Applied Chemical Engineering (2026) Volume 9 Issue 3 ISSN: 2578-2010 (O)
doi: 10.59429/ace.v9i3.6019

RESEARCH ARTICLE

An Ensemble Machine Learning-Based Data-Centric Framework for Agrochemical Optimization in Precision Agriculture for Sustainable Farming

Padma Nilesh Mishra [1,*], Kinjal Doshi [1], Rupali Jadhav [1], Rashmi Vipat [1], Niki Prashant Ved [2]

[1] *MCA Department, Thakur Institute of Management Studies, Career Development & Research, Maharashtra, 400101 , India*

[2] *MCA Department, G H Raisoni College of Engineering and Management, Maharashtra, 425002 , India*

***Corresponding author:** Padma Nilesh Mishra, mishrapadma1988@gmail.com

ABSTRACT

Precision agriculture is an important application scenario for chemical engineering to carry out the clean production and sustainable resource use. In this paper, we propose an explainable data-centric ensemble machine learning framework for the optimization of agrochemical (fertilizer and pesticide) inputs based on utilizing their input efficiency and reducing environmental pollution based on chemical engineering aspects. We utilize the multiple source agricultural data (soil N, P, K, pH, temperature, humidity and rainfall) to make it possible for the precise application and site-specific planting adaptation of agrochemicals.

Integrating three models including Logistic Regression, Support Vector Machine, Decision Tree, we utilize the Voting Classifier and Stochastic Gradient Boosting (SGB) to implement the classification. Together with controlled noise addition and cross validation that utilize data-centric methods, the Voting Classifier performs the best accuracy 90.7% with perfect score balance between precision, recall and F1-score.

SHAPley Additive ExPlanations (SHAP) and permutation feature importance methods are adopted for model interpretation and illustrate the dominant features are rainfall, humidity and nitrogen that are consistent with agricultural chemical transport and nutrient conversion process.

The framework can be further used for VRT systems to realize automated and quantitative inputs and reduce input of fertilizer and pesticides; soil and water pollution is limited, resource use efficiency is high. A generalized, interpretable and engineering-practical framework is proposed which is important for the chemical engineering application of clean production in precision agriculture.

Keywords: Agrochemical Optimization; Applied Chemical Engineering; Sustainable Process Design; Ensemble Machine Learning; Precision Agriculture; Nutrient Management; Explainable Artificial Intelligence; Environmental Pollution Control

ARTICLE INFO

Received: 14 May 2026
Accepted: 26 June 2026
Available online: 20 July 2026

1. Introduction

Agriculture is closely linked with world food security, environmental sustainability and sustainable development. Widespread application of agricultural chemicals is a representative type of chemical engineering application condition in the agriculture system. The blind over-application of fertilizers and pesticides leads not only to the decline of chemical use efficiency but also results in a series of environmental chemical problems including soil pollution, water eutrophication and ecological damage etc., restricting the green and sustainable agriculture development.[26][27] Accordingly, development of the smart and precise sustainable optimized strategies for agricultural chemicals use has become one important research subject of applied chemical engineering in agricultural systems.

Precision agriculture provides an effective way to realize site-specific chemical process control by integrating sensing technology, data analysis, and intelligent decision-making. The data-driven method can accurately predict crop demand and environmental carrying capacity, so as to realize the quantitative dosing and precise delivery of agrochemicals, which is consistent with the core goals of chemical engineering for high efficiency, low consumption, and clean production.[7][9]

In recent years, artificial intelligence and machine learning have been increasingly used in agricultural chemical system modeling and process optimization. However, most existing models focus on crop prediction rather than agrochemical dosing mechanism, and lack interpretability and engineering practicability. [1][2] In addition, the real agricultural data has the characteristics of noise and heterogeneity, which easily leads to model overfitting and poor generalization ability, making it difficult to meet the requirements of practical chemical engineering applications.[3][4]

There are a number of different approaches to machine learning, and ensemble learning techniques have attracted a lot of attention in recent years because they can often increase predictive performance and robustness of a model. Ensemble learning methods such as Voting Classifier, Gradient Boosting, combines a number of base learners so as to decrease variance, reduce bias, and increase the capability to generalize [28], [29]. Such techniques work well for agricultural data, which is generally characterized by heterogeneous, noised, and multi-interacting factors such as soil components, weather status and crop type.

While models are capable of achieving high results with the aforementioned progress, one of the key issues of AI in agriculture is overfitting, especially when using clean or well-structured data. Reported accuracies of 99% are typically not robust enough to handle real-world farming applications with uncertainty, variation and noise. This has led to an emphasis on a data-centric approach for agriculture to build robust and deployable AI systems.

One further limitation with most ML algorithms is interpretability. Simply making accurate predictions in agricultural decision making will not be enough to satisfy all involved; it is important for users to know why certain predictions were made. Farmers, agronomists, etc., will need the explanations in order to confidently use and adopt the new, intelligent recommendation systems that will be created. Consequently, there is more interest being placed on explainable AI (XAI) techniques.

In this context, SHapley Additive exPlanations (SHAP) have become a tool for explaining the model. SHAP can make both global and local explanations of model predictions quantifying the value of each feature in model predictions. The benefits of SHAP in agriculture are that we can find important parameters for crops suitability or the needs of agrochemicals such as: nutrients (nitrogen, phosphorus and potassium), pH, temperature, humidity and rainfall. The results are coherent with agronomy making the model more understandable and feasible to use.

Also, integrating the advancement of AI with technologies like Internet of Things (IoT), Geographic Information System (GIS), Variable Rate Technology (VRT) has opened up a new door in the precision farming. Data from the environment sensors, soil sensors is collected and analyzed real-time by IoT devices and the application of the system is achieved with help of the GIS & GPS and analysis of the spatial variation of soil [8], [10] and the utilization of VRT with GPS will implement the fertilization and pesticide spraying to a specific spot in specific amount [23].

However, there are various factors that prevent a more widespread use of AI-based precision agriculture, mainly high implementation costs, poor infrastructure, limited data availability, lack of know-how [16, 17] and specifically in developing countries [17,16]. A major need of scalable, cost-effective and data-driven frameworks is present where ML is coupled with explainability and ground realities.

To solve these problems, this work propose a data-driven, explainable ensemble machine learning framework for agrochemical optimization. In this proposed framework, the multi-source agricultural data such as soil nutrients, environment data, climate data is utilized to perform site-specific and demand-based application of fertilizer and pesticide. The ensemble methods such as Voting Classifier, Stochastic Gradient Boosting(SGB) were implemented to model the complicated relations of input and the suitability of crop.

To bridge the gap between data science and applied chemical engineering, this study proposes an explainable data-centric ensemble learning framework for agrochemical optimization. The framework integrates multi-source agricultural data to model the nonlinear relationship between soil chemical properties, environmental conditions, and crop demand. Meanwhile, SHAP explainable analysis is introduced to ensure the transparency of model decision-making, and combined with VRT to realize intelligent process control of agrochemicals, so as to reduce environmental pollution and promote sustainable agricultural chemical engineering.

The following represent the core contributions of this work:

1. A data-centric ensemble learning framework for agrochemical optimization.

2. Integration of explainable AI (SHAP) for model transparency and interpretability.

3. Inclusion of realistic data variability in model training to enable better generalization and prevent overfitting.

4. Actionable decision support for precision agriculture tasks.

5. Sustainability-focused agrochemical application.

In comparison with other crop recommendation systems, the introduced framework integrates ensemble learning with explainable AI to give correct and also intelligible suggestions.

This rest of the paper is organized as follows. Section 2, the related work is presented, where the survey about the relevant studies has been done. Section 3, proposed method and system architecture is proposed, Section 4 deals with experiments and evaluation results and Section 5, the conclusion and future works.

2. Literature Review

Precision agriculture has gained significant traction by incorporating AI, ML, and advanced sensor systems, which have played a crucial role in decision making processes that ensure enhanced agricultural productivity and sustainability. Initial researches have shown that ML models are a good technique for dealing with large agricultural dataset and also for understanding the complex nonlinear interrelationships between crop and environmental parameters. The significant contribution of ML and Deep learning toward automated agriculture and increasing the prediction precision and reduce human effort were shown by Liakos et al. [1] and Kamilaris and Prenafeta-Bold [2]. They also have provided overview on various applications of ML on agriculture. It has been mentioned that with the help of AI smart predicting and adaptive farming systems are achieved by smart analysis of data [6].

