	2.76 (0.22)	0.387	0.143
Random	M0 baseline	0.935	0.760	1.93 (0.06)	2.72 (0.12)	3.77 (0.15)	0.283	0.068
Random	M1 inverse-frequency	0.935	0.774	2.00 (0.06)	2.72 (0.09)	3.91 (0.13)	0.272	0.054
Random	M1 capped	0.935	0.768	2.00 (0.06)	2.77 (0.10)	3.89 (0.14)	0.270	0.055
Random	M2 adaptive	0.938	0.760	1.72 (0.04)	2.31 (0.09)	3.32 (0.18)	0.273	0.052
Random	M3 zone–crop offset	0.934	0.766	1.93 (0.06)	2.72 (0.10)	3.79 (0.14)	0.286	0.065

The levelling-down index (Eq. (13); Table 7) shows why the gain from M2 is largely apparent. Under M2 the worst cell improves by only 1.1% (temporal) and 3.5% (random), while the best cell gets worse by 6.0% and 13.6%. The mean LDI is 0.82 (SD 0.26) in the temporal protocol and 0.79 (SD 0.07) in the random protocol, so about four-fifths of the reduction in the log cell ratio comes from degrading the best-served cell. In the temporal protocol the LDI of individual seeds ranges from 0.44 to 1.12, and in one seed the worst cell also got worse. M1 behaves differently. In the temporal protocol it lowers the best cell's error by about 10% and raises the worst cell's by 4%, so its ratio rises. That is the best cell improving while the worst cell suffers a little, something the ratio alone cannot show. The LDI is undefined when the ratio does not fall, so we do not report it for M1 and M3 under the random protocol. These results rest on five seeds and one model family, and the zone and cell reductions are in-sample for the grouping used for tuning, so the paired intervals should be read

as indicative. In these runs, weighting by difficulty lowered the ratios with little change in average error, but mostly by raising the error of the best-served cells.

Table 7
Levelling-down decomposition of the mitigation results over five seeds: mean ± SD percentage change in worst-cell and best-cell NAE relative to M0, the paired difference in the cell ratio with its 95% interval, the number of seeds in which the cell ratio fell, and the levelling-down index (LDI), reported only for M2, where the ratio falls in all five seeds

Protocol	Method	Δ worst cell (%)	Δ best cell (%)	Δ cell ratio [95% CI]	Seeds with lower ratio	LDI (mean ± SD)
Temporal	M1 inverse-frequency	+4.2 ± 3.8	−10.4 ± 2.1	+0.321 [+0.234, +0.407]	0/5	n/a
Temporal	M1 capped	+0.2 ± 0.7	−10.9 ± 2.7	+0.245 [+0.163, +0.328]	0/5	n/a
Temporal	M2 adaptive	−1.1 ± 1.3	+6.0 ± 3.0	−0.132 [−0.215, −0.049]	5/5	0.82 ± 0.26
Temporal	M3 zone–crop offset	+10.1 ± 1.0	+4.0 ± 0.4	+0.114 [+0.092, +0.137]	0/5	n/a
Random	M1 inverse-frequency	−4.0 ± 1.2	−4.3 ± 1.2	+0.006 [−0.056, +0.069]	3/5	n/a
Random	M1 capped	−4.8 ± 1.2	−6.6 ± 2.1	+0.053 [−0.065, +0.170]	2/5	n/a
Random	M2 adaptive	−3.5 ± 1.6	+13.6 ± 2.9	−0.409 [−0.526, −0.293]	5/5	0.79 ± 0.07
Random	M3 zone–crop offset	+1.0 ± 0.9	+0.9 ± 1.0	+0.001 [−0.036, +0.038]	3/5	n/a

5.4 Integration, Feasibility and Explainability

Of the 22 recommendation crops, 16 can be linked to a crop in the cleaned production data. The other six are apple and watermelon (zero production in every record), coconut (mixed units), coffee and kidney beans (fewer than 50 usable records) and muskmelon (no equivalent category). Coffee matters for the South, the worst-served zone. We report these crops as unmapped and do not assign them to a lookalike. Without a filter, the recommender is location-blind. When a random soil profile is paired with a random district and season, 84.7% of the recommended crops have no record in that district–season before 2011, and 27.3% cannot be passed to the yield model at all, so this figure mixes a data-linkage failure with true infeasibility. Among mapped crops the rate is 79.0% in the sample (Ω = 79.1% over all mapped combinations). It ranges from 71.8% in Central India and 72.2% in the South to 85.9% in the North-East and 89.6% in the Andaman and Nicobar Islands, and requiring at least two or three earlier records raises Ω to 82.9% and 85.2%. These figures come from random pairing and show that the two datasets cannot be linked. They are not the failure rate of a real recommender.

