UOT: 004.67:004.8:631.6:633
DOI: https://doi.org/10.30546/09090.2025.210.005

CROPINTEL DATASET: A COMPREHENSIVE AGRO-ENVIRONMENTAL RESOURCE FOR CROP CLASSIFICATION, IRRIGATION PLANNING, AND YIELD ANALYSIS

Artughrul GAYIBOV
agayibov@beu.edu.az
https://orcid.org/0009-0009-7349-0286
Baku Engineering University
Baku, Azerbaijan

ARTICLE INFO

Article history:
Received: 2025-06-17
Received in revised form:2025-06-17
Accepted:2025-07-02
Available online

Keywords:
Precision Agriculture;
Crop Recommendation;
Machine Learning;
Irrigation Prediction;
Yield Forecasting;
Data-Driven Agriculture;
Agro-Environmental Modeling

2010 Mathematics Subject Classifications: 62H30 → 62J02; 68T05; 68T10; 62P12.

ABSTRACT

The growing demand for sustainable agricultural intensification necessitates advanced tools for optimizing crop selection and resource management. This study presents an integrated analytical framework using machine learning to derive agro-environmental intelligence from the comprehensive CropIntel dataset. We evaluated a suite of over fifteen machine learning algorithms for three critical precision agriculture tasks: crop classification, seasonal irrigation requirement prediction, and yield potential forecasting. The results demonstrate the exceptional capability of tree-based models in this domain. Notably, Decision Tree, Bagging, and Gaussian Naïve Bayes classifiers achieved near-perfect accuracy (Acc≈99.7%) in identifying suitable crops based on climatic and edaphic conditions. For regression tasks, LightGBM and Histogram Gradient Boosting models proved most effective, explaining approximately 84% of the variance in irrigation needs (R^2≈0.84) and about 70% of the variance in yield potential (R^2≈0.696). These findings underscore the potential of machine learning to create powerful decision support systems that can guide crop selection, optimize water allocation, and provide reliable yield forecasts. This research contributes to a holistic, data-driven methodology that can enhance the efficiency and sustainability of agricultural practices.

1. INTRODUCTION

With a growing population and increasing environmental pressures like climate change, water scarcity, and land degradation, the world's agricultural sector is at a turning point in its history and must boost productivity to meet these demands (FAO, 2021). According to Chlingaryan, Sukkarieh, and Whelan (2018), precision agriculture has become a paradigm shift that moves away from uniform field management and towards a data-intensive approach that optimizes and manages agricultural production at a granular level. Growing the "right crop for the right land," which optimizes yield potential while reducing resource inputs and environmental impact, is a fundamental component of this paradigm (Rani, Mishra, Kataria, Mallik, & Qin, 2023).

Machine learning (ML) has emerged as a game-changing tool in agriculture thanks to the growth of data from sources like satellites, weather stations, and soil sensors as well as improvements in processing power. Numerous agricultural problems, such as yield forecasting, weed control, crop disease detection, and intelligent irrigation, have been effectively addressed by researchers using machine learning techniques (Navarro-Hellín et al., 2016; van Klompenburg, Kassahun, & Catal, 2020). These uses show how machine learning algorithms can interpret intricate, non-linear relationships in large agro-environmental datasets that are frequently too big for conventional statistical techniques.

Despite these developments, a large portion of current research concentrates on resolving discrete issues. Many studies, for example, create models especially for crop recommendation, while others only concentrate on yield prediction or irrigation optimization. A major research gap is represented by this fragmentation: there aren't enough integrated frameworks that can offer farmers and agricultural planners comprehensive, end-to-end decision support. To enable a thorough cost-benefit analysis prior to the sowing of a single seed, a truly intelligent system should not only recommend which crop to plant but also estimate its resource requirements and potential productivity simultaneously.

Using the new CropIntel dataset, this study attempts to close this gap by creating and assessing a thorough, machine learning-based analytical framework. The three main goals of our research are to: (1) accurately identify the crop that is best suited for a particular set of agro-climatic conditions; (2) forecast the crop's seasonal irrigation water requirements; and (3) estimate the crop's potential yield under those environmental conditions. This study adds to a more comprehensive approach to data-driven agricultural intelligence by addressing these related tasks using a single methodology. The results are meant to offer a strong basis for creating decision-support tools that improve farming operations' profitability and sustainability.