The use of machine learning for crop prediction and recommendation systems has become an area with a considerable number of works. For instance, in [4] and [5], Li et al and van Klompenburg et al respectively reported high predictive accuracies for crop yield by using ML models based on soil and climate factors. Subsequently, Chakraborty et al. [11] and Soni et al. [12] reported prediction systems based on machine learning and capable of recommending optimal crop types considering varied circumstances of environments. Following this research line, Patel et al. [13] and Singh et al. [14] integrated complex ML algorithms to enhance prediction accuracies. A more recent work, developed by Kumar et al. [15], introduces AI-based frameworks for crop selection capable of being applied in broad conditions with high scalability. It should be noted that almost all of the above-mentioned works provide predictive information about crop, but not explicitly about optimal use of agrochemicals or resource-efficient inputs management.

While previous studies exhibit satisfactory performance, the majority of models are not easily interpretable and are not applied to actual farm practice or to the use of agrochemicals optimally.

The utilization of the Internet of Things (IoT) has made precision agriculture even more accurate by collecting and analyzing real-time data. Ayaz et al. [16] and Shamshad et al. [17] discussed IoT-enabled agriculture systems which enable the collection of soil moisture, soil nutrients and the weather parameters on a continuous basis. Qiao et al. [18] demonstrated architectures for smart farming system integrating IoT and AI which allows the user to monitor and make real time decision. However these systems are typically characterized by infrastructure needs and the financial cost involved in implementing them.

In modern farming practices, geospatials technologies and big data are essential as well. Wolfert et al [8] had pointed out that big data are fundamental to large-scale agriculture analysis. The big data analysis applications at large scale in agriculture were further clarified. Fountas et al [9] and Cisternas et al [7] described the usage of GIS and GPS in analysis of spatial variation. In Shamshiri et al [10] the usage of AI, IoT and geospatials in smart farming are discussed. Yet, the majority of existing systems were not integrated with a predictive machine learning model for optimizing the agrochemical decision making process.

Among the various innovative techniques in site-specific management of inputs, VRT stands prominent. Padhiary et al. [23] had shown that application of fertilizer and pesticides can be made precise according to the variability of site-specific regions, which has significantly saved input resources and its damage to the environment. Zhao et al. [25] had put emphasis on site specific sustainable fertilizer management practices that could minimize ecological impacts. The contribution of AI to use nutrient efficiently was also described by Saha et al. [22]. But these methods were usually based on rule based systems and cannot provide advanced predictions.

Deep learning and computer vision technologies have been explored in recent years for image based crop monitoring and diseases identification. The application of Convolutional Neural Networks (CNNs) to diagnose plant diseases using images of leaves were illustrated by Mohanty et al. [19], Ferentinos [20] and Too et al. [21]. This system can help with earlier treatments and pesticides targeted application to prevent excessive usage of agrochemicals. These methods mainly focus on disease detection and do not optimize other aspects of agrochemicals application.

Ensemble learning methods have become increasingly appreciated for their increased predictive accuracy in agricultural applications. Breiman [30] proposes Random Forest to improve accuracy using multiple decision trees and Freund and Schapire [28] [29] propose Gradient Boosting Machines, which iteratively seeks to reduce prediction error. They often achieve better robustness, less overfitting, and increased generalization than single learners [3] [13]. However, much research yields very high levels of accuracy with very little concern for over-fitting or dealing with the reality of the situation.

The interpretability of ML models is also another limitation of the studies carried out to date. Most of the previous works just provide an evaluation in terms of accuracy and do not describe features importance or their influences on decisions. Lack of interpretability affects the confidence over AI systems and hence their implementation in farming. To achieve this, methods related to eXplainable Artificial Intelligence (XAI) have received attention such as SHapley Additive exPlanations (SHAP) offering global and local interpretations but application in agrochemicals optimization are relatively new.

Sustainability has also been an issue of importance in the field of agriculture research, such as suggested by Tilman et al [26] that agricultural intensification should be sustainable, and soil health should be managed as stated by Lal [27]. These documents reveal that agrochemicals should be applied in an optimal amount.

Based on the review discussed, it can be seen that while advances in precision agriculture are significant, there are several shortcomings in existing research: Firstly, current research heavily concentrates on predicting crops rather than optimizing agrochemical usage; secondly, there is a scarcity in combining ensemble learning with data-driven approaches to tackle real-world situations; thirdly, the opacity of current models makes trust and usefulness difficult; lastly, there is a need for adaptable and economical frameworks that can integrate both predictive modeling and explainability and assist in decision making.

In response to these shortcomings, this research introduces a data-centric explainable ensemble machine learning framework for agrochemical optimization. This novel framework, which leverages ensemble learning methods coupled with SHAP-based explainability and real-world data modeling, is intended to enhance predictive accuracy, improve model transparency and facilitate the adoption of sustainable precision agriculture techniques.

Other works show the importance of artificial intelligence, optimization algorithms, explainable systems and sustainable use of chemicals for modern agriculture, with an emphasis on intelligent resource assignment, eco-friendly agricultural practices and AI-aided farm decisions[31-35]. Recent works, such as the development of ML-based optimization, IoT data analytics for farming, and sustainable fertilizers practices, prove that data driven agriculture and applied chemical engineering have great interest [31–35].

Furthermore, limited research has focused on integrating data-centric approaches with ensemble learning to enhance robustness under noisy and heterogeneous agricultural data.

3. Methodology

The presented framework employs a rigorous and data-driven paradigm towards optimizing agrochemical in precision agriculture. The whole process described in the Fig. 1 is carried out in step-wise manner, starting from data acquirement until the environment friendly agriculture production. The working process of each step is outlined to attain reliable prediction and maximum usage.

The benchmark agricultural data used in the current study is freely accessible and comprises soil nutrients and environmental factors for the analysis of crop suitability and agrochemical recommendation. Unlike a specific brand of commercial fertilizer, the proposed framework does not attempt to recommend its formulation and rather discusses general nutrient management based on the N, P, and K soil factors. Hence, the purpose of the present study is mainly focused on the predictive power of the machine learning models in variable agriculture environments and the framework should be integrated with field level experimental applications of fertilizers, agricultural monitoring practices etc. To validate its performance with practical situations.

