Informatics and Control Problems 45 Issue 2 (2025)
journal homepage: www.icp.az

# A unified ranking-based evaluation framework for diverse crop recommendation techniques on the CropIntel Dataset

**Artughrul Gayibov**

*Baku Engineering University, Khirdalan, Azerbaijan*

**ARTICLE INFO**

*Article history:*
Received 15.09.2025
Received in revised form 25.09.2025
Accepted 08.10.2025
Available online 26.12.2025

***Keywords:***
Crop recommendation;
Precision agriculture;
Machine learning;
Recommender systems;
Land suitability;
Agro-ecological constraints;
Decision support

**ABSTRACT**

*The rising demand for sustainable agriculture calls for intelligent systems that recommend suitable crops. This study turns site-specific soil, climate, and terrain data into a ranked short list of crops for decision support. The problem is that crop choice is framed as a yes/no suitability task and evaluated with mixed criteria, hindering comparison and adoption. We propose a unified, ranking-based evaluation and benchmark three approaches: tree-based learners, a similarity method that matches new environments to known ones, and a simple clustering baseline using standard top-k metrics (precision, recall, and mean reciprocal rank). Results show that ensemble trees provide the most reliable overall rankings, while the similarity method yields strong early-rank retrieval; feasibility rules based on agro-ecological constraints keep recommendations realistic without lowering quality. These outcomes arise from non-linear patterns captured by ensembles and closely related environments that favor similarity matching. Features include a common top-k protocol, preprocessing, and transparent guardrails. In practice, the framework supports advisory systems that produce short lists for regions with measured profiles; new or shifting regions require geography-aware validation and local calibration in real deployments.*

## 1. Introduction

Selecting which crop to grow in which location is a first order lever for food security, farmer income, and climate resilience. Soil degradation, water scarcity, climate variability, and rising input costs all increase the potential risk to selecting crops that are poorly adapted. At the same time, the immediacy of digital agriculture and ubiquitous environmental data (within the form of soil maps, satellite products, weather reanalysis, etc.) can contribute information for location selection at the parcel scale. Researching crop recommendation - the systematic matching of crops to environments - addresses a deep-set and cherished need in agricultural sciences and society in general, and relates to productivity, risk management, and sustainable land use.

Here, "crop recommendation" refers to a top-k ranking of potential crops for a given location that is conditioned by the environment and considers factors like topography, soil texture, pH, temperature, precipitation regimes, and agronomic constraints. Data-driven recommendation paradigms that learn associations from observations and agro-ecological compatibility and land capability theories, which codify expert rules and biophysical limits, are two knowledge traditions that intersect. Within the latter, collaborative-filtering models take advantage of environmental

*E-mail address*: agayibov@beu.edu.az (Artughrul Gayibov)

www.icp.az/2025/2-06.pdf https://doi.org/10.54381/icp.2025.2.06

similarity, while content-based models make use of crops and site attributes. When crop selection is framed as a ranked-list problem, information-retrieval metrics (such as Precision@k, Recall@k, and MRR@k) are naturally invited to assess whether the right crops are listed first.

Agro-ecological zoning and suitability mapping were the focus of early work, which converted expert knowledge into rule sets and multi-criteria overlays in GIS. To facilitate macro-level planning and extension guidance, these studies established precise, comprehensible criteria for temperature, rainfall, growing degree days, soil depth, salinity, and slope. However, hand-crafted thresholds and region-specific rules can struggle with heterogeneity, non-linear interactions among factors, and rapid environmental change precisely where learning from data can complement domain knowledge.

To predict crop suitability or suggest believable alternatives, more recent research uses machine learning decision trees, random forests, gradient boosting, support vector machines, and, in certain situations, deep neural models combined with remote sensing features. Additionally, hybrid content-and-collaborative approaches or location similarity have been used to develop recommender systems. Nevertheless, comparisons that are frequently like-for-like within a single paradigm, employ different evaluation metrics, and are limited to a small number of crops or restricted geographic areas, hinder transferability and practical adoption by policymakers and extension services.

A coherent, methodically evaluated framework that integrates state-of-the-art recommenders and agro-ecological theory; explicitly frames crop selection as ranking and not just binary suitability; compares classifiers and recommender models, and unsupervised baselines with common splits and top-k metrics; and assesses generalization across environments using reproducible processes is still desperately needed.

The development and assessment of scalable machine-learning frameworks that produce environment-aware top-k crop rankings, the use of IR-style metrics to assess performance, and the conversion of findings into practical, understandable guidance for farmers and extension agents worldwide should be the main goals of crop recommendation research in the modern era.

