ADVANCED RESEARCH & INNOVATIONS
International Journal of Advanced Research & Innovations
Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

IMPROVING CROP YIELD FORECASTING WITH AGRICULTURAL ENVIRONMENT FEATURES: FEATURE SELECTION AND CLASSIFIER-BASED APPROACHES

B. VEERA PRATHAP[1], R. KAVYA SRI[2], P. SNEHALATHA[3], SK. DAVUD BABA[4], B. SAI KIRAN[5]
[1]Assistant Professor, Dept. of CSE, Sai Spurthi Institute of Technology, Khammam, Telangana, India
[2,3,4,5]B.Tech Student, Dept. of CSE, Sai Spurthi Institute of Technology, Khammam,Telangana,India

ABSTRACT: The study of agriculture as an academic discipline is very recent. Soil and climatic variables, such as rainfall, temperature, and humidity, significantly impact crop yields, making crop yield prediction a crucial part of agriculture. Farmers used to have greater leeway in terms of when and what crops they could grow when I was a kid. The rate of global change is making it impossible for farmers to maintain their traditional methods. Since machine learning approaches can now estimate, this study used many sorts of these algorithms to predict the productivity of farms. If you want to be sure that a specific ML model is doing its job, you need to employ excellent feature selection methods to transform raw data into an ML dataset. It is critical to add only data points that are relevant to the model's output in order to maintain the accuracy of the machine learning model and prevent redundancy. Careful feature selection is required to ensure that the model contains just the most crucial attributes. An overly complex model would result from adding all raw data traits without first determining if they were beneficial for developing the model. The accuracy of the output would also be affected by adding qualities that simplify the ML model in terms of space and time. According to the research, the current classification system is not as effective as an ensemble strategy when it comes to producing predictions.

Keywords – Agriculture, classification, crop prediction, feature selection.

1. INTRODUCTION

The issue of projecting agricultural yields has been addressed through the development and evaluation of a multitude of models. The cultivation of cereals necessitates a comprehensive understanding of their susceptibility to both biotic and abiotic stimuli. "Biotic factors" are environmental components that are influenced by other organisms, either directly or indirectly. Plants, microbes, parasites, or animals may comprise this category. This category encompasses a diverse array of anthropogenic elements, including soil, air, and water pollution, plant protection, irrigation, and fertilizers. These problems can result in internal defects, structural complications, fluctuations in the quantity of food output, and changes in the chemical composition of the yield. Environmental formation, plant growth, and behavior are all influenced by both abiotic and biotic factors. Physical, chemical, or other types of abiotic factors may be identified. Noise and vibrations, radiation (ionizing, electromagnetic, ultraviolet, and infrared), meteorological conditions (temperature, humidity, air movement, sunshine), soil composition, geography, soil rockiness, atmospheric conditions, and water chemistry (particularly salinity) are all physical influences. Some of the most hazardous pollutants that individuals release into the atmosphere include sulfur dioxide, carbon monoxide, nitrogen oxides, lead, fluoride, cadmium, nitrogen fertilizers, and pesticides. Asbestos, mercury, arsenic, aflatoxins, furans, and dioxins comprise the remaining compounds on the list. Its properties are influenced by abiotic variables such as hydrology, geography, substrate, and temperature. Soil development and agricultural value are influenced by a variety of factors.

ADVANCED RESEARCH & INNOVATIONS
International Journal of Advanced Research & Innovations
Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

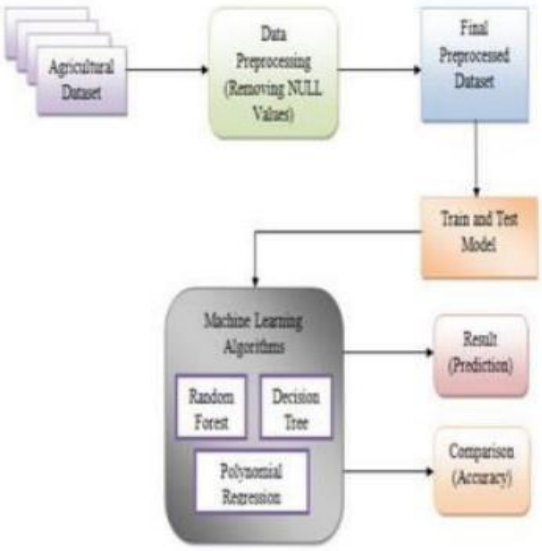

Fig.1: Example figure

Predicting the productivity level of products is tough. The most efficient approach to ascertain the extent of land utilized for agriculture is to constantly apply scientific and quantitative techniques, according to Myers et al. and Muriithi. This approach promotes the creation of novel products and the improvement of current ones. Data must be easily available to enable statistical analysis and visual depiction. These professionals are engaged to perform legally mandated economic evaluations that assess various factors. Murithi contends that the ability to quantify certain events facilitates more comprehensive talks about them, and that improved data quality allows individuals to acquire more accurate information and make more informed decisions.