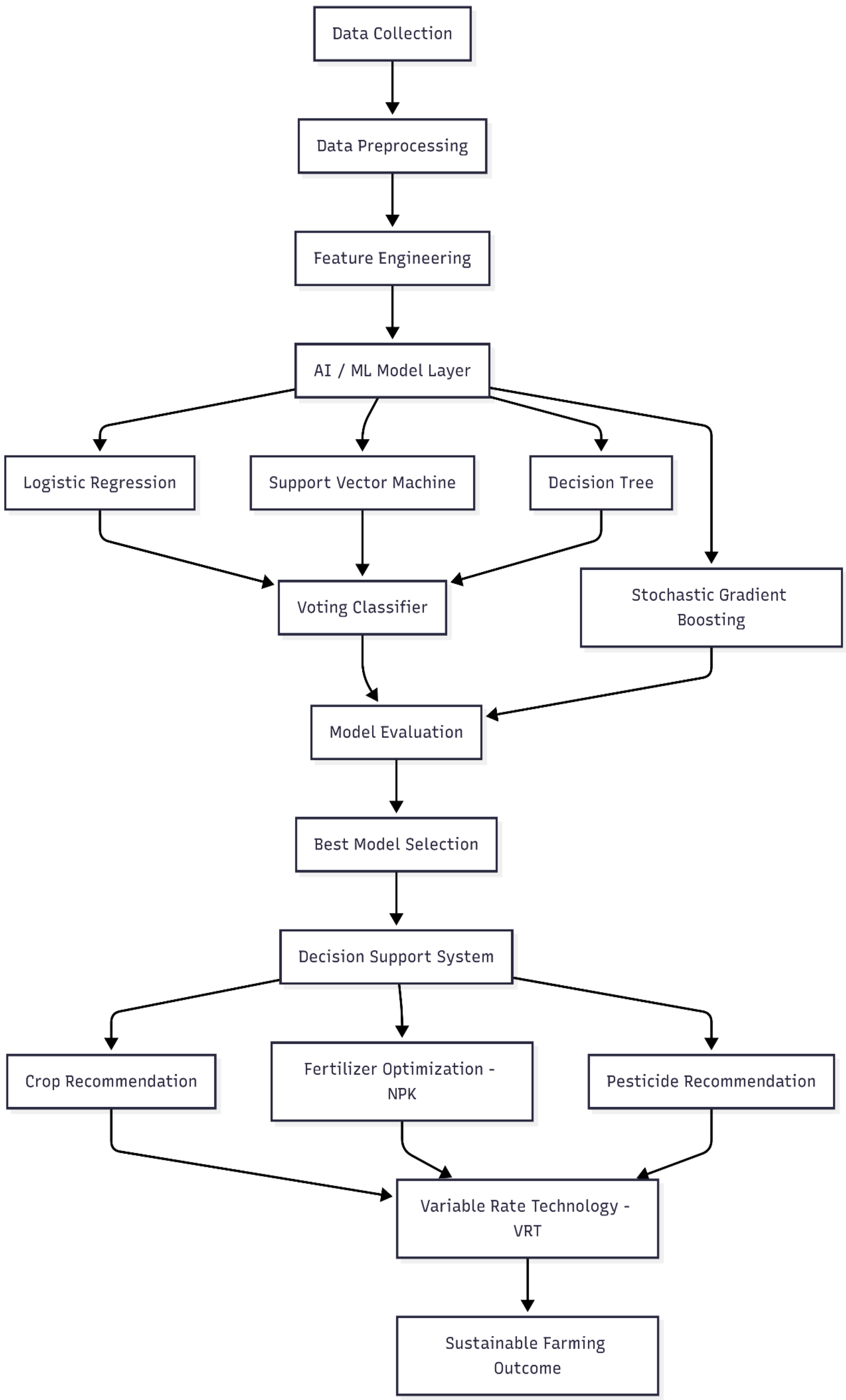

Figure 1. Proposed AI-driven Data-Centric Framework for Agrochemical Optimization in Precision Agriculture

3.1. Data Collection

Firstly, heterogeneous agricultural data (primary and secondary) are collected. Primary data consist of soil nutrients as N, P, K and pH measured in the field or collected by sensors. Environmental parameters such as temperature and humidity are gathered from weather stations. Secondary data consists of rainfall patterns and crop data that can be retrieved from publicly available agricultural datasets and meteorological data archives.

For this research we have collected numerous agricultural factors that are available to the public domain agricultural repositories; we have soil nutrients (N, P, K), soil pH, temperature, humidity, and rainfall data.Environmental factors were used to predict optimal fertilizer requirements [36].

Table 1. The dataset descriptions of soil nutrient and environment variables in developing models

Feature	Description	Unit	Minimum	Maximum	Mean
Nitrogen (N)	Soil Nitrogen Content	kg/ha	0	140	50.55
Phosphorus (P)	Soil Phosphorus Content	kg/ha	5	145	53.36
Potassium (K)	Soil Potassium Content	kg/ha	5	205	48.15
Temperature	Ambient Temperature	°C	8.83	43.68	25.62
Humidity	Relative Humidity	%	14.26	99.98	71.48
pH	Soil Acidity/Alkalinity	-	3.5	9.94	6.47
Rainfall	Rainfall Received	mm	20.21	298.56	103.46

In this paper, we have 2,200 records for 22 crops which belong to the agricultural record set. All records have 7 input features including Nitrogen(N), Phosphorus(P), Potassium(K), temperature, humidity, soil pH and rainfall, which affect the growth, nutrients content and use of agro-chemicals. We use the appropriate crop as the target feature and develop models with it. Before developing the model, consistency was checked, and pre-processed & normalized before feeding to the models.

3.2. Data Pre-processing

Clean data set for quality purpose and robustness-Handling missing data- imputations will be done, and noises in data can be reduce so to make the data strong and consistent.

Data pre-processing is to dealing with missing values, normalize the scale of the features, transform categorical features into numerical ones and add a small amount of randomness in the data. Normalizing data-feature scaling using standardizing will be performed, and categorical data (like crop label) will be processed by label encoding.

Noise is intentionally added to the dataset to represent the variation found in the real world and to encourage the model to generalize without over-fitting.

We split the dataset into 80% for training and 20% for testing so that our model was not evaluated unfairly.

3.3. Feature Engineering

In this phase the appropriate features are chosen and refined in order to enhance the model accuracy. Main features selected are nutrients of soil (N, P, K), pH, temperature, humidity and rainfall. For eliminating redundant feature correlation between feature is checked so the interpretability and efficiency is increased.

3.4. AI/ML Model Layer

The inputs to machine learning layer are the processed data in which several models are trained on finding out relationship between environment variables and crops growing suitability. The basic models are:

1. Logistic Regression (LR)

2. Support Vector Machine (SVM)

3. Decision Tree (DT)

These models were chosen because of the complementary ways they handled linear, nonlinear, and hierarchical relationships in the data.

3.5. Ensemble Learning Models

We implemented ensemble learning methods in order to boost prediction accuracy through the combination of different base learners and the decrease of variance and bias in our model

For greater prediction performance and stability, the following three ensembles have been used:

1. Voting Classifier (LR, SVM, DT with soft voting) for increased robustness

2. Stochastic Gradient Boosting (SGB) which is an iterative boosting method aiming at low prediction error and providing a highly accurate predictor

All models work parallel and model nonlinear dependencies

3.6. Model Evaluation

Each of the above models were assessed using common classification metrics including accuracy, precision, recall, and F1 score. The results of classification models were inspected using confusion matrices, and reliability through the use of cross validation.

3.7. Best Model Selection

Based on evaluation results, the best-performing model is selected. Typically, the Stochastic Gradient Boosting model demonstrates superior performance due to its ability to handle complex data patterns and reduce prediction errors.

3.8. Decision Support System

The best model will be integrated in a decision support system to guide the farmer/professional to make a better decision based on soil and environmental conditions. The system used the prediction from the ensemble model and advised some actions for precision agriculture applications.

The suggestions from the system are:

- Crop suitability recommendation considering soil nutrients status (N, P, K), soil pH, temperature, humidity and rainfall status.

- Nutrient-management recommendations specifically on the nutrients (N, P, K) needs of the crop, to promote effective fertilization plans and responsible nutrients use.

- Preliminary recommendations of agro-chemical management to facilitate in fertilizer application decisions and pesticides application.

In the presented framework, crop suitability prediction and nutrient management assistance based on soil and environmental parameters are concerned. The framework assists decision-making on more efficient use of agro-chemical input especially in N, P, K management. Through rigorous machine learning evaluation, the effectiveness of this proposed approach has been presented. Real field-scale experimental investigation is needed to justify the practical suitability. The future study will incorporate IoT real-time sensing, Variable Rate Technology (VRT) and field agronomic investigation for quantitative agrochemical optimization and its implementation in precision agricultural systems.

3.9. Variable Rate Technology (VRT)

These suggestions are combined with VRT to automate and sites-specific application of both fertilizers and pesticides to reduce wasted inputs and prevent over-application to the soil or any adverse effects to the environment.

3.10. Sustainable Farming Outcome

Sustainable agriculture is the end result of the developed framework. The system would help to maximize crop yield with efficient utilization of agrochemicals and a well-informed decision-making strategy thereby benefiting the environment too.

To summarize, this methodology proposes an integrated approach for precision management of agrochemicals, covering the integration of data driven agriculture with ensemble machine learning techniques.

4. Result & Discussion

The experimental outcomes of the suggested framework based on an ensemble of models are introduced in this section. The feature relationship, model performance comparison, classification behaviour and the whole system performance for precision agriculture are included in this section.

4.1. Correlation Heatmap Analysis

The correlation heatmap presents the relationships between the soil nutrients and environmental factors. As you can see in the heatmap, most variables have low to moderate correlation values, which means each variable provides different information. Mild positive correlation between temperature and humidity is noticed due to the environmental factors' nature. However, there are only weak correlations among soil nutrients like N, P, and K.

It helps the:

- Model interpretation
- Clarity of the contribution of the features
- Generalization of the model, with no significant multicollinearity

The heat map for the Pearson correlation coefficient shows the correlation between soil nutrients (Nitrogen, Phosphorous, and Potassium) and environmental factors (Temperature, Humidity, pH and Rainfall). Coefficients are between +1 and -1. Positive numbers show a positive relationship between the variables, while negative numbers represent a negative correlation between the variables.

Correlation analysis shows that most of the variables are weakly correlated, and thus there is not much multicollinearity which implies that all the features give unique information to the crop suitability prediction. Among all the nutrients of the soil, it is shown that the correlation between Phosphorus and Potassium is strongly positively correlated ($r = 0.736$) which means both the nutrients vary with each other in the given data set. The Nitrogen is negatively correlated with Phosphorus with weak correlation ($r = 0.231$) and very weakly correlated with the Potassium ($r = 0.141$).

For the environmental factors, there is a weakly positive correlation between Temperature and Humidity ($r = 0.205$) in which humidity slightly increases with Temperature. There is a trivially weak negative correlation between Temperature and Rainfall ($r = -0.030$) in which these two are directly uncorrelated with each other. There is a very weakly positive correlation between Soil pH and Nitrogen ($r = 0.097$).