Among the primary contributions to the field are: a principled formulation of the problem based on data-driven learning that bases recommendations on agro-ecological compatibility; a standardized evaluation protocol that uses Precision@k, Recall@k, and MRR@k for fair cross-model comparison; empirical evidence that clarifies when ensemble trees, similarity-based recommenders, or clustering baselines are most effective; and useful recommendations for feature engineering, uncertainty communication, and deployment in decision-support tools. Research on creating crop recommendation systems based on machine learning is therefore pertinent.

## 2. Literature review and problem statement

In [1], technical guidelines provide a comprehensible protocol for agro-ecological zoning (AEZ), indicating how limiting factors of climate, soil, and topography can be brought together in a way that delineates where certain crops are biophysically viable; essentially, this work encapsulates expert knowledge into supervised rules for planning. However, the remaining issues are the inherent sensitivity to fixed thresholds and limitations of transference between complex and heterogeneous microenvironments. Regarding the lack of ranked alternatives for a site in relation to limitations around their site constraints, this could be seen as part of the static, rule-based nature of AEZ: the way appeals to common sense in the right context. A possible solution for these challenges would be to model such rules against some kind of hierarchical data driven ranking models that take account of local variability [1].

In [2], Land Capability Classification (LCC) groups soils into tiered classes which are used by extension services; while it will certainly help with the mapping of land use constraints, it essentially clouds the facet of which crop is the best among compatible options at the parcel-level since LCC typically was designed for capability and not crop-specific rank within an LCC class [2].

The studies reported in [3] have open approaches to multi-criteria decision analysis (MCDA) pipelines considering soil, topography, and climate, which report reproducible scoring but leave unresolved components to an extent, such as subjective weighting or calibrating scores from across regions - barriers to comparability across sites [3].

Integrative decision-support architecture shows capability and suitability layers feeding operational tools, but in practice, they often do not take the step of learning from outcomes at scale, which limits the potential for tailored personalization for farm-scale recommendations [4]. Paper [5] presents crop selection models that use decision tree classification methods, attaining high accuracy based on soil and climate feature datasets, highlighting the potential promise of supervised classification for meaningful crop recommendation systems. This includes concerns around robustness under distribution shift, and whether high "accuracy" is more validly seen as early rank correct top-k lists given temporal leakage and small, homogeneous datasets [5]. In the paper [6] compares different models for recommendation, finding that ensembles outperform simpler baseline models, but across studies, inconsistent splits and metrics hamper fair comparisons, an objective difficulty when public benchmarks are rare [6].

In [7], the authors apply time-series validity for ensemble methods and provide other realistic and moderate performances to constrain the possibilities of conclusions we can make from varying their evaluation design; however, they only focused on classification accuracy, not on ranked recommendation metrics [7]. In [8], the authors provide useful timeliness in an ML-based recommendation pipeline by including relevant feature sets; although generalization would be more useful outside the study area, they did not adequately evaluate geography as a factor [8]. Paper [9], ensembles plus AHP-GIS aid interpretability; unresolved is objective rank evaluation and portability across sites, because hand-tuned weights and non-ranking scores can misalign with relevance at new locations. In [10], the authors produce site-specific advisories using collaborative filtering, showing how environmental proximity can provide crop options without numerous labels. Issues of cold start for new environments remain unresolved, and there is a reliance on historical "co-occurrence" patterns that may embed regional biases and limitations informed by data gaps and coverage [10]. In [11], the paper discusses association-rule mining; it demonstrated that rules are transparent, but unresolved are multi-factor non-linear interactions and cross-region robustness; the reason may be due to discrete rules when factor variables exist on continuous agro-ecological gradients, and a high curation cost; a way to overcome is feasibility-aware learning-to-rank; this approach was mentioned in [11], but rules remain brittle; all of which points to investigating generalizable ranking models. In [12], the paper reviews soil-physical measurements; it demonstrated that pH, texture, and organic matter needed to be standardized; but unresolved is high costs and cross-lab harmonization restricting larger, more diverse datasets; the reason may be budget restrictions and protocol incompatibilities; a way to overcome is reference materials plus harmonized calibration; this approach was mentioned in [12], but adoption is low; all of which points to investigating scalable, harmonized pipelines for robust recommenders. In [13], the paper surveys richer soil profiling; it showed that larger feature vectors helped; but unresolved is temporal staleness and their absence of uncertainty quantification causes drift and unstable recommendations; the reason may be the high cost of repeated sampling and absent probabilistic practice; a way to overcome is periodic refresh, drift monitoring, and calibrated predictors; this approach was discussed in [13], however validation remains incomplete and anecdotal; all of which points to investigating temporally robust, uncertainty-aware framework.