The feasibility filter is conservative but not unreasonable. Of the crop–district–season combinations of mapped crops observed in 2011–2015, 88.9% had also appeared before 2011, so the filter would have admitted most real cultivation and blocks mainly new introductions. In the simulation (Table 8), the unconstrained pipeline recovers the true crop 98.4% of the time but still recommends a crop with no local history in 8.9% of cases. The filtered pipeline (threshold 1) cuts that to zero, at the cost of giving no estimate for 6.9% of records (coverage 93.1%), with crop agreement

of 98.1% among covered records (91.4% over all records). At least 7.3 of the 8.9 percentage points of no-history recommendations are correct recommendations of the recorded crop that simply had no earlier record in that district and season. The filter therefore blocks mainly crops that farmers really did start growing. Raising the threshold to two and three earlier records lowers coverage from 93.1% to 91.4% and 89.8% and agreement among covered records from 98.1% to 97.9% and 97.7%, while the normalised yield error stays at 0.290, 0.290 and 0.287 (oracle 0.265; unconstrained 0.308). By zone, coverage is lowest in the North-East (86.3%) and highest in the East (99.6%), so the data-poor North-East is served worst by the filter. The soil profile is drawn from the true crop's own pool, and the production record serves as the reference, so the simulation cannot validate the soil–location pairing or agronomic suitability. We treat the filter as a diagnostic of how location-blind the two datasets are, not as a validated component of a deployable system.

Table 8
Integration diagnostics. The simulation uses the 17,521 production records of 2011–2015 whose crop maps to a recommendation crop, with yield from XGBoost trained on 1997–2010; normalised yield error is the mean absolute error divided by the crop-mean yield. Thresholds are the minimum number of earlier records that count as local precedent

Quantity	Value
Recommendation crops mappable to production data (of 22)	16
Infeasibility rate Ω over mapped district–season–crop combinations (threshold 1 / 2 / 3)	79.1% / 82.9% / 85.2%
Location-blind recommendations that cannot be mapped to a production crop	27.3%
Location-blind recommendations with no local history (unmapped counted as no history)	84.7%
Observed 2011–15 combinations already seen before 2011 (filter recall), mapped crops / all crops	88.9% / 84.9%
Simulation, oracle crop: normalised yield error	0.265
Simulation, unconstrained: crop agreement / no-history share / error	98.4% / 8.9% / 0.308
Simulation, filtered (threshold 1): coverage / agreement among covered / agreement over all records / no-history share / error	93.1% / 98.1% / 91.4% / 0.0% / 0.290
Simulation, filtered (threshold 2): coverage / agreement among covered / agreement over all records / no-history share / error	91.4% / 97.9% / 89.4% / 0.0% / 0.290
Simulation, filtered (threshold 3): coverage / agreement among covered / agreement over all records / no-history share / error	89.8% / 97.7% / 87.7% / 0.0% / 0.287

For the 22-class recommender, humidity is the most influential feature by contribution towards the predicted class (mean |SHAP| 0.205, first in 100% of bootstrap resamples), and permutation importance also ranks it first. The baseline's headline claim about humidity therefore holds up within this model. The order of the remaining features is not resolved. Potassium (0.178) and rainfall (0.172) have overlapping intervals, impurity importance puts rainfall (0.220) marginally above humidity (0.217), and temperature and pH contribute little under every method. Agreement between methods shows internal consistency, not agronomic correctness, because all methods explain the same model and, on augmented data, importances partly reflect how the classes were generated. For yield, crop identity dominates the temporal XGBoost model (mean |SHAP| 0.565 on the log scale, first in 100% of random subsets), followed by state (0.123), district (0.076) and crop year (0.058), with season

(0.028) and area (0.016) last by SHAP (Fig. 4). This differs from the baseline's conclusion that season dominates. Recomputing importances for the baseline configuration also puts crop first, with state ahead of season. Over five refits, crop, state and district keep ranks 1–3 by both SHAP and permutation importance. Crop year has a non-zero SHAP value but exactly zero permutation importance. A likely reason is that every test year (2011–2015) lies beyond the training range, so tree splits treat all test years identically and the model cannot represent the trend. The state effect is positive for the southern zone (+0.106 on the log scale), the North (+0.059) and the West (+0.044) and negative for the North-East (−0.060). It reflects regional productivity and crop mix, not any ordering of integer codes.