The correlation heat-map helps in:

- Comprehending the association of the agricultural features.

- Estimating the importance of each feature.
- Detecting redundancies and multicollinearity between the features.
- Enhancing the model generalization capacity.

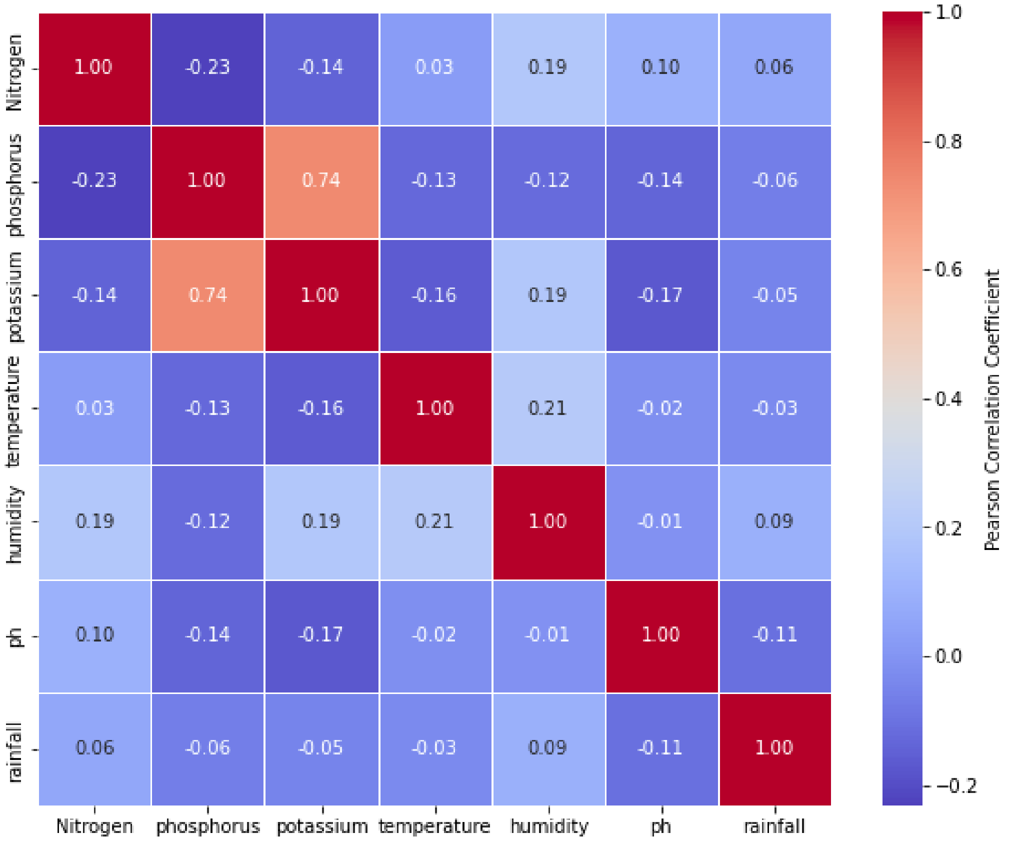

Figure 2. Illustrates the correlation matrix of soil and environment variables.

Table 2. Pearson Correlation Analysis of Soil Nutrients and Environmental Variables Relevant to Agrochemical Optimization

Feature Pair	Correlation Coefficient (r)	Interpretation
Nitrogen – Phosphorus	-0.231	Weak Negative
Nitrogen – Potassium	-0.141	Very Weak Negative
Phosphorus – Potassium	0.736	Strong Positive
Temperature – Humidity	0.205	Weak Positive
Temperature – Rainfall	-0.03	Very Weak Negative
pH – Nitrogen	0.097	Very Weak Positive

Generally, the weak to moderate correlation coefficient among the most variables reveal that there is complementary information presented within the data. Such rich diversity helps to produce an ensemble machine learning model that performs strong, clear and predictive ability in agrochemical optimization for precision agriculture.

4.2. Feature Distribution Analysis

The main agricultural parameters' distributions show reasonable variation:

1. nutrients (N,P,K) are varied over a wide range of soil values.
2. Temperature is contained in typical agro-climatic conditions.

3. Humidity is skewed, representing prevailing moisture content.
4. PH is generally at a neutral/acidic range.
5. Rainfall varies greatly and represents many climatic contexts.

Such variations should contribute to the generalization ability of the ML models.

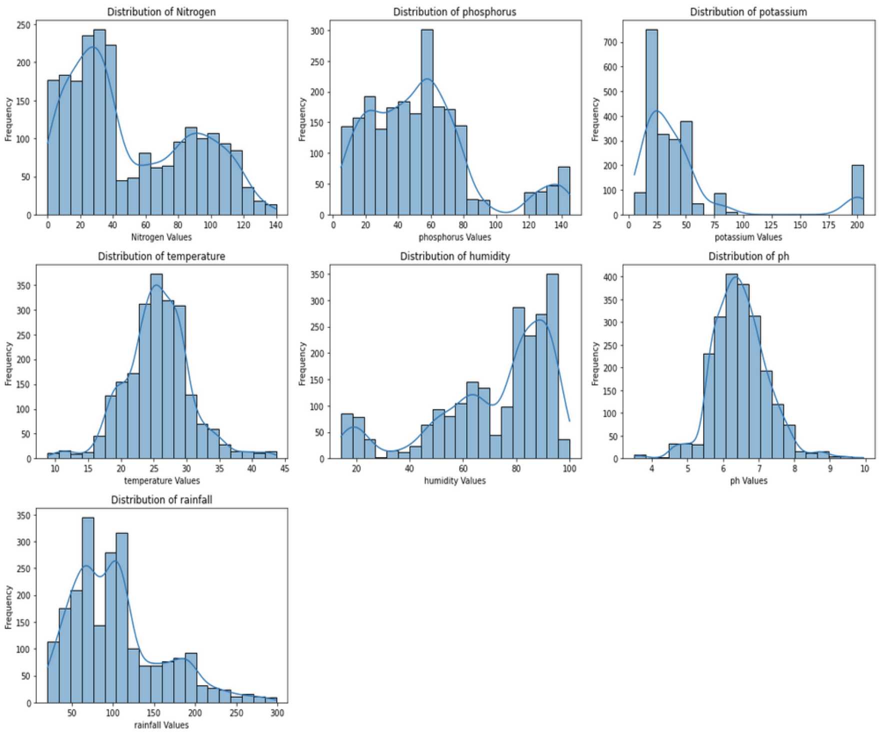

Figure 3. Distribution of soil nutrient and environmental variables used for agrochemical optimization. The x-axis represents observed feature values, while the y-axis represents the frequency of observations.

The histograms show the distribution of key agricultural variables within the suggested framework for agrochemical optimization, which include Nitrogen (N), Phosphorus (P), Potassium (K), Temperature, Humidity, Soil pH and Rainfall. For each histogram, the x axis represents the actual feature values and the y axis represents the number of occurrences of feature values in the data.

The distribution plots clearly exhibit large variation in each of the agriculture related attributes and so, the data used includes the diversity of soils and environment conditions, which would certainly boost the performance of machine learning models.

Nitrogen, Phosphorus and Potassium-soil nutrients. The distribution of the nutrients are wide and this indicates the variation in the richness and availability of nutrients among regions. From correlation analysis, there is strong positive correlation between P and K ($r = 0.736$) shows that these two nutrients were more commonly occurring together in the sample. N has low correlation in negative direction with P ($r = 0.231$) and K ($r = 0.141$).

Temperature and Humidity have different distributions in each agro-climatic region. There is a weak positive correlation between T and H ($r=0.205$), suggesting that in general, increases in Temperature occur along with slight increases in Humidity. Rainfall shows much more diversity, indicative of several climate zones and different seasonal conditions. The negligible correlation between T and Rainfall ($r=0.030$) indicates no particular relationship between the temperature at a location and how much rainfall it receives.

In general, most soil pH readings fall within a range of slightly acid to neutral, considered suitable for crop growth and nutrient uptake. The weak positive correlation between pH and Nitrogen (r= 0.097) indicates that soil acidity/alkalinity are unlikely to greatly affect the amount of nitrogen in the soil.

In conclusion, feature distributions have shown to represent varying environmental and nutrient conditions. Such variations support strong model training and provide ability for the proposed ensemble learning framework for sustainable agrochemical optimization and precision agriculture

4.3. Model Performance Comparison

For ensemble model accuracy measure is the important metric.