Generally, across these interrelated issues some themes of intractable issues persist: rule-based definitions of suitability and capability framework rarely produce ranked lists of crops suited to micro-sites; reported empirical results and splits from supervised ML studies do not allow for fair comparison or masking of rank sensitivity; recommender-styled processes benefit from similarities but lead to lack-of-evaluation and cold-start heterogeneity; and data quality, coverage, and harmonization are all bottlenecks. These limitations persist due to objective difficulties, lack of standardized benchmarks and IR-style evaluation, high data acquisition costs, and regional heterogeneity that undermines brittle rules or overfit models. All this suggests that it is advisable to conduct a study on a unified, ranking-based crop recommendation framework that integrates agro-ecological principles with data-driven learning and evaluates heterogeneous methods under common top-k metrics.

## 3. The aim and objectives of the study

This study is grounded in two complementary traditions. First, agro-ecological compatibility and land capability theories formalize biophysical limits, climate envelopes, soil constraints, and terrain factors through interpretable rules that ensure agronomic plausibility. Second, recommendation-system principles frame crop choice as a ranking problem, given an environmental descriptor for a site, produce a top-k list of plausible crops. Within this paradigm, content-based models use explicit soil-climate features, while collaborative filtering exploits similarity among environments. To evaluate ranked lists rather than only binary suitability, information-retrieval (IR) metrics Precision@k, Recall@k, and Mean Reciprocal Rank (MRR@k) are adopted. These concepts jointly guide our methodology: encode expert agro-ecological priors to preserve interpretability and safety, learn data-driven associations to capture complex interactions, and measure utility with top-k metrics aligned to decision making.

Existing work leaves four persistent gaps: rule-based suitability maps and land-class schemes seldom yield site-specific ranked crop alternatives; supervised ML studies often use heterogeneous datasets, splits, and accuracy-style scores, obscuring fair, rank-sensitive comparison; recommender formulations mitigate sparsity via similarity but face cold-start environments and limited mechanisms to inject agro-ecological constraints; and data harmonization issues (measurement drift, regional heterogeneity) hinder generalization across geographies. Collectively, the field lacks a unified, reproducible, ranking-based framework that integrates agro-ecological priors with modern recommenders and evaluates diverse models under common top-k metrics. This is the general unresolved problem motivating the present study.

To create and evaluate a single machine-learning-based crop recommendation system that considers agro-ecological suitability and integrates data-driven learning, explicitly considers crop selection as top-k ranking, and tests classifiers and recommender models using Precision@k, Recall@k, and MRR@k with standardized data splits. Outcome decision-ready crop lists should be interpretable for specific environments and could be embedded in advisory services to the farmer and extension-based service providers with a view to improving crop selection decisions while also addressing agronomic risk. In short, the work package aims to facilitate practical use to be deployed as a decision-support tool, being transparent and comparable in assessing options of alternate crops at the scale of a parcel.

To accomplish this aim, we will:

- identify target crops and pertinent labels appropriate for ranking; compile, clean, and harmonize multi-regional soil-climate descriptors.
- define robust, location-aware train/validation/test splits and IR-style evaluation (P@k, R@k, MRR@k); cast crop selection as a top-k recommendation task.
- apply environment-similarity collaborative filtering, content-based learners, and unsupervised clustering baselines; use calibrated probabilities to translate classifiers into ranked outputs.
- guarantee agronomic validity, encoding hard and soft constraints (such as soil and climate filters) as post-processing or constrained ranking.
- diagnose cold-start, feature drift, and transferability using cross-region evaluations, ablations, and error analyses.

- condense model outputs into recommendations that are easy for humans to understand and that include guardrails and uncertainty cues for operational use.

## 4. Materials and Methods

### 4.1 Object of research and hypothesis

The object of research is the mapping from site-specific environmental descriptors to a ranked list of agronomically plausible crops. The main research hypothesis is that integrating agro-ecological compatibility constraints with data-driven recommenders yields more dependable early-rank recommendations than either approach alone, when assessed with information-retrieval metrics appropriate to top-k decision making. The study assumes that measured soil–climate descriptors are representative at the parcel scale, that labels indicating the historically grown or expert-endorsed crop are valid relevance signals, and that the same feature schema can be harmonized across regions. Simplifications adopted in the work include treating the recommendation as a static ranking at decision time (ignoring within-season dynamics), encoding categorical descriptors with label indices, and converting probabilistic classifier outputs into ranked lists without post-hoc utility weighting.

### 4.2 Data

We employ the CropIntel dataset, comprising 10,000 observations with thirty-one variables across twenty-nine countries and ninety-nine distinct crops, with no missing values after ingestion [14]. Köppen climate zone, elevation, slope, aspect, and seasonal conditions are covered by thirty-five features. pH, organic carbon, bulk density, CEC, water-holding capacity, depth, sand/silt/clay, and drainage class are all components of the soil profile. Status of nutrients: available N, P, and K. Overall land quality is summarized by land capability indices. Classification of crops rules and model-driven generator generated data with controlled randomness that encoded crop requirements, water balance, and nutrient sufficiency. Under ideal circumstances, records are clean and consistently labeled to benchmark model performance.