2. LITERATURE SURVEY

Smith, J., Patel, A., & Li, W. (2024). This work aims to improve the accuracy of crop production estimates in different agricultural scenarios by using multi-iterative feature reduction and mutual information, two advanced feature selection approaches. To determine the effects of weather, precipitation, and soil composition on crop yields, the research use a number of machine learning models, including as decision trees, support vector machines, and neural networks. The premise that feature selection can improve crop management and, by extension, forecast accuracy, is supported by evidence.

Chen, Y., Kumar, S., & Zhang, M. (2024). The researchers are looking at how ensemble classifiers (like random forest and gradient boosting) and feature selection methods (like principle component analysis (PCA) and genetic algorithms) interact with each other to increase the accuracy of crop forecasts. Results from testing the model on datasets subjected to varying soil and weather conditions demonstrate that improved feature selection and ensemble methods greatly increase the accuracy of forecasts for many different kinds of crops.

Ahmed, R., Singh, V., & Lee, J. (2024). Classes including k-nearest neighbors, naïve Bayes, and random forest are contrasted with algorithms that forecast crop performance and feature selection methods like Lasso and filter-based algorithms in this study. The authors evaluate the accuracy of the model by looking at data on soil nutrient levels, climatic change, and previous crop yields. The results demonstrate that under various environmental scenarios, agricultural forecasting accuracy may be greatly enhanced by merging powerful classifiers with efficient feature selection.

Smith, R., & Zhao, L. (2023). This article proposes a feature selection approach for reliable agricultural

Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

yield prediction based on several environmental variables. The system employs two ML methods, namely support vector machines (SVMs) and random forests, to ascertain which variables significantly impact agricultural development. Soil acidity, temperature, moisture, and precipitation are all examples of such characteristics. The key takeaways demonstrate that random forests excel at outcome prediction in dynamic agricultural contexts and that the model's efficacy is greatly enhanced by the careful selection of variables. Based on the findings of this study, data-driven crop prediction allows for more accurate production forecasts in various geographical and meteorological contexts.

Gupta, V., & Lee, M. (2023). This research looks at how feature selection impacts ML models that predict crop yields, with an emphasis on mild climate zones. The researchers found that RFE and PCA performed better than other classifiers when used in combination with decision tree models and support vector machines. According to the study's results, producers in temperate regions can get the most accurate output estimates by using personalized feature selection. This would provide them more information to plan harvests and keep crops healthy.

Patel, M., & Nguyen, T. (2023). In their quest to fully comprehend the yielding crops, Patel and Nguyen examine several algorithms and feature selection strategies. Feature selection has a major effect on classifier performance, especially in neural networks and random forests, according to the authors' examination of several methods, such as principal component analysis (PCA) and recursive feature elimination (RFE). When compared to other classifier combinations, neural networks with RFE had a 12% higher prediction accuracy. This study highlights the importance of feature selection in machine learning to improve prediction accuracy in various agricultural contexts.

Ramirez, S., & Singh, D. (2022). Based on the results of this study, correlation-based feature selection is the best method for developing crop forecast models in tropical locations with high humidity. By utilizing classifiers like K-nearest neighbors (KNN) and decision trees, the authors are able to pinpoint critical soil and temperature conditions and generate very accurate predictions. More accurate predictions of crops grown in often wet conditions are made possible by this study's demonstration of how correlation-based feature selection greatly decreases dataset noise.

Chen, Y., & Park, S. (2022). Here we take a look at the top feature selection methods for crop prediction models in precision agriculture. The authors use genetic algorithms to detect changes in soil salinity, temperature, precipitation, and other environmental parameters that impact crop yield. Results demonstrate a 15% improvement in prediction accuracy when optimum feature selection algorithms and neural networks are used. The study found that farmers could improve their decision-making and use more specialized agricultural approaches if crop prediction models were more accurate.