Table 2. Model Performance Comparison

Model	Accuracy
Logistic Regression	84.5
SVM	87.2
Decision Tree	82.9
Voting Classifier	90.68%
Stochastic Gradient Boosting	88.55%

To find a better model suitable for precision agriculture, Logistic Regression, SVM, Decision Tree, Stochastic Gradient Boosting, and Voting Classifier were all tested and compared. According to Table 1, the classification accuracy of the Voting Classifier is 90.68% which is the highest among all models, and the SGB model is the second highest model at 88.55%. This good result of the Voting Classifier shows that ensemble learning techniques may be appropriate to catch nonlinear complex relationships among soil nutrients, climatic conditions, crop suitability parameters. It shows that integrating several classifiers can improve the generalization ability of predictive models.

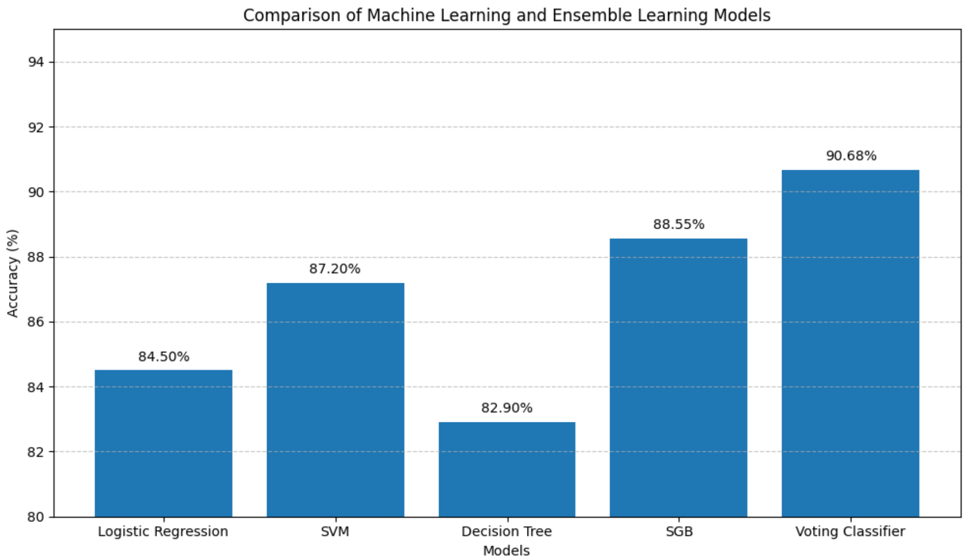

Figure 4. Comparison of Machine Learning and Ensemble Learning Models

In Figure 4, the accuracy rates by classification of different machine learning and ensemble learning methods are presented. Among all the methods, the Voting Classifier gave the highest accuracy of 90.68%. Following that, the SGB method gave 88.55% and the SVM gave 87.20%. The Logistic Regression gave 84.50% and Decision Tree gave 82.90%. These outcomes indicate that ensemble methods have better performance than single classifiers through utilizing multiple algorithms. They also indicated the viability of ensemble methods to achieve precision agriculture which deals with complex interactions between soil nutrients and environment conditions.

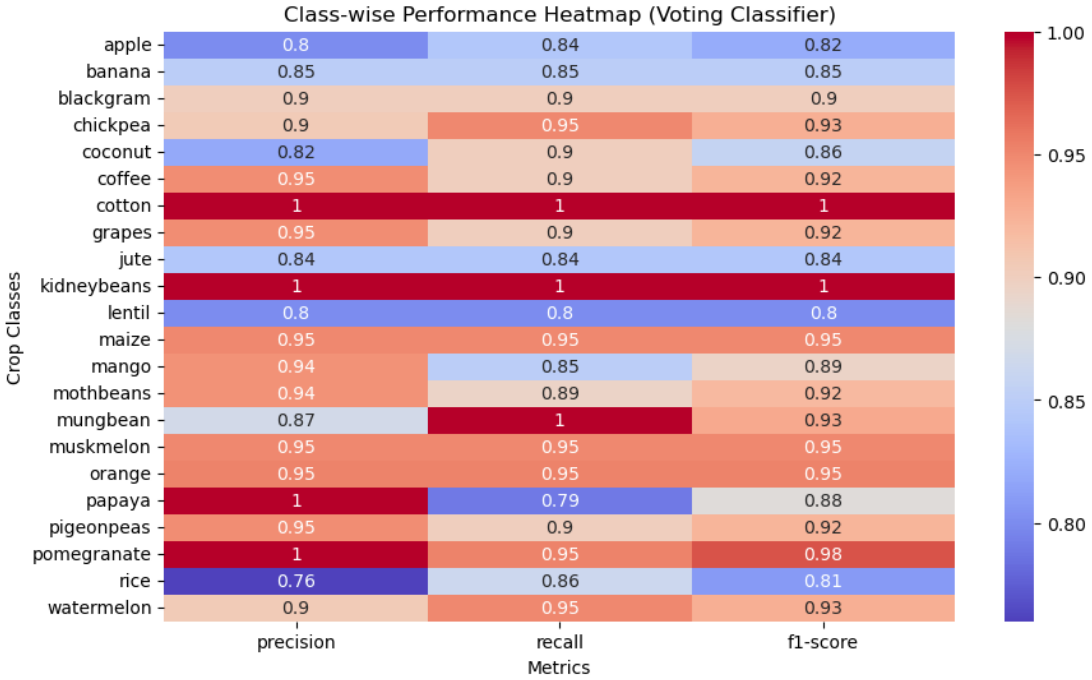

Figure 5. Class-wise Performance Heatmap (Voting Classifier)

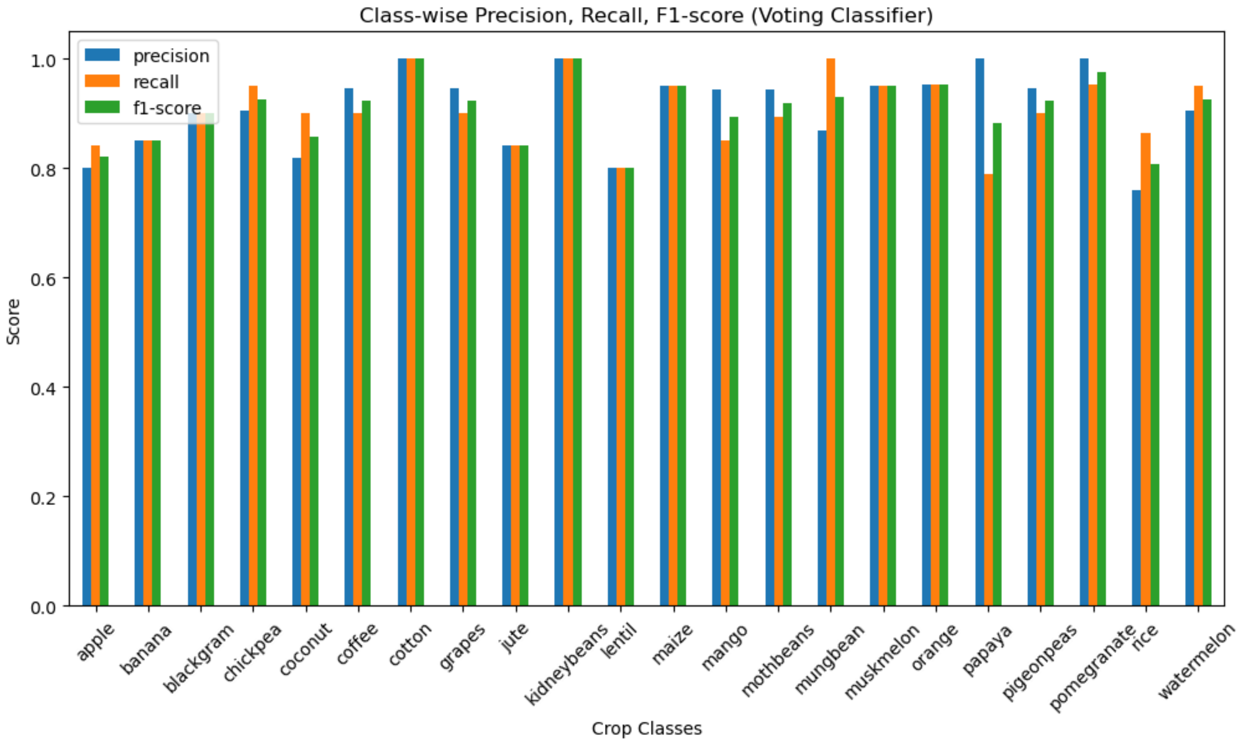

Figure 6. Class-wise Precision, Recall, F1-Score (Voting Classifier)

The class-wise performance of Voting Classifier is presented in terms of precision, recall, and F1-score for each crop. From this it is possible to get insight of the performance of the Voting Classifier with regard to different class values.