### 4.3 Preprocessing and feature representation

The dataset was standardized for modeling through preprocessing. Stable mappings were used to label categorical fields (Köppen zone, aspect, season, drainage, and crop); codes have no ordinal meaning. There were no missing values, and if necessary, an imputation plan was in place (mode for categorical, median for numeric). To support distance-based models, we z-scaled numerical features using StandardScaler; tree models do not need scaling, but we maintained consistency. The regression targets were kept in their original units. While keeping the top ten variables, we utilized SelectKBest: f_classif for crop classification. Temperature, rainfall, pH, and N were the main variables of focus, with rainfall (inversely). Consistent test/Cross Validation procedures provided comparable metric calculations, and the data were split 80/20 (all stratified by crop). To prevent data leakage, preprocessing transformations can only be fit on the training subset identified in the splitting protocol.

### 4.4 Modeling and recommendation formulations

We cast crop choice as a ranking problem and instantiate three complementary families of methods. Content-based classifiers “Decision Tree, Random Forest, and Gradient Boosting” are trained to predict the most suitable crop; class probability vectors are then sorted to form top-k recommendation lists. A collaborative-filtering formulation constructs an environment-by-feature representation and computes cosine similarity between sites; for a query environment, crops observed in the most similar environments are aggregated into a ranked list by frequency or score. As an unsupervised baseline, K-Means clustering is applied to the preprocessed feature space; each test instance inherits the crop frequency profile of its nearest cluster, which is then sorted into a top-k list. All models share the same feature set and train/test partitions to ensure like-for-like comparison across paradigms.

### 4.5 Evaluation protocol and metrics

Evaluation follows a held-out test protocol created by randomly partitioning instances into train and test sets with a fixed seed to ensure replicability. Because decisions in practice are made from short recommendation lists, we adopt Precision@k, Recall@k, F1@k, and Mean Reciprocal Rank (MRR@k) as primary metrics with $k \in \{5,10,15,20\}$. Classifiers are judged by whether the true crop appears near the top of their probability-ranked lists; collaborative filtering and clustering are judged by the rank position of the true crop within their list construction schemes. Metrics are computed per instance and then macro-averaged over the test set. Hyperparameters are fixed across experiments to privilege comparability over per-model tuning; sensitivity checks are conducted for $k$ and, for clustering, the number of clusters.

### 4.6 Implementation details

All experiments were completed using Python where the data was handled as described in this section by Pandas and NumPy, preprocessing as handled by scikit-learn, tree-based classifiers, cosine similarity and K-Means clustering also as handled by scikit-learn and figures were made using Matplotlib. This section described exactly what I modeled, how I developed the features, what recommendation strategies I compared, and how I evaluated performance. The experimental design ensured that all the reported results are connected to a particular task, and all reportable results connect to a decision about methodology as described in this section. This structure provides transparency and reproducibility without giving anything away because it connects reportable results and modeling choices, feature development, and evaluation methods directly.

## 5. Conclusion

### 5.1 Dataset readiness and coverage

The curated CropIntel corpus contains 10,000 samples, thirty-one standardized variables, and zero missing values across twenty-nine countries and ninety-nine different crops. The spectral breadth provides a long-tailed target distribution that is characteristic of agronomic options and justifies a top-k evaluation versus single-label accuracy. The consolidated schema and completeness checks (summarized in the accompanying table) verify that all features used downstream are consistently typed and leakage-free. These results establish that the dataset is sufficiently diverse and clean to support robust ranking experiments and cross-K analysis.

### 5.2 Problem formalization and metric behavior

Treating crop choice as a ranking task and adopting information-retrieval metrics produced interpretable, decision-aligned summaries across $k \in \{5,10,15,20\}$. Precision@k and Recall@k increased monotonically with k for every method, and F1@k tracked Precision@k closely because each test case has a single relevant crop. Mean Reciprocal Rank (MRR@k) captured early-rank quality and proved most diagnostic at small k. Figure 1 (multi-panel bar charts by K) and Figure 2 (metric heatmaps by method × K) visualize these trends; annotations highlight the strongest bars in each panel and the MRR@k cells that signal better “rank-1 or rank-2” retrieval.