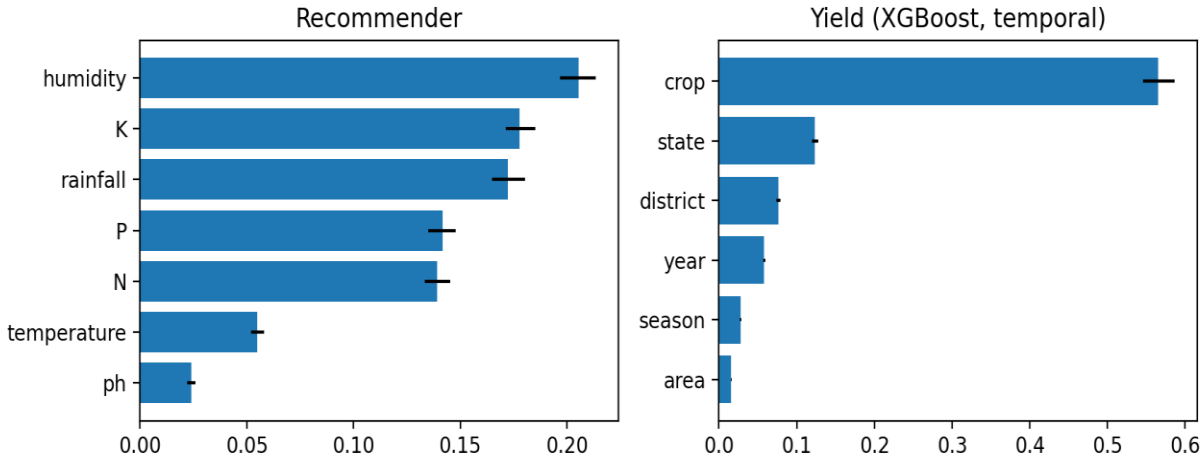


Fig. 4. Global SHAP importance with 95% bootstrap intervals. (a) Crop recommendation, predicted-class contributions. (b) Yield model

6. Discussion

Re-running the baseline framework shows which of its claims survive. The recommendation accuracy of about 99.5% reproduces, but it is saturated, the ranking among the top classifiers is unresolved, and it comes from data that appear to be augmented. The yield R^2 of 0.948 also reproduces under a random split, but it reflects interpolation inside known series. Under forecasting, Random Forest falls behind the historical mean, and only XGBoost with native categorical splits improves on the simple floors. The humidity claim holds within the recommender, whereas the claim that season dominates yield does not.

Four practical lessons follow. First, state the evaluation protocol next to every accuracy figure and include simple baselines. Without a historical mean and, for forecasting, last-year and trend baselines, a Random Forest R^2 of 0.95 looks like a success even though the model may be worse than a lookup table for next year's yields. Second, report group metrics and test them against a reference. A pooled yield error hid a 1.8-fold difference between zones, and a permutation null with crop-stratified contrasts shows how much of such a ratio is structure rather than chance or crop mix. Third, never present a lower disparity ratio on its own, because it can fall when the best-served group gets worse. The levelling-down index makes that difference explicit, and in our five-seed runs it attributes about four-fifths of the apparent gain from adaptive reweighting to this cause. Fourth, think about the likely cause of a disparity before choosing a mitigation. Our runs suggest that when groups differ in difficulty and not in size, difficulty-based reweighting is the more promising direction, while

inverse-frequency weighting may fail to help. That needs confirming across model families. These findings fit a broader view that fairness in agricultural AI concerns regions, production scales and data quality at least as much as individual attributes [9]. We use the word disparity for error heterogeneity across administrative groups. We do not claim that any community is treated unfairly, because no decision, harm or affected population is defined and the groups are proxies.