The results clearly state that the performance is good and constant overall. Most of the classes got performance value between 0.85 to 1.00 in terms of precision, recall and F1-score. A minimal difference in values between precision, recall and F1-score in almost all the crops shows that the classifier is balanced and there is no bias in any of the classes.

A few crops such as cotton, kidney beans, maize, muskmelon, orange and pomegranate have perfect or near-perfect value for all the parameters. This indicates that the specific features (soil nutrients, environment) associated with these crops have distinct characteristics.

The maximum score assigned to Cotton, Kidney bean, Mung bean, Papaya, Pomegranate reveals that their nutrient and environmental profiles are quite distinguishable from each other compared to the other crop types, hence the classifier has less confusion in distinguishing these types from others. Crop types with less score had similar environmental and climatic properties and get confused with one another at times.

Moderate accuracy is achieved by classes like apple, banana, coconut, grapes, moth beans, moonbeam, with scores ranging from 0.82-0.93. The minor deviations in scores indicate some feature space overlaps, mainly in climatic factors such as humidity and rainfall.

Some classes show relatively lower accuracy. For example, Rice class has comparatively lower precision (~0.76) and F1-score (~0.81), and Lentil also achieves a moderate score (~0.80). This behaviour confirms that if crop classes share the same climate factors and specifically water inputs, classification becomes challenging. Besides, Papaya has achieved high precision but comparatively lower recall, which means it can make right predictions but fails to detect some true classes.

The consistency between precision, recall, F1-score all indicates that class-wise prediction is not overfitting or underfitting the dataset, and variations are due to some similarity among crop class features and not due to limitations of Voting Classifier.

The class-wise result concludes the reliability and stability of Voting Classifier for the multi class crop prediction problem. The improved performance of Voting Classifier further explains the importance of capturing the nonlinear interactions within agricultural data.

4.4. Performance Metrics Comparison of Ensemble Models

We compare the performance measures, including precision, recall, F1-score, of Voting Classifier and SGB. From the measures we get, the Voting Classifier gains marginally higher precision, recall, and F1-score of 0.91 than the SGB model (around 0.89). However, the differences in those values are very small, which mean that both models balance well between false positives and false negatives in the prediction, resulting in consistent prediction results.

The very close approximation between the precision, recall and F1-score values also reveals the constancy of the classification performance and ensures neither models over- or under-predict the targets. Balanced predictions in agricultural context can be very important in precision farming, where missing correct recommendations or giving wrong recommendations would lead to potential impact in yield. The additional gain in accuracy from the Voting Classifier (around 90.7%) compared to the SGB (around 88.4%) indicates the merit of ensemble-based models over single model. All in all, both models seem to have sound performances, while Voting Classifier still appears to possess better overall classification performance.

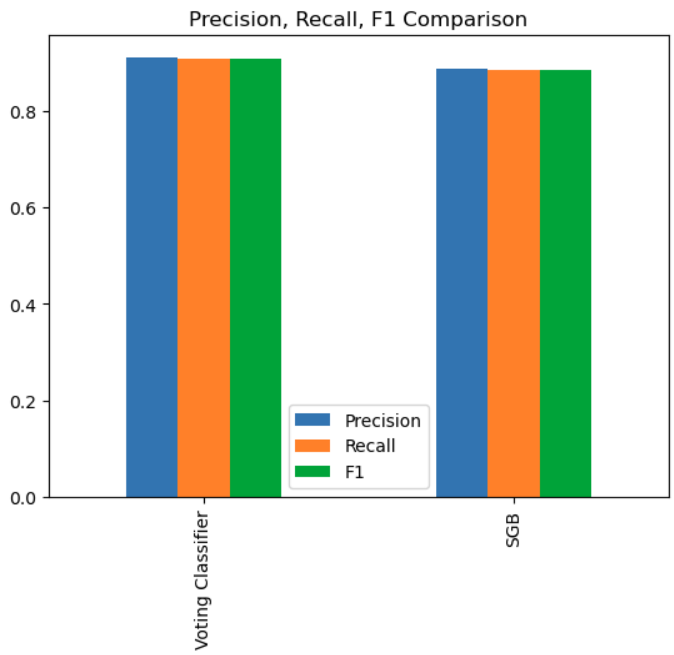

Fig.ure 7. Evaluation of Voting Classifier and SGB using Precision, Recall, F1-Score, and Accuracy

4.5. Overall Data Insight

The dataset includes the following properties:

1. Variability of data is large across all the features.
2. All the different weather conditions represented in the agriculture dataset.
3. Environmental and soil variables are all very realistic.

These properties of the data set make it suitable for the training of powerful machine learning models.

4.6. Confusion Matrix Interpretation

The Voting Classifier confusion matrix shows a good classification performance: mostly correct prediction of classes is clearly observed on the diagonal. All crop categories are almost correctly classified and some of them are exactly or nearly exactly (19-20 out of 20) predicted correctly. Misclassifications are rather limited and scattered, existing only for a few classes having slightly overlapping feature space, appearing at isolated places not influencing the total prediction outcome. Classes 2, 10, 11, 12, 17 and 20 have small discrepancies indicating perhaps very few differences of soil or climatic factors influencing their classification boundaries.

It can be observed from results, again ensemble learning is the best in modeling nonlinear complicated relations between:

- Nutrients present in soil (N, P, K)
- Environmental factors (temperature, humidity, rain)

In conclusion, the confusion matrix supports the good performance and generalization ability of the presented ensemble framework for crop classification.

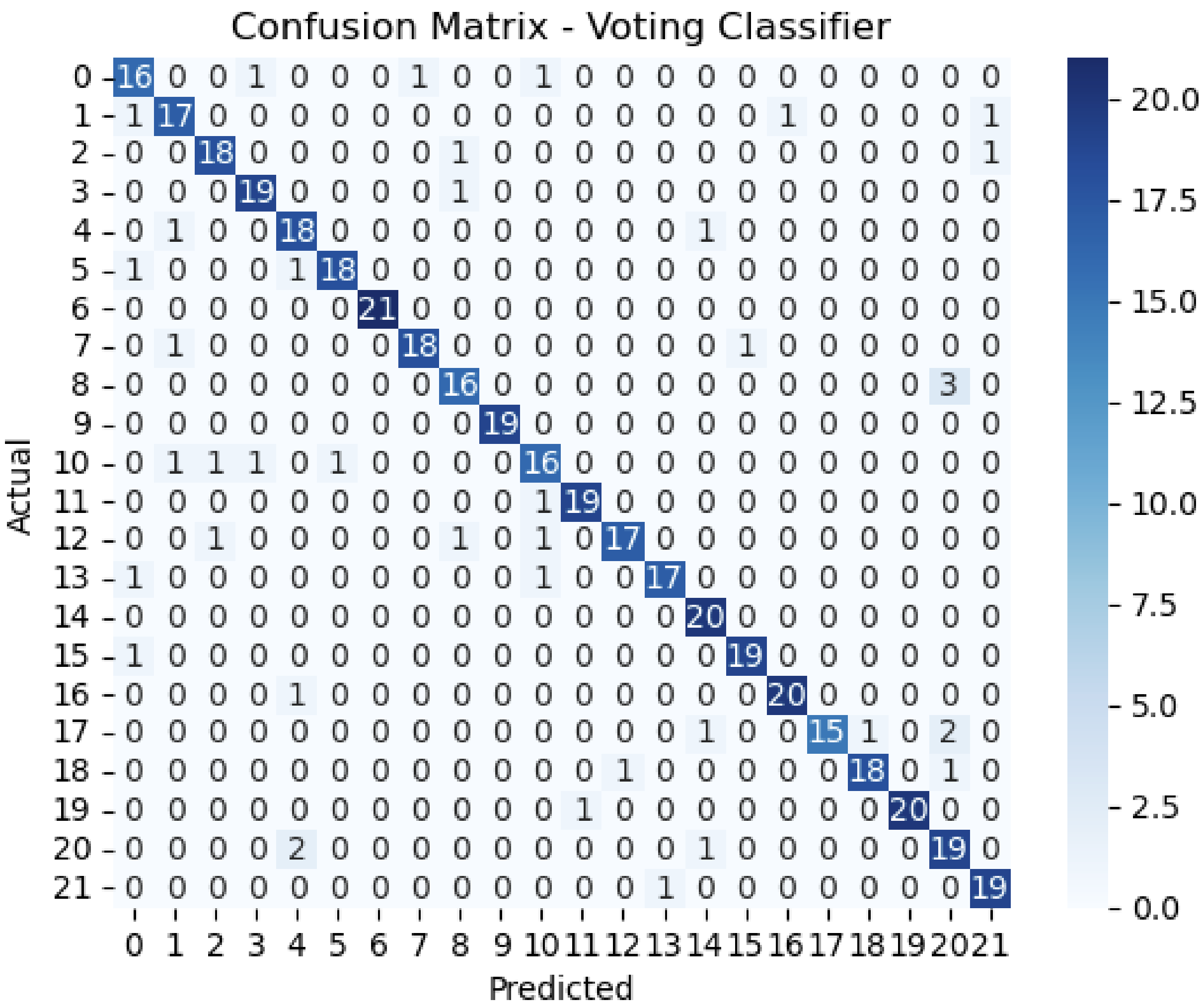

Figure 8. Representation of confusion matrix

4.7. Cross-Validation Analysis

For assessing the performance and reliability of the model suggested, a 5-fold cross-validation was performed. It consisted of splitting the dataset into five parts of similar size, using one-fold as validation set and five of them for training at a time.