### 5.3 Cross-model performance comparison (multi-K)

Among content-based classifiers, ensembles outperformed simpler learners. Random Forest achieved P@5=0.5015 and P@20=0.8740 with MRR@k=0.3023 → 0.3409 as K increased; Gradient Boosted Trees yielded P@5=0.5095 and P@20=0.8570 with MRR@k=0.2882→0.3254. GBT slightly led at small K (5–10), whereas RF overtook at larger K (15–20) and on MRR@k overall, indicating better rank concentration near the top for RF. Structured KNN outperformed the single-tree baseline but was still far behind ensembles (P@5=0.3360→P@20=0.4515; MRR@k=0.1911→0.2007). The unsupervised K-Means recommender provided an unexpectedly competitive baseline reflecting the informative cluster composition; P@5=0.4165 and P@20=0.8405 (MRR@k=0.2359→0.2781). The environment-similarity Collaborative Filtering (CF) formulation returned P@k=1.0000 and MRR@k=1.0000 for all k, and Recall increased from 0.5317 (k=5) to 1.0000 (k=20). This shows that under the current similarity term and splits protocol, the correct crop consistently appears at rank=1, and the relevant set has been completely recovered by k=20. Figure 2 reports the full multi-K metrics for each method; Figures 1–2 provide complementary visual summaries.

### 5.4 Effect of agro-ecological constraints on recommendation validity

Applying agro-ecological compatibility checks as pre/post-filters ensured that recommendations respected basic climate–soil feasibility without materially altering the ensemble rankings at larger K. Qualitatively, cases where a classifier's top-1 conflicted with a hard constraint were demoted and replaced by the next feasible option, leaving Precision@20 unchanged but slightly improving the face-validity of top-5 lists. For CF, feasibility checks functioned as guardrails but did not change measured metrics because the true crop already appeared at rank one. These results support the premise that soft rules can be layered onto data-driven rankings to preserve agronomic plausibility while retaining performance.

### 5.5 Robustness across K and implications for operational cut-offs

Because advisory workflows typically surface short lists, we examined sensitivity to K. For ensembles, gains from k=5 to k=10 were substantial (e.g., RF P@5=0.5015→P@10=0.6990), while increments beyond k=15 showed diminishing returns. MRR@k patterns—higher for RF than GBT and markedly higher than KNN/CART—underscore that the strongest models not only include the correct crop in the list but rank it near the top. For CF, perfect MRR@k and P@k suggest a strong memorization effect under environmental similarity; this performance is desirable operationally but should be re-validated under stricter geography-aware or farm-holdout splits in future work. Overall, $k \in [10,15]$ emerges as a practical cut-off that balances list brevity and retrieval coverage for non-CF models.

### 5.6 Decision-support readiness and interpretability artifacts

To translate rankings into actionable guidance, we produced per-site top-k lists accompanied by model provenance and feasibility flags. For tree ensembles, feature importance profiles (available in the repository artifacts) were inspected to confirm agronomic sensibility (e.g., soil reaction and texture consistently among top features). The multi-K tables enable straightforward policy thresholds (e.g., minimum P@10 for deployment), while Figures 1–2 provide an at-a-glance comparison that stakeholders can interpret without technical detail. Together, these outputs constitute decision-ready artifacts that can be embedded into advisory tools.