2. LITERATURE REVIEW

The use of machine learning in agriculture has expanded significantly, as evidenced by the abundance of research showing its applicability in a variety of fields. Key findings in the three main areas that are pertinent to our study—crop recommendation, irrigation management, and yield prediction—are summarized in this review.

2.1. Crop Recommendation

Based on soil properties and climate, crop recommendation systems seek to pair crops with the best available land. Rule-based logic was frequently used in early systems, but more recently, ML classifiers have been used to produce predictions that are more accurate and dynamic. For instance, Pudumalar et al. (2017) suggested crops using algorithms such as K-Nearest Neighbors (KNN) and Naïve Bayes. More sophisticated research, like that conducted by Rani et al. (2023), used a Random Forest classifier on a mix of soil and weather data and achieved an accuracy of more than 97%. These studies demonstrate that crop suitability can be accurately predicted by environmental characteristics. They might not always incorporate projections of future resource requirements, though, and frequently function in regional contexts.

2.2. Irrigation Management

In contemporary agriculture, water efficiency is crucial. ML has played a key role in creating "smart" irrigation systems that maximize water use. Navarro-Hellín et al. (2016) created a

decision support system that automated and managed irrigation scheduling using sensor data and predictive models, showing a great deal of promise for water savings. Like this, scientists have forecasted crop water requirements based on meteorological data using methods ranging from adaptive neuro-fuzzy inference systems to support vector regression (SVR) (Goap et al., 2018). Although useful, these studies frequently treat irrigation as a stand-alone issue, unrelated to the initial crop selection procedure or the implications for final yield.

2.3. Crop Yield Prediction

Because of its significant economic ramifications, crop yield prediction is one of the most researched applications of machine learning in agriculture. Numerous algorithms, such as Random Forest, SVR, and deep neural networks, have been used to forecast yields for different crops, according to a systematic review by van Klompenburg et al. (2020). Chlingaryan et al. (2018) pointed out that by more accurately representing the intricate interactions between variables influencing growth, data-driven models usually perform better than conventional process-based or statistical models. Paudel et al. (2022) demonstrated the scalability of machine learning by successfully implementing it for regional crop yield forecasting in Europe. However, yield prediction remains a complex challenge, as performance is highly dependent on the quality and diversity of input data, and models often do not concurrently predict the resource inputs required to achieve that yield.

The literature currently in publication attests to machine learning's effectiveness for agricultural tasks. However, it also highlights a recurring weakness in the integration of these tasks. Using a single, unified dataset, our study contributes to the field by explicitly connecting crop selection (classification) with predictions of its primary resource need (water; regression) and its output (yield; regression).

2. THEORETICAL FRAMEWORK

Agroecology and land evaluation science, which hold that agricultural productivity is a direct result of the interplay between a crop's genetic potential and its surroundings, serve as the foundation for this study. We use machine learning as a tool to model these intricate, well-established relationships empirically.

3.1. Agro-Climatic Suitability

Agro-climatic suitability is the fundamental idea driving our crop classification task. For optimum growth, each crop species has a different set of environmental requirements, such as ranges for soil pH, moisture, temperature, and sunlight (FAO, 1976). The ecological niche of a crop is defined by these conditions. Our framework assumes that an ML model can learn the decision boundaries that define the niche for each crop if it is given a rich set of features that describe these environmental parameters. As a result, the classification task is an empirical approximation of the agroecological principle of matching a crop to its ideal environment rather than just a pattern recognition exercise.

3.2. Land Capability and Productivity

The idea of Land Capability Classification (LCC) serves as the theoretical foundation for the regression tasks of yield and irrigation prediction. According to its innate ability to support agricultural production without degrading, the LCC is a methodical framework for evaluating land (Helms, 1992). Classes I through VIII are used to grade land, with lower classes having

greater productivity potential and fewer restrictions. LCC ratings for both irrigated and non-irrigated conditions are included in the CropIntel dataset. This gives our models a solid theoretical foundation. We predict that while irrigation requirement is a function of the difference between a crop's water requirements and the environment's capacity to supply them (a function of rainfall, soil water holding capacity, etc.), yield potential is highly connected with land capability. Our regression models are thus designed to learn these functional relationships, predicting how land quality and climate translate into specific resource needs and production outcomes.