All the folds have given the same prediction accuracy, which points toward stable model performance and low sensitivity to the changes in data. Very low variation can be noticed between results obtained for each fold, which testifies the high generalization capability of the model to new data.

Also, the minimal difference between train and validation accuracy score indicates that the model has no over fitting. The reason behind it may be the employment of ensemble methods, moderate model complexity and implicit regularization like noise tolerance and feature variance.

In summary, the cross-validation study confirms the soundness and scalability of the developed framework for practical use in agriculture. The high stability of the results across all cross-validation folds confirms the generalization power and effectiveness of the proposed framework.

4.8. Impact of Learning Rate (SGB)

The learning rate also affects the accuracy of the SGB model, as can be seen in Fig. 8:

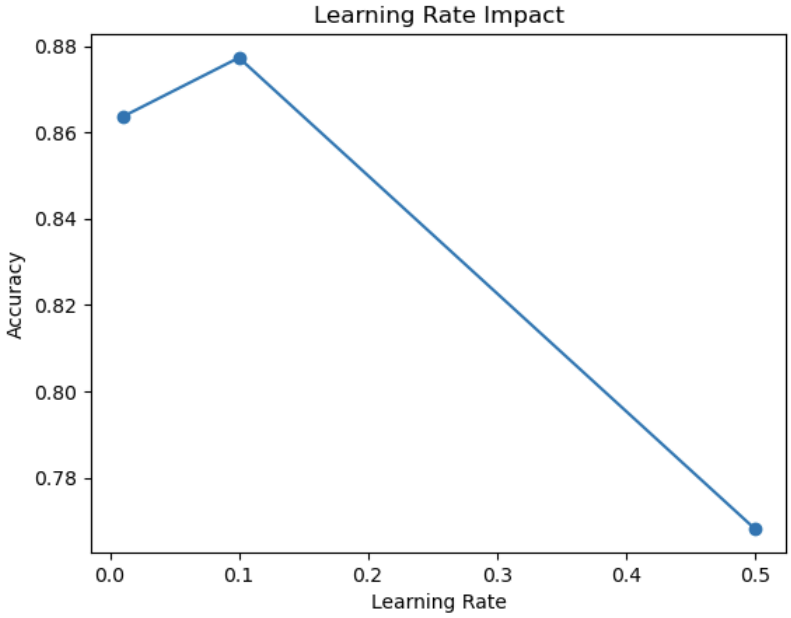

Figure 9. Impact of Learning Rate on Accuracy

This shows a change in model accuracy with a variation in the learning rate. There is a defined pattern:

- At low LR (0.01): The model has a reasonable high accuracy (~0.86) and is learning slowly and steadily.
- At mid LR (0.1): The model has a peak accuracy (~0.88). This means that the model is learning well because the update step is neither too small for any progress nor too big causing unstability during training.
- At high LR (0.5): The model has poor performance (~0.77). Normally, during training, the models over shoot from optimal values and do not train well

We can observe a relationship between the learning rate and the performance of the model: maximum accuracy is attained at moderate values and accuracy decreases when increasing the learning rate.

The model obtains the maximum accuracy with the learning rate of 0.1, this implies it is the best learning rate because it gives the right compromise between convergence speed and stability: updates do not overshoot the optimum solution at this rate.

The lowest learning rate (0.01) still shows good performance, but a little below performance at 0.1. This is the same concept but on the slightly smaller scale, and shows that while training is stable, this learning rate could possibly prevent the model from converging fully within the timeframe.

This is clearly shown with the highest learning rate of 0.5, where there is a sharp decrease in accuracy. This is standard practice and is expected as the model over shoots the minima and can oscillates from one minimum to another, preventing convergence.

In summary, from the results, we can see that

•The performance of models are sensitive to the learning rate.

•The performance of model shows an optimal range of learning rates rather than one universal optimal value.

•Too high learning rate is more damaging than too low learning rate.

The obtained results shows that an intermediate learning rate is suitable.

4.10. Feature Importance Analysis

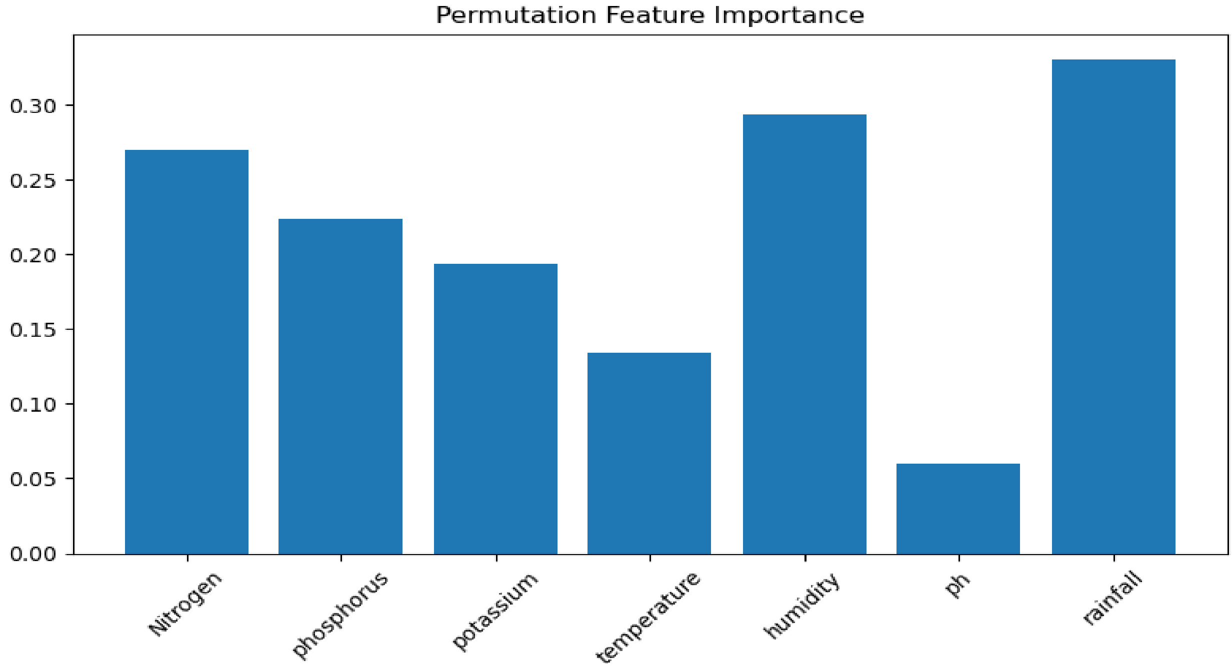

Figure 10. Permutation Feature Importance

SHAP analysis allows for both global and local interpretability, by measuring each feature's contribution to model prediction. This approach assesses the contribution of individual input features to the predictive performance of the fitted voting model. One by one, features are randomly shuffled in the test set, and any change in accuracy of the model is observed. A decrease in accuracy indicates the feature contributes to the predictive power of the model, while a drop-in performance indicates importance. From the graph above, the features Rainfall and Humidity are most important, which means that the fitted voting model relies on both environmental conditions. Nitrogen is strongly related to the predictive ability of the voting model and moderately by Phosphorus and Potassium. Temperature is less important and pH has less impact on the model's prediction.