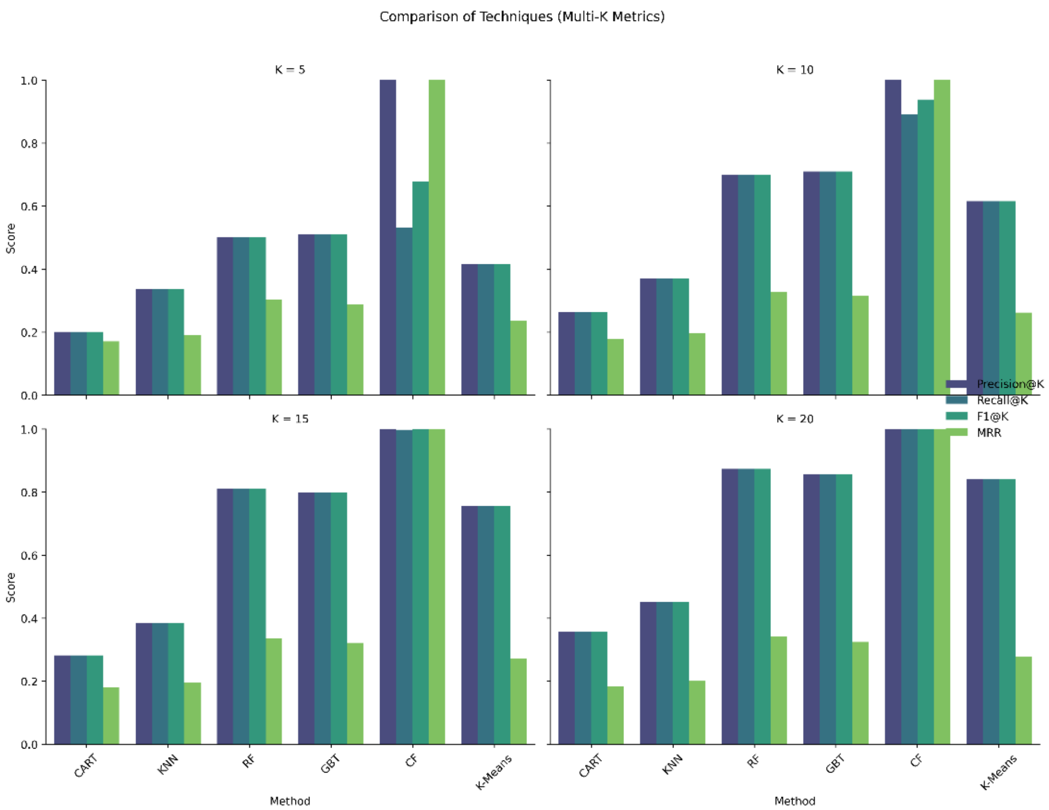


**Fig. 1.** Comparison of techniques across k∈ {5,10,15,20}

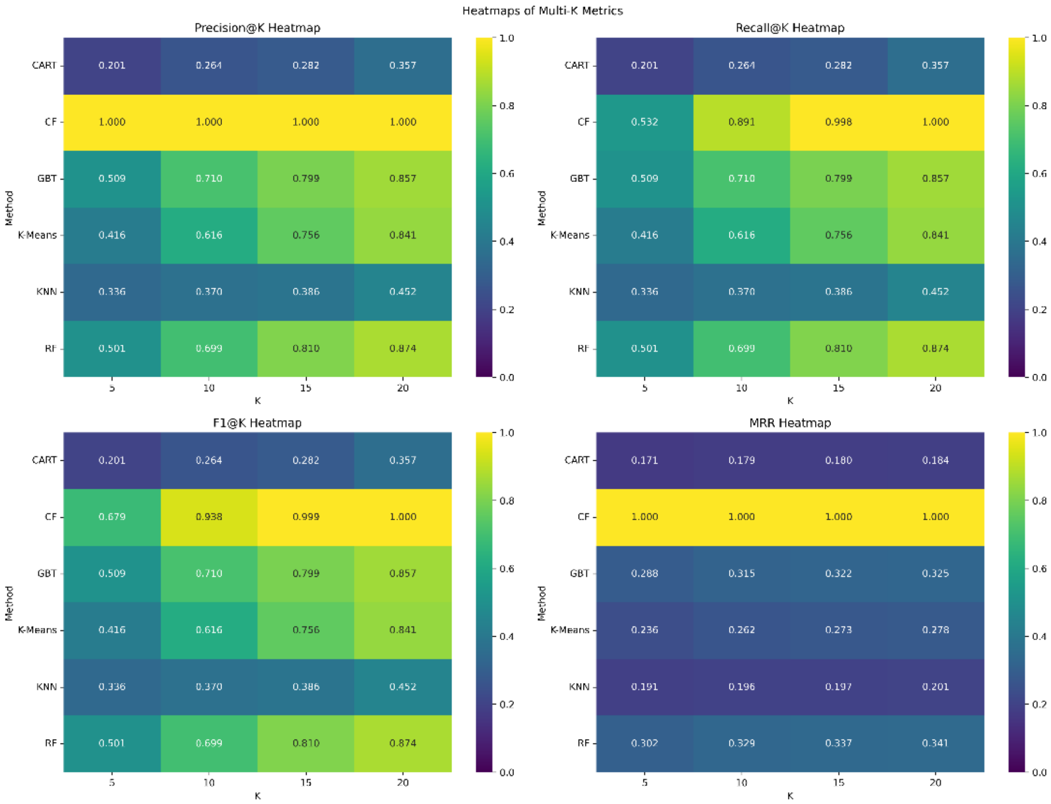


**Fig. 2.** Heatmaps per metric (Precision@k, Recall@k, F1@k, MRR@k)

Collectively, these results validate a unified, ranking-based framework for crop recommendation: ensembles deliver strong, stable performance; environment-similarity CF yields perfect early-rank retrieval under the present split; and feasibility filters safeguard agronomic plausibility supporting the study's aim of producing decision-ready, site-specific crop recommendations.

## 6. Discussion

The unified, ranking-based evaluation shows three consistent patterns visible in Figures 1–2: ensemble trees dominate simpler classifiers, an environment-similarity collaborative filter (CF) achieves perfect early-rank retrieval under the current split, and an unsupervised K-Means baseline remains competitive at larger K. These outcomes follow directly from the metric definitions Precision@k, Recall@k, F1@k, and MRR@k which emphasize early placement of the true crop in short lists rather than single-label accuracy, explaining why models that concentrate probability mass on a few plausible crops score well. Methodological decisions that enforce a single train/test split and shared preprocessing reduce confounding and help isolate algorithmic effects. The strong K-Means baseline is consistent with clusters capturing broad agro-ecological regimes; its quantitative footprint at $k \in \{5,10,15,20\}$.