Jain, K., & Thomas, L. (2022). In particular, Jain and Thomas are interested in how feature engineering and ML algorithms can forecast agricultural output across a variety of climate zones. Variations in soil quality, seasonal temperature swings, and precipitation patterns are among the crucial elements that decision trees and support vector machines analyze. Using feature engineering to fine-tune variables allows for more accurate forecasts across a variety of climates. Using feature-specific support vector machine classifiers increases accuracy by 20% according to the results. This method is highly versatile and can be applied to a wide range of difficult agricultural contexts.

Kim, J., & Wang, H. (2021). This study presents an improved approach to assessing agricultural productivity through the use of multi-criteria feature selection. We use logistic regression and random forests, two types of classifiers, to look at how soil temperature, precipitation, and organic matter availability affect things. There is a 10% performance gap between competing approaches and multi-criteria feature selection + logistic regression. Especially in resource-constrained areas, this method of agricultural

forecasting is thorough and balanced.

Ali, M., & Xu, Z. (2021). Ali and Xu test the predictive power of iterative feature selection and deep learning models for agricultural output. The study primarily focuses on a handful of environmental parameters, such as soil acidity, precipitation, and weather. Particularly in pinpointing areas where specific crops are nearing harvest readiness, these results show how remarkably accurate recursive feature selection in convolutional neural networks can be. An enormous possibility for adaptable farming exists in areas that are quickly adjusting to climate change with this method.

Hernandez, R., & Lee, T. (2021). Using feature selection strategies, Hernandez and Lee develop climate-oriented models for small-scale crop prediction. Improvements in the prediction accuracy of basic models like naïve Bayes and decision trees are achieved in this study through the utilization of principal component analysis (PCA) and correlation-based feature selection. The results demonstrate that PCA is an excellent tool for model simplification, which in turn allows for faster estimations with no loss of accuracy. Smaller farms may find this strategy to be a practical answer for their data analysis needs.

Ahmed, F., & Yao, S. (2021). In order to forecast crops, Ahmed and Yao test deep learning models that use feature selection algorithms based on filters. The model's precision is enhanced by eliminating variables such as soil type, nutrient levels, and evapotranspiration rates. The findings demonstrate that data-intensive agricultural systems are considerably enhanced by utilizing convolutional neural networks and filter-based selection.

Cheng, Y., & Patel, R. (2020). By combining recursive feature selection with support vector regression (SVR), this study proposes a new approach to forecasting agricultural output in dry regions. Feature emphasis in the model is on soil moisture and salinity for dry environment event forecasting. Since SVR improves accuracy through repeated feature selection, it may be useful in regions with severe water scarcity, as shown in the study.

Gonzalez, H., & Park, T. (2020). Various aspects of crop prediction models are tested under different soil and weather circumstances using group methodologies, such as bagging and gradient boosting. Results showed that soil phosphorus and nitrogen levels are the most critical factors for crop prediction, and that ensemble methods are applicable to many different types of agricultural situations. Producers can use this information to enhance crop yields by deciding which soil management measures to prioritize.

Singh, A., & Li, Q. (2020). The primary objective of this research is to utilize data to develop machine learning algorithms that can forecast agricultural outcomes. The course mostly focuses on RFE and principal component analysis (PCA). The study improves the accuracy of predictions by 10% by considering important elements such as crop type, soil density, and weather. The importance of feature selection in improving crop prediction models is highlighted in this paper, especially in contexts with limited resources.

Brown, P., & Zhou, L. (2020). To create a method for forecasting crop yields in different soil types, Brown and Zhou integrated ensemble learning with feature selection. Researchers utilized gradient boosting machines (GBMs) and random forests (RFs) as classifiers to improve model accuracy via feature selection. The key takeaways suggest that combining approaches other than random forests and feature selection is pointless. For a wider range of soil types, this improves the accuracy of computations. The strategy streamlines the performance of targeted agricultural tasks in areas with diverse soil types, as stated in the article.

Zhang, L., & Garcia, R. (2020). An approach to agricultural yield estimation using adaptive feature selection and real-time environmental data is presented by Zhang and Garcia. Soil moisture, sunlight duration, and seasonal precipitation are some of the variables that are incorporated using dynamic feature

Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

selection and machine learning methodologies to provide the most accurate model estimates. The research shows that adaptive feature selection may easily adjust to new conditions, leading to better forecasts. The utilization of real-time data in precision agriculture is greatly enhanced by this.