3. METHODOLOGY

This study employs a quantitative, data-driven approach to build and evaluate predictive models for crop selection, irrigation needs, and yield potential.

4.1. The CropIntel Dataset

The empirical foundation of this research is the CropIntel dataset, a large and comprehensive collection of agro-environmental data. It contains 10,000 unique records, each representing a specific crop scenario defined by a combination of geographic, climatic, topographic, and edaphic (soil) factors. The dataset's strength lies in its breadth, covering 29 countries across North America, Europe, and Asia, encompassing a wide range of Köppen climate zones (e.g., Cfa, Cfb, Dfb, Bsk, Csa) (Arnfield, 2019). This diversity ensures that the models are trained in a wide spectrum of agricultural conditions.

The dataset includes over 35 features for each record, which can be categorized as follows:

- **Locational and Climatic Features:** Country, Subregion, Köppen Climate Zone, Season, Seasonal Average Temperature (°C), Seasonal Total Rainfall (mm), Seasonal Average Humidity (%), and Seasonal Average Daily Sunlight (hours).
- **Topographic Features:** Elevation (m), Slope (degrees), and Aspect (e.g., North-facing, South-facing).
- **Soil Properties:** A detailed soil profile including Soil pH, Soil Organic Carbon (%), texture (Sand, Silt, and Clay %), Bulk Density (g/cm³), Cation Exchange Capacity (meq/100g), Water Holding Capacity (mm/m), Soil Depth (cm), and Drainage Class.
- **Nutrient Status:** Levels of primary macronutrients, including Nitrogen (N), Phosphorus (P), and Potassium (K), measured in kg/ha.
- **Land Capability Indices:** Land Capability Class for both non-irrigated and irrigated scenarios, providing a synthesized measure of land quality.
- **Target Variables:**

1. **Crop:** The specific crop type (over 120 types, such as 'Maize', 'Wheat', 'Apples'), which serves as the target for our classification task.
2. **Seasonal_Avg_Irrigation_Needed_mm:** The estimated volume of irrigation water (in mm) required during the season to supplement rainfall for optimal growth. This is a target for the first regression task.
3. **Yield_Potential_Score:** A normalized index (0-100) that represents the potential productivity of the crop in the given environment. This is the target for the second regression task.

4.2. Data Preprocessing and Feature Selection

The data was preprocessed in several ways prior to model training. First, label encoding was used to transform categorical features (such as Country, Subregion, and Crop) into numerical representations. Second, the median and mode were used to impute numerical and categorical columns, respectively, to account for any missing values. For algorithms that are sensitive to feature magnitudes (such as SVM and MLP) to function at their best, all numerical features were finally standardized using StandardScaler (scikit-learn, StandardScaler, 2019) to have a mean of zero and a standard deviation of one.

We used SelectKBest (sklearn, SelectKBest, 2021) for a feature selection step to simplify the model and concentrate on the most important variables. We determined the ten features with the greatest variation across crop classes for the classification task using the F-statistic from ANOVA (f_classif). We chose the ten features that were most correlated with the corresponding target variables (yield and irrigation) for the regression tasks using the F-statistic from univariate linear regression (f_regression).

4.3. Machine Learning Models and Evaluation

We evaluated a comprehensive suite of over 15 ML models for both classification and regression to ensure a thorough comparison of different algorithmic approaches.

- **Classification Models:** The list included Decision Tree, Support Vector Machine (SVC), K-Nearest Neighbors, Random Forest, Extra Trees, AdaBoost, Histogram Gradient Boosting, Bagging, Logistic Regression, Ridge Classifier, SGD Classifier, Gaussian Naïve Bayes, and a Multi-Layer Perceptron (MLP) Classifier, as well as XGBoost and LightGBM classifiers.
- **Regression Models:** A parallel set of regressors was used, including Decision Tree, SVR, K-Nearest Neighbors, Random Forest, Extra Trees, AdaBoost, Histogram Gradient Boosting, Bagging, Ridge, Lasso, ElasticNet, SGD Regressor, Huber Regressor, MLP Regressor, XGBoost, and LightGBM.

A rigorous three-fold cross-validation process was used to evaluate the model's performance. The dataset was divided into three folds, two of which were used for training and one of which was used as a validation set. After that, the three folds' performance metrics were averaged.