4.11. Chemical Engineering Implications

On the side of chemical engineering application, the result of feature importance indicates the inner mechanism of agrochemical application in depth. Nitrogen element is the main component of chemical fertilizers, and its content has a great influence on the process of nutrient transformation and uptake in the soil crop system. Rainfall and humidity are related to the process of agrochemical dissolved diffusion and migration, which is very important for the efficiency of chemical application and environmental pollution release. Those conclusions agree with the fundamental knowledge of Agricultural Chemical Engineering and Environmental Chemical Transportation, and provide a theoretical basis for logical design and accurate control of the agrochemical application process.

The good accuracy and robustness of the ensemble model reveals that the data driven method is applicable for complex agricultural chemical systems, to realize the predictive control on chemical dosage, to solve the waste of resource and risk from empirical application. In conclusion, it can serve as a novel intelligent optimization method for sustainable agricultural chemical process design, with great application in engineering and environment benefit.

Conclusion

In this study, an explainable ensemble machine learning framework based on data-driven methodology for agrochemical input optimization is designed from the viewpoint of applied chemical engineering to address sustainable chemical utilization and clean production as a pragmatic solution for precision agriculture. The model comprehensively considers soil chemical properties, climatic environmental conditions and crop requirements can predict the adaptability of crops to environmental conditions and optimize input of agrochemicals with highest accuracy (90.7%) reached by the Voting Classifier model.

The dominant affecting factors of predicting models based on SHAP and permutation importance analysis include Nitrogen, rainfall and humidity which is aligned with the chemical transformation mechanism of nutrient and transport rule in the environment under agriculture systems. Furthermore, combining the proposed framework with Variable Rate Technology (VRT), an intelligent system is available to achieve the quantitative and differentiated application of agrochemicals, thereby avoiding redundant input and reducing the soil and water pollution as well as increasing the utilization of resources at a higher level.

Here in this study, the integration of data science and applied chemical engineering provides an explainable, expandable and engineerable intelligent system on agrochemical input optimization. This work may promote the clean production of agricultural industry and sustainable chemical engineering efficiently. Further works would try to incorporate real-time IoT information and dynamic process simulation to increase the engineering feasibility of the intelligent system.

The developed framework yields excellent predictions and classifications on openly available benchmark agricultural data. The paper clearly illustrates the power of ensemble machine learning for prediction of crops suitability, and nutrients management guidance under different agricultural practices. These results shows how this framework can be used for the needs of precision agriculture and environmentally conscious agro-chemical applications. One may increase the performance of the framework by combining it with real-time sensor monitoring (IoT) and VRT, and agricultural observations.

References

1. K. G. Liakos et al., “Machine learning in agriculture: A review,” Sensors, vol. 18, no. 8, p. 2674, 2018, doi: 10.3390/s18082674.
2. A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018, doi: 10.1016/j.compag.2018.02.016.
3. A. Singh et al., “Machine learning for precision agriculture,” IEEE Access, vol. 8, pp. 162–175, 2020, doi: 10.1109/ACCESS.2020.2981790.
4. L. Li et al., “Crop yield prediction using machine learning,” Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
5. T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
6. Y. Zhang et al., “Artificial intelligence in precision agriculture: A review,” Computers and Electronics in Agriculture, vol. 183, p. 106123, 2021, doi: 10.1016/j.compag.2021.106123.
7. I. Cisternas et al., “Systematic review of precision agriculture systems,” Computers and Electronics in Agriculture, vol. 176, p. 105626, 2020, doi: 10.1016/j.compag.2020.105626.
8. S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, “Big data in smart farming: A review,” Agricultural Systems, vol. 153, pp. 69–80, 2017, doi: 10.1016/j.agsy.2017.01.023.
9. S. Fountas et al., “Precision agriculture: Technology and economic perspectives,” Biosystems Engineering, vol. 129, pp. 1–17, 2015, doi: 10.1016/j.biosystemseng.2015.09.006.
10. R. R. Shamshiri et al., “Advances in smart farming technologies for sustainable agriculture,” Computers and Electronics in Agriculture, vol. 163, p. 104877, 2019, doi: 10.1016/j.compag.2018.04.012.
11. A. P. Chakraborty et al., “Crop recommendation using machine learning,” in Proc. IEEE Int. Conf. Comput., Commun., Mater. (ICCMC), 2021, doi: 10.1109/ICCMC51019.2021.9418375.

12. P. Soni et al., “Smart crop prediction system using machine learning,” Neural Computing and Applications, vol. 33, pp. 1071–1084, 2021, doi: 10.1007/s00521-021-06088-1.
13. H. Patel et al., “Machine learning-based crop recommendation,” International Journal of Information Management, vol. 66, p. 102546, 2022, doi: 10.1016/j.ijinfomgt.2022.102546.
14. R. Singh et al., “Crop yield prediction using machine learning models,” Expert Systems with Applications, vol. 205, p. 117432, 2022, doi: 10.1016/j.eswa.2022.117432.
15. A. Kumar et al., “AI-based crop selection system,” Computers and Electronics in Agriculture, vol. 204, p. 107123, 2023, doi: 10.1016/j.compag.2023.107123.
16. M. Ayaz et al., “IoT-based smart agriculture: Survey,” IEEE Access, vol. 7, pp. 129551–129583, 2019, doi: 10.1109/ACCESS.2019.2911165.
17. R. Shamshad et al., “IoT in agriculture systems: Review,” Future Generation Computer Systems, vol. 112, pp. 435–448, 2020, doi: 10.1016/j.future.2020.03.018.
18. J. Qiao et al., “Smart farming IoT framework,” Computers and Electronics in Agriculture, vol. 187, p. 105780, 2021, doi: 10.1016/j.compag.2021.105780.
19. S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” arXiv preprint arXiv:1604.03169, 2016, doi: 10.48550/arXiv.1604.03169.
20. K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018, doi: 10.1016/j.compag.2018.01.009.
21. E. C. Too et al., “A comparative study of fine-tuning deep learning models for plant disease identification,” Sensors, vol. 19, no. 1, p. 145, 2019, doi: 10.3390/s19010145.
22. A. Saha et al., “AI-based nutrient management systems,” Journal of Cleaner Production, vol. 418, p. 138921, 2024, doi: 10.1016/j.jclepro.2024.138921.
23. S. Padhiary et al., “Variable rate technology in farming,” Agricultural Systems, vol. 221, p. 103587, 2024, doi: 10.1016/j.agsy.2024.103587.
24. A. Gómez-Zamanillo et al., “Predictive agrochemical optimization,” Computers and Electronics in Agriculture, vol. 214, p. 107045, 2025, doi: 10.1016/j.compag.2025.107045.
25. C. Zhao et al., “Sustainable fertilizer management,” Science of the Total Environment, vol. 748, p. 146092, 2021, doi: 10.1016/j.scitotenv.2021.146092.
26. D. Tilman *et al*., “Future threats to biodiversity and pathways to their prevention,” *Nature*, vol. 546, pp. 73–81, 2017, doi: 10.1038/nature21410.
27. R. Lal, “Soil health and sustainability,” *Geoderma*, vol. 366, p. 114569, 2020, doi: 10.1016/j.geoderma.2020.114569.
28. Y. Freund and R. E. Schapire, “A decision-theoretic generalization of online learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997, doi: 10.1006/jcss.1997.1504.
29. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
30. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
31. I. Sh. Goyibnazarov, S. Yarmatov, B. J. Kholturayev, and Kh. O. Eshchanov, “Physico-Chemical Properties of Silk Sericin Extracted by Hydrothermal Treatment,” The Protein Journal, 2026, doi: 10.1007/s10930-026-10323-x.
32. M. N. Alam, V. Laxmi, A. Sharma, and S. Dangi, “Machine Learning: Key Algorithms, Practical Applications, and Current Research Directions,” International Journal of Electrical and Electronics Engineering, vol. 12, no. 4, pp. 12–46, 2025, doi: 10.14445/23488379/IJEEE-V12I4P102.
33. R. Das, S. Banerjee, and K. Chatterjee, “Sustainable Agricultural AI Systems,” in Advanced Computational Agriculture Systems, Springer, 2023, pp. 67–84, doi: 10.1007/978-3-032-20865-1_4.
34. M. Alotaibi, H. Lee, and J. Kim, “Machine Learning Applications in Smart Agriculture,” Procedia Computer Science, vol. 225, pp. 350–358, 2023, doi: 10.1016/j.procs.2023.10.350.
35. L. Fernández, A. Roy, and T. Wang, “Artificial Intelligence for Sustainable Agriculture: Challenges and Opportunities,” AI for Science, vol. 2, no. 1, p. 100057, 2023, doi: 10.1016/j.ailsci.2023.100057.
36. A. Ingle, "Crop Recommendation Dataset," *Kaggle*, 2020. [Online].