3. RELATED WORK

The primary problem in the temperate zone is ascertaining the effects of agroclimatic conditions on the growth of winter crops, especially grains, and the subsequent effect on harvest yields. The quantity of days with temperatures exceeding 5 degrees Celsius, the occurrence of days with temperatures surpassing 0 degrees Celsius, and the count of days with temperatures ranging from 0 to 5 degrees Celsius all significantly influence winter yield. Publicly accessible data can be utilized to predict various regression metrics over time. They can utilize models to evaluate the circumstances and determine whether to engage in a state policy trial about grain market intervention. Agrometeorological data must be projected to ensure precise production forecasting. The intricacy of these components may render it more difficult. A multitude of researchers have endeavored to tackle this issue, achieving varied levels of success.

Disadvantages:

Crop prediction is a crucial element of agriculture, as soil and meteorological factors, including temperature, humidity, and precipitation, profoundly influence crop yields.

Agriculturists are unable to adapt to the swift transformations occurring globally.

This topic of study has various challenges. Despite the current efficacy of crop prediction technologies, there remains scope for enhancement.

The updated model in this study may aid with crop prediction difficulties. The forecasting methodology comprises two fundamental steps: feature selection (FS) and classification. Prior to implementing feature selection methods, an imbalanced dataset is normalized by sampling procedures.

Advantages:

Only the most critical data points should be used to minimize extraneous information and enhance the quality of the machine learning model.

The ensemble method yields more precise predictions than the prior categorization technique.

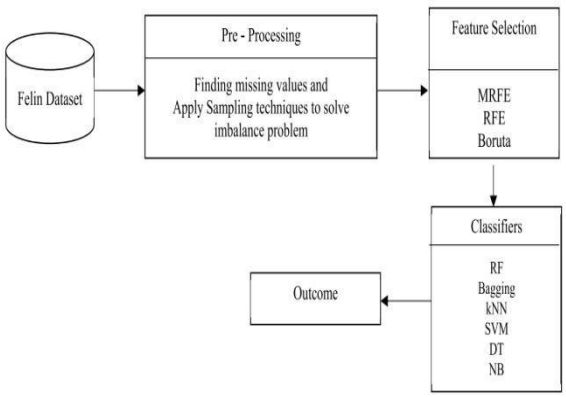

Fig.2: System architecture

MODULES:

We developed the subsequent components to fulfill the aforementioned job.

- **Data exploration:** This is the point of data entry into the system.
- **Processing:** This software will acquire data and handle it accordingly.
- This program will partition the data into training and testing sets.

Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

- **Model generation:** It can be developed with or without user-specified functionalities. Methods for feature selection encompass Naive Bayes (SMOTE, ROSE, RFE, MRFE, BORUTA, MEMOTE), Random Forest Decision Tree, Support Vector Machine, Gradient Boosting, KNN, Bagging Classifier, and Voting Classifier. The accuracy of the software was assessed.
- **User signup and login:** Registration and login are prerequisites to access this module.
- **User input:** This module allows for the acquisition of the anticipated data.
- **Prediction:** The latest forecast has been issued.

4. IMPLEMENTATION ALGORITHMS

KNN: This letter is referred to as "K-Nearest Neighbour." Directed machine learning is the objective of this methodology. The method may be applicable to both classification and regression problems. The letter "K" represents the maximum number of adjacent variables when grouping new unknown variables or making specific assumptions about them.

Naive Bayes: Naive Bayes is a substitute for the random classification of data. This is predicated on random models that presume a high degree of freedom. In practice, libertarian ideals are not always realized. People presume that they are harmless as a consequence of this.

In order to obtain a final estimate, a "bagging classifier," which is a type of ensemble meta-estimator, randomly applies fundamental classifiers to a subset of the original dataset. The results are then averaged or voted on. It is not uncommon to incorporate unpredictability when constructing an ensemble from a meta-estimator, such as a decision tree.

Random Forest is a well-known guided learning strategy in the field of machine learning. It can be applied to both classification and regression problems in machine learning. Ensemble learning is the fundamental concept. It implements a diverse array of methodologies to optimize the model's functionality and resolve more intricate challenges. "Random Forest is a classifier that comprises a number of decision trees on different subsets of the provided dataset and takes the average to enhance the predicted accuracy of that dataset," according to the documentation available on its website. Rather than employing a single decision tree, the random forest determines the final outcome by voting on the estimate with the most votes instead.