- **Evaluation Metrics:**
  - For **classification**, we measured **Accuracy**, and weighted **Precision**, **Recall**, and **F1-Score** to account for the multi-class nature of the problem.
  - For **regression**, we measured the **Coefficient of Determination (**R^2**)**, **Mean Absolute Error (MAE)**, and **Mean Squared Error (MSE)**.

4. RESULTS

Below are the empirical findings from our three analytical tasks: yield prediction, irrigation prediction, and crop classification. Tables and figures provide a summary and visual representation of the performance of the different models.

5.1. Crop Classification Performance

The crop classification task yielded exceptionally clear results, with a distinct separation between high-performing and low-performing models. As shown in Table 1 and Figure 1, several algorithms were able to predict the correct crop type with near-perfect accuracy.

Table 1. Performance Metrics for Crop Classification Models

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree Classifier	0.997	0.997	0.997	0.997
Bagging Classifier	0.997	0.997	0.997	0.997
Gaussian Naive Bayes	0.997	0.994	0.997	0.995
Extra Trees Classifier	0.956	0.954	0.956	0.950
Random Forest Classifier	0.896	0.890	0.896	0.886
MLP Classifier	0.803	0.791	0.803	0.788
Logistic Regression	0.524	0.474	0.524	0.467
K-Nearest Neighbors	0.362	0.334	0.362	0.341
LightGBM Classifier	0.224	0.238	0.224	0.195

The most remarkable outcome is the Acc≈99.7% accuracy attained by Gaussian Naïve Bayes classifiers, Bagging (which starts with decision trees), and the standard Decision Tree classifier. This suggests that the CropIntel dataset's environmental and soil characteristics include distinct, distinct signals that specify each crop's suitability. With accuracies of 95.6% and 89.6%, respectively, ensemble techniques like Random Forest and Extra Trees also demonstrated strong performance. The non-linear character of the decision boundaries was highlighted by the difficulties faced by distance-based models (KNN) and linear models (such as logistic regression).

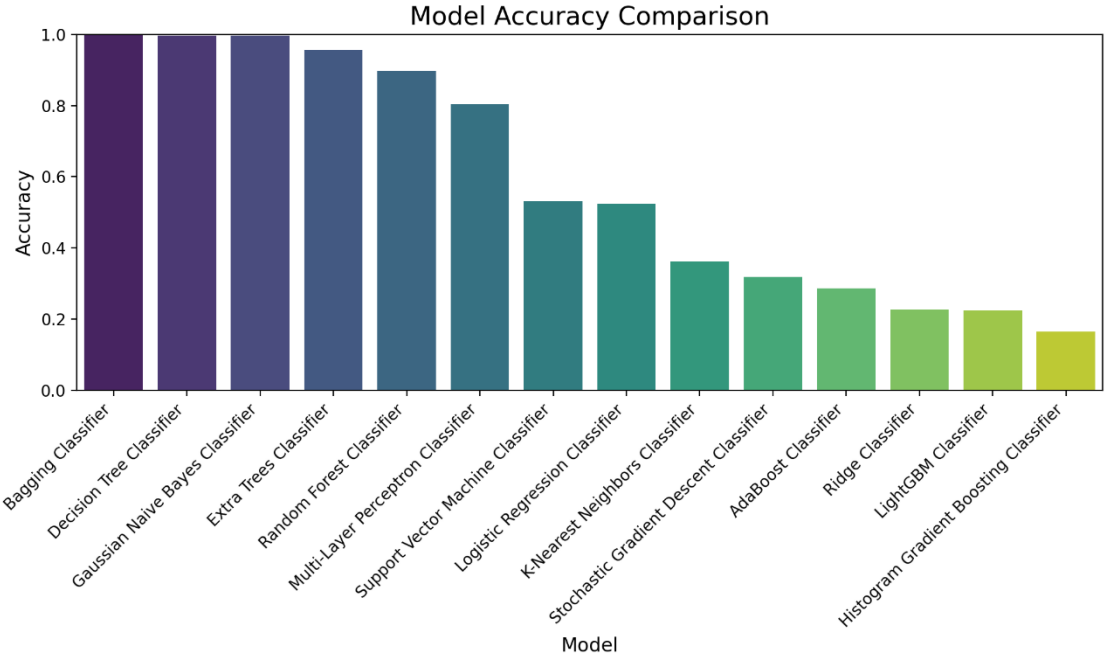

Figure 1. Crop Classification Accuracy Across Models

5.2. Irrigation Requirement Prediction Performance

Ensemble regression models showed good predictive power for the task of forecasting the seasonal average irrigation required. The performance is displayed in Table 2 and Figure 2, where each model's R^2 score represents the percentage of variance it can account for.