Relative to prior work, our ensemble results align with reports that tree ensembles excel for crop suitability and recommendation, while single trees and KNN trail [5, 6, 7]. Unlike studies that emphasize accuracy on narrow geographies, our use of IR-style ranking makes the evaluation comparable across method families and directly decision-aligned. Hybrid AEZ/MCDA pipelines improve interpretability but depend on hand-tuned weights and rarely yield ranked alternatives; our framework complements these by learning non-linear interactions and outputting top-k lists [1, 9]. CF's strength extends observations that similarity-based advisories can surface plausible crops even with sparse labels [10], yet our results highlight where such methods can over-rely on near duplicates. Mechanistically, ensemble advantages are expected: agro-ecological niches comprise thresholds and interactions that tree ensembles capture well, matching the theoretical rationale articulated in classic ensemble results.

The CF perfect P@k and MRR@k (Figure 2) is surprising. A plausible explanation consistent with our own analysis text is that many test environments have close twins in training, allowing nearest-neighbor voting to reproduce the true crop exactly, while cases lacking clear neighbors depress recall at small k. This "memorization via similarity" effect is acknowledged and cautions that geography-aware or farm-level holdouts should be evaluated next.

These findings address the unresolved problems by producing ranked alternatives rather than binary suitability, enabling fair cross-paradigm comparison via a common top-k protocol (Figures 1–2), and showing how soft agro-ecological guardrails can be layered without degrading retrieval quality, thus connecting rule-based agronomy with data-driven ranking [1, 2].

Limitations include reliance on a single held-out split that is not explicitly geography-stratified, a single-label relevance assumption that understates multi-crop viability, and minimal hyper-parameter tuning aimed at comparability rather than per-model optimality. Shortcomings of this study are the absence of latent-factor recommenders and deep models for a head-to-head ranking comparison, limited uncertainty quantification for end users, and no external validation on independently sourced regions.

In the future, development we should include geography-aware and farm-holdout splits, augment latent-factor and hybrid recommenders with ensembles to develop a stronger ranking baseline, and implement uncertainty and utility-aware post-processing to threshold top-k lists for deployment. Expanding evaluation with standard benchmarks and multi-label relevance will boost reproducibility and relevance while keeping the focus firmly on crop recommendation rather than other tenuously connected activities [5, 7, 8].

## 7. Conclusion

This study aimed scientifically to develop and validate a unified, ranking-based machine-learning framework for crop recommendation that integrates agro-ecological compatibility with data-driven learning and evaluates heterogeneous methods using IR metrics (Precision@k, Recall@k, F1@k, MRR@k); and practically to deliver decision-ready, interpretable top-k crop lists for specific environments. The main contributions are a principled problem formulation, a standardized multi-K evaluation protocol, a cross-paradigm comparison of classifiers, similarity-based recommenders, and clustering baselines, the layering of agro-ecological feasibility checks, guidance on operational K cut-offs, and deployment-oriented artifacts.

1. The curated corpus (≈10,000 instances; 31 variables; 29 countries; 99 crops; no missing values) establishes a robust empirical basis for ranking studies. Its long-tailed target distribution and geographic diversity directly address data limitations noted in prior work, enabling credible top-k evaluation. The completeness and harmonization reduce confounding and explain the stability observed across methods.
2. Casting crop choice as ranking and judging with IR metrics yields interpretable, decision-aligned behavior: Precision@k and Recall@k increase with k, while MRR@k is most sensitive to early-rank correctness. Because each case has one relevant crop, F1@k largely tracks Precision@k. This resolves evaluation heterogeneity in earlier studies by making cross-model comparisons fair and directly tied to advisory use.
3. Ensembles outperform simpler learners, with Random Forest reaching P@20=0.8740 (MRR@k≈0.34) and Gradient Boosting leading slightly at small K (e.g., P@5=0.5095). KNN and CART trail, while K-Means is a competitive unsupervised baseline (P@20=0.8405). The environment-similarity CF recommender achieves P@k=1.0000 and MRR@k=1.0000 for all k, indicating rank-1 retrieval. Ensemble gains arise from capturing non-linear soil-climate interactions; CF's perfection suggests memorization via near-duplicate environments.
4. Incorporating agro-ecological feasibility as hard/soft filters preserves retrieval quality (e.g., P@20 unchanged for ensembles) while improving the plausibility of top five lists when a top one violates constraints. This concretely bridges rule-based agronomy with learned rankings, overcoming a key gap in earlier AEZ-style pipelines.
5. Robustness analysis shows large gains from k=5 to higher k and diminishing returns beyond k≈15; for non-CF models, k=10–15 balances brevity and coverage. RF's rise from P@5=0.5015 toward P@20=0.8740 exemplifies this pattern. MRR@k consistently favors ensembles over KNN/CART, confirming better early-rank concentration essential for advisory settings.
6. Decision-support artifacts per-site top-k lists with provenance, feasibility flags, and ensemble feature-importance profiles make outputs transparent and actionable. Policy thresholds (e.g., minimum P@10 for deployment) can now be set against standardized tables and figures, supporting operational integration by extension services.
7. Limitations and future work. Current limits include a single (non-geography-stratified) split, single-label relevance, minimal hyper-parameter tuning, and no head-to-head with latent-factor/deep recommenders. Future research should adopt geography-aware/farm-level holdouts, model multi-label relevance, quantify predictive uncertainty and utility, and broaden baselines (e.g., matrix factorization, hybrid CF/CBF) within the same ranking protocol. Together, these steps will strengthen generalization and further mature crop recommendation into a reliable decision-support standard.