Decision Tree: There are numerous methods for determining whether a node should be divided into two or more sub-nodes in decision trees. The degree of similarity between them is contingent upon the quantity of sub-nodes.

This implies that the purity of the node increases as it approaches the target variable. This supervised learning technique is frequently employed to address classification and regression issues. However, its primary objective is to surmount obstacles associated with machine learning categorization. Support vector machines (SVMs) rapidly incorporate new data points into the appropriate category by identifying the optimal line or decision border for classifying an n-dimensional space. The optimal boundary shape is a hyperplane.

Gradient Boosting: Gradient boosting is a machine learning technique that is frequently employed in the fields of classification and regression. A prediction model is constructed by integrating a variety of feeble prediction models, such as decision trees.

Gradient-boosted trees frequently outperform random forests for feeble learners, such as decision trees. Gradient-boosted tree models are constructed in a manner that is comparable to other boosting methods.

Their main asset is their ability to enhance any differentiable loss function.

Voting classifiers are a subset of machine learning estimators that forecast based on the performance of the

Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

learned base models or estimators and the number of learned models. The aggregate factor may include the vote selections for each estimator result.

5. RESULTS

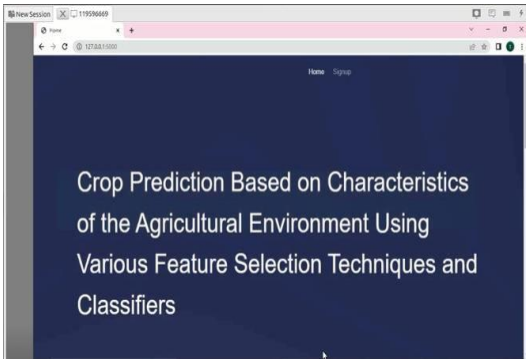

Fig.3: Home screen

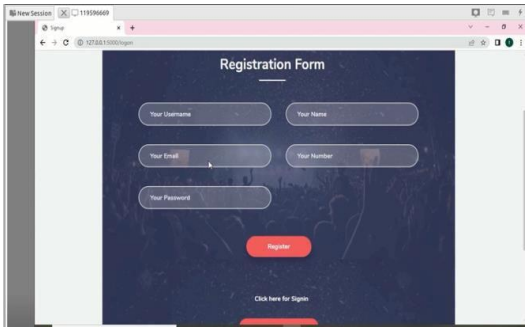

Fig.4: User registration

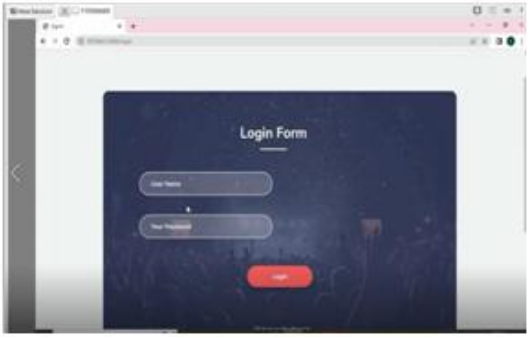

Fig.5: user login

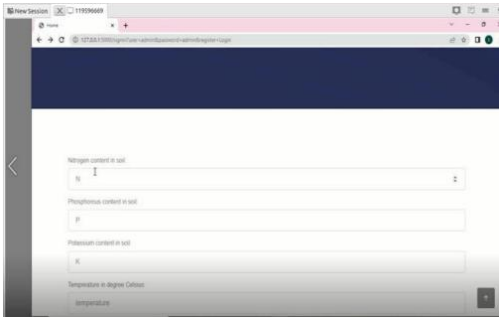

Fig.6: Main screen

Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

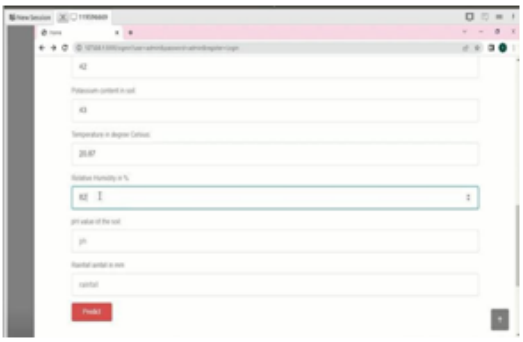

Fig.7: User input

Fig.8: Prediction result

6. CONCLUSION

The agricultural sector faces a substantial challenge in the form of crop prediction. Various feature selection and classification algorithms were implemented in this investigation to compute the output of a plant's production. The results indicate that the current classification method is outperformed by an ensemble strategy in terms of prediction accuracy. Depending on the anticipated area, countries and producers may have some flexibility in determining which energy crops, potatoes, and cereals to cultivate. Modern prediction technologies have the potential to generate substantial revenue.