Table 2. Performance Metrics for Irrigation Requirement Regression Models

Model	R^2-Score	MAE (mm)	MSE (mm^2)
LightGBM Regressor	0.842	5.95	215.90
Hist. Gradient Boosting	0.841	5.82	217.61
Random Forest Regressor	0.781	6.30	300.37
Extra Trees Regressor	0.771	6.90	314.07
Bagging Regressor	0.754	6.54	337.13
Decision Tree Regressor	0.620	6.60	521.24
MLP Regressor	0.536	12.37	635.82
K-Nearest Neighbors	0.450	12.40	754.28
Support Vector Regressor	0.234	13.10	1049.63

With an R^2 of roughly 0.84, the LightGBM and Histogram Gradient Boosting regressors were the best performing. This means that 84% of the variation in irrigation needs based on input features can be explained by these models. A high level of precision was indicated by these models' extremely low Mean Absolute Error, which was between 5.8 and 5.9 mm. Linear and distance-based models were much less successful than other tree-based ensembles, such as Random Forest and Extra Trees.

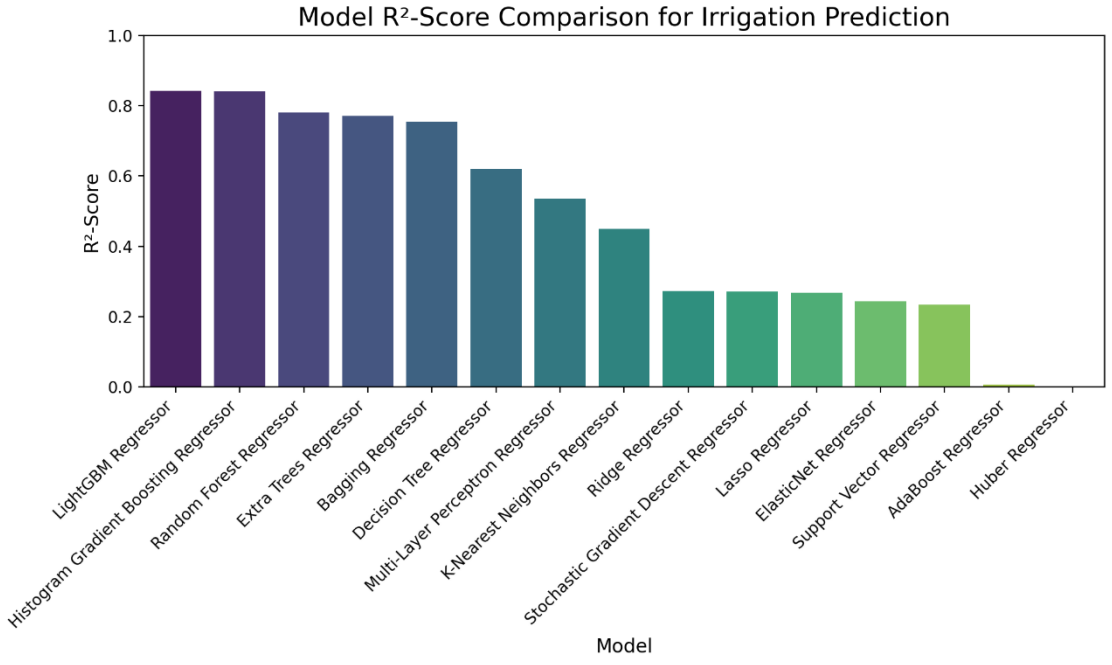

Figure 2. R^2 Scores for Irrigation Prediction Models

5.3. Yield Potential Prediction Performance

Of the three tasks, predicting the Yield Potential Score was the most difficult, but the models still performed admirably. Table 3 and Figure 3 provide a summary of the findings.