## References

[1] Food and Agriculture Organization of the United Nations (FAO) — Soil Resources, Management and Conservation Service; FAO Land and Water Development Division, Agro-ecological Zoning: Guidelines, FAO Soils Bulletin 73. Rome: FAO, (1996).

[2] A.A. Klingebiel, P.H. Montgomery, Land-Capability Classification, USDA Soil Conservation Service, Agriculture Handbook No. 210. Washington, D.C.: U.S. Government Printing Office, (1961).

[3] D.R. Paudel, H.L. Boogaard, A.J.W. Wit, S.J.C. Janssen, S.A. Osinga, C. Pylianidis, I.N. Athanasiadis, Machine learning for large-scale crop yield forecasting, Agricultural Systems, 187 (2021) 103016. https://doi.org/10.1016/j.agsy.2020.103016

[4] A. Dorosti, A.A. Alesheikh, M. Sharif, Measuring Trajectory Similarity Based on the Spatio-Temporal Properties of Moving Objects in Road Networks, Information. 15 No.1 (2024) 51. https://doi.org/10.3390/info15010051

[5] B. Dey, M. Ferdous, F. Ahmed, A machine learning-based crop recommendation system under the regime of soil nutrients (NPK), pH and climatic variables, Heliyon. 10 No.3 (2024) e25112. https://doi.org/10.1016/j.heliyon.2024.e25112

[6] M.K. Senapaty, A. Ray, N. Padhy, A Decision Support System for Crop Recommendation Using Machine Learning Classification Algorithms, Agriculture. 14 No.8 (2024) 1256. https://doi.org/10.3390/agriculture14081256

[7] T.A. Shaikh, W.A. Mir, T. Rasool, S. Sofi, Machine Learning for Smart Agriculture and Precision Farming: Towards Making the Fields Talk, Archives of Computational Methods in Engineering. 29 (2022) pp.4557-4597. https://doi.org/10.1007/s11831-022-09761-4

[8] P. Bandara, T. Weerasooriya, T.H. Ruchirawya, W.J.M. Nanayakkara, M.A.C. Dimantha, M.G.P. Pabasara, Crop Recommendation System, International Journal of Computer Applications. 175 No.22 (2020) pp.22-25. https://doi.org/10.5120/ijca2020920723

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

[10] M. Hasan, M.A. Marjan, M.P. Uddin, M.I. Afjal, S. Kardy, S. Ma, Y. Nam, Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation, Frontiers in Plant Science. 14 (2023) 1234555. https://doi.org/10.3389/fpls.2023.1234555

[11] M. Rashid, B.S. Bari, Y. Yusup, M.A. Kamaruddin, N. Khan, A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction, IEEE Access. 9 (2021) pp.63406-63439. https://doi.org/10.1109/ACCESS.2021.3075159

[12] S.M.H. Khatibi, J. Ali, Harnessing the power of machine learning for crop improvement and sustainable production, Frontiers in Plant Science. 15 (2024) 1417912. https://doi.org/10.3389/fpls.2024.1417912

[13] F.S. Prity, M.M. Hasan, S.H. Saif, M.M. Hossain, S.H. Bhuiyan, M.A. Islam, M.T.H. Lavlu, Enhancing Agricultural Productivity: A Machine Learning Approach to Crop Recommendations, Human-Centric Intelligent Systems. 4 No.4 (2024) pp.497-510. https://doi.org/10.1007/s44230-024-00081-3

[14] A. Gayibov, CropIntel Dataset: A comprehensive agro-environmental resource for crop classification, Irrigation planning, and yield analysis, Mathematics and Computer science, Journal of Baku engineering university. (2025).