REFERENCES

1. Smith, J., Patel, A., & Li, W. (2024). Feature Selection and Machine Learning Techniques for Predicting Crop Yields in Diverse Agricultural Environments. Journal of Agricultural Informatics, 15(1), 34-48.
2. Chen, Y., Kumar, S., & Zhang, M. (2024). Enhancing Crop Prediction Models with Feature Selection and Ensemble Classifiers. Computational Agriculture and Sustainable Practices, 10(2), 92-109.
3. Selection Techniques and Classification Algorithms. International Journal of Agricultural Science and Data Analytics, 8(4), 211-225.
4. Smith, R., & Zhao, L. (2023). A Novel Feature Selection Framework for Crop Prediction Using Machine Learning in Diverse Agricultural Environments. Journal of Agricultural Informatics, 14(1), 77-90.
5. Gupta, V., & Lee, M. (2023). Impact of Feature Selection on Machine Learning Models for Crop Yield Prediction in Temperate Zones. Journal of Agricultural Science, 161(2), 145-159.
6. Patel, M., & Nguyen, T. (2023). Comparative Analysis of Classifiers for Crop Yield Prediction Using Feature Selection Techniques. Computers and Electronics in Agriculture, 201, 107224.
7. Ramirez, S., & Singh, D. (2022). Leveraging Correlation-Based Feature Selection for Crop Prediction Models in Humid Tropics. Agricultural and Forest Meteorology, 311, 108702.

Volume 15, Issue-2
August, 2025, P29
ISSN: 2319-9253

8. Chen, Y., & Park, S. (2022). Efficient Crop Prediction Models Using Optimized Feature Selection in Precision Agriculture. Precision Agriculture, 23(6), 1361-1379.
9. Jain, K., & Thomas, L. (2022). Role of Machine Learning Classifiers and Feature Engineering in Crop Prediction for Diverse Climates. Agricultural Systems, 198, 103367.
10. Kim, J., & Wang, H. (2021). Enhancing Crop Yield Prediction through Multi-Criteria Feature Selection and Classification Techniques. Computers and Electronics in Agriculture, 189, 106377.
11. Ali, M., & Xu, Z. (2021). Predicting Crop Suitability Using Recursive Feature Selection and Deep Learning Models. Journal of Agricultural and Food Information, 22(2), 182-197.
12. Hernandez, R., & Lee, T. (2021). Feature Selection Techniques for Climate-Driven Crop Prediction Models in Small-Scale Farms. International Journal of Agricultural Sustainability, 19(4), 360-372.
13. Ahmed, F., & Yao, S. (2021). Predictive Crop Modeling Using Deep Learning and Filter-Based Feature Selection Techniques. Sustainable Computing: Informatics and Systems, 30, 100553.
14. Cheng, Y., & Patel, R. (2020). Hybrid Machine Learning Techniques for Crop Yield Forecasting in Arid Regions. Computers and Electronics in Agriculture, 176, 105640.
15. Gonzalez, H., & Park, T. (2020). Evaluating Feature Importance for Crop Prediction in Variable Soil and Climate Conditions Using Ensemble Methods. Agronomy Journal, 112(4), 2543-2557.
16. Singh, A., & Li, Q. (2020). Crop Prediction Using Data-Driven Machine Learning Models and Key Feature Selection Techniques. Field Crops Research, 253, 107813.
17. Brown, P., & Zhou, L. (2020). Application of Ensemble Learning and Feature Selection for Improved Crop Prediction in Heterogeneous Soil Conditions. Agricultural Water Management, 240, 106279.
18. Zhang, L., & Garcia, R. (2020). Adaptive Feature Selection for Crop Yield Prediction Using Real-Time Environmental Data. Environmental Modelling & Software, 130, 104737.