Table 3. Performance Metrics for Yield Potential Regression Models

Model	R^2-Score	MAE	MSE
LightGBM Regressor	0.696	5.48	49.84
Hist. Gradient Boosting	0.695	5.47	50.04
Extra Trees Regressor	0.630	6.09	60.69
Random Forest Regressor	0.620	6.15	62.24
Bagging Regressor	0.582	6.45	68.52
MLP Regressor	0.518	7.05	78.94
Support Vector Regressor	0.456	7.57	89.06
AdaBoost Regressor	0.429	7.79	93.55
Decision Tree Regressor	0.228	8.64	126.40

With an R^2 of almost 0.70, LightGBM and Histogram Gradient Boosting once again demonstrated their superior performance. This indicates that they were able to account for roughly 70% of the Yield Potential Score's variation. At about 5.5 points on the 0-100 scale, the MAE was remarkably low. Although this is an impressive result, the lower R^2 when compared to the irrigation task implies that more intricate interactions or variables not fully represented in the dataset affect yield.

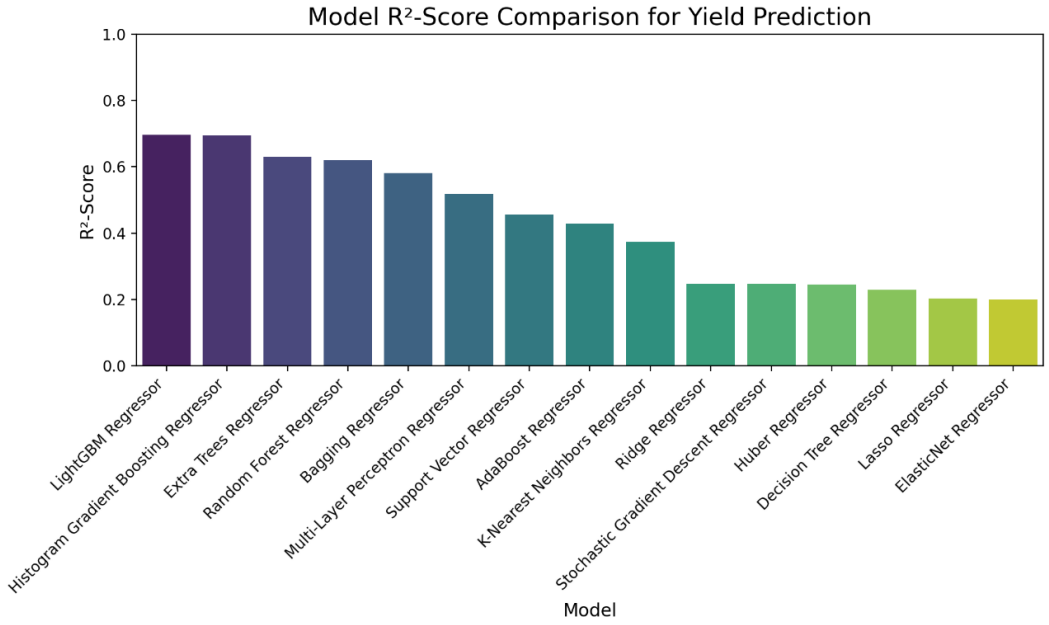

Figure 3. R^2 Scores for Yield Potential Prediction Models

5. DISCUSSION

The study's findings provide important new information about how machine learning can be used to achieve integrated agro-environmental intelligence. Agro-climatic niches for crops are clearly defined and learnable, as evidenced by the exceptional performance in the classification task, with nearly perfect accuracy from multiple models. The fundamental idea of data-driven crop recommendation is validated by the features in the CropIntel dataset, which are adequate to define these niches. Because of the dataset's thoroughness and cleanliness, this result is consistent with—and even surpasses—the high accuracies documented in the body of existing literature (e.g., Rani et al., 2023). Practically speaking, automated systems can offer farmers extremely trustworthy crop recommendations, which could increase land use efficiency and avoid crop

failures brought on by environmental mismatches.

Ensemble tree-based models—especially LightGBM and Histogram Gradient Boosting—perform better in the regression tasks. When modelling agricultural systems, their capacity to capture intricate, non-linear relationships and threshold effects is essential. With an R^2 of 0.84 for irrigation, it is possible to predict water requirements with high confidence. This ability serves as the basis for developing tools that assist farmers in scheduling irrigation events, allocating water resources, and ultimately conserving water—a crucial objective in sustainable agriculture (Navarro-Hellín et al., 2016).