The study has several limitations. Zones are administrative regions, and scale groups use district-level area, not farm size, caste or income, so the audit shows regional and scale disparities in model error but does not measure harm to specific farming communities. The recommendation data are probably synthetic, so the accuracy, robustness, disparity and SHAP results describe the dataset and not real farms. Leave-regime-out testing also withholds whole crops that no closed-set classifier can predict. The "agreement" in the integration simulation measures consistency with historical practice, and the soil–location pairing is random. The yield stage has no weather or soil inputs and the data end in 2015, so the results should not be extrapolated to present-day yields, and part of the temporal degradation may come from unmodelled seasonal weather. Units are only partly resolved, and administrative boundaries changed during 1997–2015 (for example, Telangana was created in 2014), so series continuity is only approximate. The mitigation results rest on five seeds and one model family, the tuning groups coincide with the reporting groups, and the official agro-climatic zones were not audited. The multi-fold Random Forest has 60 trees and we ran one configuration per encoding, so a broader search could change the ranking of the Random Forest variants. The disparity ratio is descriptive and upward-biased for noisy groups, and record-level intervals ignore dependence across states and years, so state- and year-clustered intervals should be preferred. The re-evaluation targets one baseline paper and two public datasets, so generalising needs replication on other frameworks and on datasets with co-located soil, weather and production data. Finally, the feasibility filter locks in past practice: it cannot recommend a crop that is new to a district, and the data-poor North-East has the lowest coverage and agreement under it.

7. Conclusion

We reproduced a published crop recommendation and yield framework exactly, which let us ask what its numbers actually mean. Through the LEAF-Audit protocol, recommendation accuracy turns out to be high but saturated. It comes from data that appear to be augmented, and it does not carry over to crops outside the sampled climate range. The yield R^2 owes much to random splitting. In 96.8% of random-split test rows the series was already seen in training, and on the remaining rows Random Forest does not beat a historical-mean lookup.

Once the task becomes forecasting, Random Forest falls behind the historical mean, and the last-year and rolling-mean floors are hard to beat. Only XGBoost with native categorical splits improves on them (MAE 1.10 against 1.25 for the historical mean; intervals exclude zero under district and year clustering). The same booster with integer codes, and Random Forest with one-hot or target-encoded categories, do not. Detrended and tuned variants do not help, and every tree model struggles to transfer to unseen districts and states unless unseen levels are handled explicitly. The production file also contains unit, zero-value and aggregate-row problems that distort both training and evaluation.

The disparity audit adds another layer. Error differs by up to roughly 1.8-fold across zones, far beyond a permutation null and not explained by crop mix. Predictions for future years are low in every group, and large-area districts have higher normalised error although their production-weighted error is lower. The levelling-down index shows that about four-fifths of the apparent

benefit of adaptive reweighting comes from making the best-served cells worse, while inverse-frequency reweighting, with or without a cap, does not lower the ratios. Both findings come from five seeds and one model family. A feasibility filter built on local cropping history gives a minimal form of integration, at the cost of covering 93% of records and blocking genuinely new cultivation.

Future work should use co-located soil, weather and production data, the official agro-climatic zones, farm-level attributes that allow community-level analysis, and multi-seed mitigation tests across model families.

Author Contributions

Conceptualization, V.B.Y. and A.K.; methodology, V.B.Y.; software, V.B.Y.; validation, V.B.Y. and A.K.; formal analysis, V.B.Y.; investigation, V.B.Y.; resources, V.B.Y.; data curation, V.B.Y.; writing—original draft preparation, V.B.Y.; writing—review and editing, V.B.Y. and A.K.; visualization, V.B.Y.; supervision, A.K.; project administration, A.K. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

Both datasets are public (Kaggle) and are cited as references [41] and [42]. The corrections applied to the production file and the record flow are given in Table 1, and the state-to-zone and crop-name mappings are in Sections 4.3 and 5.4, so the corrected data can be regenerated from the public files. The experiment code, seeds, environment files and the corrected production file are available from the corresponding author (vipin884060@gmail.com) on genuine and reasonable request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Statement

The study relies on aggregated, public agricultural statistics and involves no human-subject data. Group labels are administrative regions and production-scale terciles, and no individual or community is identified.

Acknowledgement

This research was not funded by any grant.