With a top R^2 of about 0.70, the yield prediction task demonstrates the inherent difficulty of predicting agricultural output. Although most of the variance was explained by the models, the remaining 30% points to the impact of variables not included in the dataset, such as the pressure from pests and diseases, severe weather conditions (like hail or frost), or fine-grained management choices. This outcome is in line with the larger body of research on yield prediction, which recognizes that crop growth is multifactorial (Chlingaryan et al., 2018; van Klompenburg et al., 2020). However, a model that accounts for 70% of yield variance is a useful tool for risk assessment, policymaking, and planning.

This work's integrated approach is its main contribution. We show the potential for a comprehensive decision support system by tackling classification, irrigation, and yield prediction in a single framework and dataset. A more thorough and informed planning process could be possible with such a system, which could, for example, not only suggest planting wheat but also estimate its probable water requirements (e.g., 150 mm) and potential yield (e.g., a score of 85/100).

6. CONCLUSION

This study successfully demonstrated the power of an integrated machine learning framework for enhancing agro-environmental intelligence. Using the comprehensive CropIntel dataset, we have shown that it is possible to:

1. **Classify suitable crops** with exceptional accuracy (Acc≈99.7%), providing a reliable basis for crop recommendation systems.
2. **Predict seasonal irrigation needs** with high fidelity (R^2≈0.84), offering a valuable tool for water resource management.
3. **Forecast crop yield potential** with strong performance (R^2≈0.70), enabling better planning and risk assessment.

Our results demonstrate that models based on ensemble trees are especially effective at representing the intricate, non-linear dynamics present in agricultural systems. This study offers a comprehensive approach that goes beyond discrete prediction tasks, opening the door for increasingly complex, comprehensive precision agriculture decision support systems.

There are significant practical ramifications for farmers, agronomists, and legislators. The models created here can optimize the distribution of limited water resources, guide strategic crop selection decisions, and offer useful projections for production scheduling. Although the dataset used in this study is well-structured, a critical next step is to validate and modify these models using actual, in-field data. Future research might also concentrate on integrating economic

factors to optimize profitability in addition to agronomic suitability, as well as dynamic variables like real-time weather forecasts and remote sensing data. We can open new possibilities for creating a food system that is more resilient, productive, and sustainable if we keep bridge the gap between data science and agricultural science.

REFERENCES

1. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69.
2. FAO. (1976). *A framework for land evaluation*. Food and Agriculture Organization of the United Nations.
3. FAO. (2021). *The State of the World's Land and Water Resources for Food and Agriculture – Systems at breaking point (SOLAW 2021)*. Rome: FAO.
4. Goap, A., Sharma, D., Shukla, A. K., & Krishna, C. R. (2018). An IoT based smart irrigation management system using machine learning and open-source technologies. *Computers and Electronics in Agriculture*, *155*, 41-49.
5. Helms, D. (1992). The tonnage of the land: A history of the Soil Conservation Service. US Department of Agriculture, Soil Conservation Service.
6. Arnfield, A. J. (2019). Koppen climate classification | Description, Map, & Chart. In Encyclopædia Britannica. https://www.britannica.com/science/Koppen-climate-classification
7. Navarro-Hellín, H., Martínez-del-Rincon, J., Domingo-Miguel, R., Soto-Valles, F., & Torres-Sánchez, R. (2016). A decision support system for managing irrigation in agriculture. *Computers and Electronics in Agriculture*, *124*, 121–131.
8. Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., & Athanasiadis, I. N. (2022). Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, *276*, 108377.
9. Pudumalar, S., Ramanujam, E., Rajashree, R., Kavya, C., Kaviya, T., & Monisha, B. (2017). Crop recommendation system for precision agriculture. *In 2016 Eighth International Conference on Advanced Computing (ICoAC)* (pp. 32-36). IEEE.
10. Rani, S., Mishra, A. K., Kataria, A., Mallik, S., & Qin, H. (2023). Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, *13*(1), 15997.
11. 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.
12. scikit-learn, StandardScaler. (2019). *StandardScaler*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
13. sklearn, SelectKBest — *scikit-learn 0.23.0 documentation*. (2021.). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html