References

[1] Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. *Heliyon*, *10*, e25112. https://doi.org/10.1016/j.heliyon.2024.e25112

[2] Alam, M. S. B., Esichaikul, V., Lameesa, A., Ahmed, S. F., & Gandomi, A. H. (2025). An approach for crop recommendation with uncertainty quantification based on machine learning for sustainable agricultural decision-making. *Results in Engineering*, *26*, 105505. https://doi.org/10.1016/j.rineng.2025.105505

[3] Shastri, S., Kumar, S., Mansotra, V., & Salgotra, R. (2025). Advancing crop recommendation system with supervised machine learning and explainable artificial intelligence. *Scientific Reports*, *15*, 25498. https://doi.org/10.1038/s41598-025-07003-8

[4] Ahmed, R. A., El-Shafai, W., Ahmed, Z. A., El-Rabaie, E. S. M., & Abd El-Samie, F. E. (2025). High-precision crop recommendation system with stacking ensemble classifiers for optimizing agricultural productivity. *Scientific Reports*, *15*, 43377. https://doi.org/10.1038/s41598-025-09640-5

[5] Sankareswari, More, R., Kazi, A., & Zagade, M. (2026). Smart crop selection and yield prediction through machine learning and explainable artificial intelligence. *International Journal of Artificial Intelligence and Machine Learning*, *6*(13s), 444–458.

[6] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), 15. https://doi.org/10.1145/2382577.2382579

[7] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

[8] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29* (pp. 3315–3323). https://arxiv.org/abs/1610.02413

[9] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115. https://doi.org/10.1145/3457607

[10] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

[11] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

[12] Zhai, Z., Martínez, J. F., Beltran, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, *170*, 105256. https://doi.org/10.1016/j.compag.2020.105256

[13] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks Track)*. https://arxiv.org/abs/2207.08815

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

[15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[16] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

[17] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

[18] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

[19] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PLOS ONE*, *11*(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

[20] Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, *10*, 1750. https://doi.org/10.3389/fpls.2019.01750

[21] Badshah, A., Alkazemi, B. Y., Din, F., Zamli, K. Z., & Haris, M. (2024). Crop classification and yield prediction using robust machine learning models for agricultural sustainability. *IEEE Access*, *12*. https://doi.org/10.1109/ACCESS.2024.3486653

[22] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*, 4540. https://doi.org/10.1038/s41467-020-18321-y

[23] Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

[24] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[25] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). https://arxiv.org/abs/1705.07874

[26] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

[27] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[28] Shams, M. Y., Gamel, S. A., & Talaat, F. M. (2024). Enhancing crop recommendation systems with explainable artificial intelligence: A study on agricultural decision-making. *Neural Computing and Applications*, *36*(11), 5695–5714. https://doi.org/10.1007/s00521-023-09391-2

[29] Turgut, O., Kok, I., & Ozdemir, S. (2024). AgroXAI: Explainable AI-driven crop recommendation system for Agriculture 4.0. In *Proceedings of the 2024 IEEE International Conference on Big Data*. https://doi.org/10.1109/BigData62323.2024.10825771

[30] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25. https://doi.org/10.1186/1471-2105-8-25

[31] Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 5491–5500). https://arxiv.org/abs/2002.11097

[32] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

[33] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). https://doi.org/10.1145/2090236.2090255

[34] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

[35] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (Article 43). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[36] Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 120–129). https://arxiv.org/abs/1905.12843

[37] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*. https://arxiv.org/abs/1911.08731

[38] Hashimoto, T. B., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 1929–1938). https://arxiv.org/abs/1806.08010

[39] Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. https://doi.org/10.1145/3465416.3483305

[40] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, et al. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

[41] Kaggle. (n.d.). *Crop recommendation dataset* [Data set]. Retrieved October 10, 2026, from https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[42] Kaggle. (n.d.). *Crop production in India* [Data set]. Retrieved October 10, 2026, from https://www.kaggle.com/datasets/abhinand05/crop-production-in-india

[43] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[44] Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, *37*(3/4), 256–266. https://doi.org/10.1093/biomet/37.3-4.256

[45] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

[46] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

[47] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

[48] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

[49] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

[50] Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, *3*, 1. https://doi.org/10.2466/11.IT.3.1

[51] Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414–427. https://doi.org/10.1162/rest.90.3.414

[52] Bouckaert, R. R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in knowledge discovery and data mining (PAKDD 2004)* (Lecture Notes in Computer Science, Vol. 3056, pp. 3–12). Springer. https://doi.org/10.1007/978-3-540-24775-3_3

[53] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

[54] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

[55] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

[56] Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177