Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 45 No. 4
April 2024

AgriForecast: A Machine Learning Solution for Crop Yield and Fertilizer Prediction for Developing Countries

Anindita A Khade[1], Avaneesh Karthikeyan Iyer[2]
[1]Assistant Professor, School of Technology Management and Engineering, SVKM's NMIMS Deemed to be University, Navi Mumbai, Maharashtra, India
Email: aninditaac1987@gmail.com; ORCID: https://orcid.org/0000-0003-2616-5092
[2]Student, Department of Computer Engineering, SIES Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra, India. Email: avaneesh.karthik@gmail.com

Abstract
In developing countries like India, agriculture is vital for the livelihoods of its massive population, yet it grapples with inefficiencies and outdated equipment. Bridging the gap between traditional farming practices and modern technological solutions is imperative to boost agricultural productivity. Leveraging advanced machine learning algorithms like Logistic Regression, Support Vector Machine, XGBOOST and Random Forest holds tremendous promise in this regard. These algorithms offer precise forecasts and insights, revolutionizing crop forecasting and yield estimation processes. While farmers traditionally relied on experience for projections, machine learning enables data-driven decision-making, facilitating optimized planting strategies and risk mitigation. Moreover, the adoption of machine learning fosters sustainable practices by enhancing resource allocation and minimizing environmental impact. Ultimately, integrating machine learning into agriculture represents a shift towards smarter and more sustainable farming practices in India. This transition is expected to unlock the agricultural sector's potential, ensuring food security and economic prosperity for farming communities.

***Keywords**: Machine Learning, Prediction, SVM, Logistic Regression, Random Forest,XGBoost*

1. INTRODUCTION

Since the inception of agricultural practices, farming has evolved into a cornerstone of human existence, transcending mere survival to become a pivotal component of global economies. In India, agriculture holds paramount importance, not only for economic prosperity but also for sustaining livelihoods[1]. As the country's economy diversifies, with manufacturing gaining prominence, there's a growing tendency to misuse technology[2].

Presently, scholars are increasingly turning to deep learning, machine learning, and data mining techniques to enhance agricultural productivity and quality. These methods, devoid of explicit programming, enable machines to proficiently learn and improve performance by identifying discernible trends within vast datasets[3]. Utilizing algorithms like Logistic Regression, Support Vector Machine, and Random Forest, researchers aim to forecast crop and fertilizer yields across various regions, considering factors such as topography, soil composition, climate, and nutrients [4].

The objective is to aid farmers in maximizing agricultural output by recommending suitable crop varieties and fertilizers, while accounting for variables like soil health, weather conditions, and resource availability[5]. Sustainable farming practices necessitate efficient resource management, where crop prediction systems can advise on selecting crops that minimize resource consumption while optimizing yields[6].

Given the unpredictable climate patterns, identifying resilient crop varieties becomes crucial for ensuring stability and sustainability in agricultural production. Recommending high-value crops not only enhances farmers' incomes but also contributes to their economic empowerment. Additionally, promoting low-impact crops fosters sustainable farming practices, reducing reliance on pesticides and fertilizers, thus mitigating environmental impact[7].

Leveraging data and analytics provides valuable insights for farmers to make informed decisions about their farming operations, ultimately driving agricultural productivity and sustainability.

The rest of the paper deals with the literature review, followed by the proposed approach, the results and then the conclusions followed by references.

2. RELATED WORK

The survey findings hold potential for enhancing the functionality of crop and fertilizer projections, as well as addressing user concerns and challenges. Moreover, they offer valuable insights for designing future crop-fertilizer prediction systems that cater to user preferences and needs, while also tackling potential constraints.

The authors in [8] introduce the Crop Selection Method (CSM), aimed at optimizing crop selection in agricultural planning to maximize net crop yield ratio, thereby supporting food security and economic growth in agricultural nations. Their approach integrates various machine learning methods like Support Vector Machine (SVM), Decision Tree Learning, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), GBDT, and Regularized Greedy Forest (RGF) to enhance the precision of forecasts and improve the effectiveness of CSM.

The authors in [9] emphasize the use of machine learning techniques to enhance rice crop production prediction in Indian agricultural regions, particularly in Maharashtra. They employ the SMO classifier in WEKA software, analyzing data from Indian government records to assess factors like temperature, precipitation, land area, evapotranspiration, and production. Comparative evaluations with other algorithms like Multilayer Perceptron, Bayes Net, and Naïve Bayes reveal the effectiveness of the SMO classifier.

The researchers in [10] propose a methodology for augmenting crop productivity in Indian agriculture through machine learning. Their approach involves two phases: seasonal weather forecast using Recurrent Neural Network (RNN) Long Short-Term Memory (LSTM) networks, and crop selection using the Random Forest classification system. By focusing on agricultural regions in Telangana and recommending crops like rice, cotton, and maize, the methodology aims to assist farmers in making informed decisions regarding crop selection and cultivation.

The authors in [11] address the challenges faced by Indian farmers in selecting suitable crops based on soil conditions, leading to decreased efficiency. Their solution revolves around smart farming, employing an integrated model using machine learning methods such as K-Nearest Neighbour, Naive Bayes, Random Tree, and CHAID. Through soil-specific data collection for the Madurai district and utilizing ensemble techniques, the methodology aims to improve crop recommendation accuracy in precision agriculture.

The researchers in [12] investigate crop type prediction using sensor data and assess the efficacy of various classification algorithms. They evaluate algorithms like k-means, fuzzy c-means, Bayesian estimation, and Support Vector Machine, proposing a hybrid classifier method that combines multiple layers to accurately categorize crops from hyperspectral images. Their study underscores the importance of algorithm selection for accurate crop classification, crucial for agricultural planning and management, demonstrating the superiority of their proposed classifier over traditional systems.

The literature review suggests that the Crop Selection Model may encounter challenges in accurately predicting influential factors, potentially leading to suboptimal crop selection outcomes due to uncertainties in climate and soil data. Moreover, the machine learning algorithms utilized in the study, particularly the SMO classifier, might not fully capture the intricacies of agricultural dynamics, resulting in varying performance across different meteorological and agronomic variables not considered within the selected timeframe. Reliance on precise data inputs and model assumptions introduces errors and biases, undermining the reliability of the proposed technique. Additionally, managing large and dynamic agricultural datasets in real-time could pose computational challenges, impacting the scalability and efficiency of the projected crop prediction model. The research solely focuses on classification algorithms, overlooking practical implementation and user-friendly integration, and is limited by small datasets, potentially hindering its applicability across diverse agricultural contexts.

3. PROPOSED METHODOLOGY

3.1 System Design: To collect data for the crop prediction and fertilizer recommendation system, we have utilized datasets available from sources like Kaggle and the UCI Machine Learning Repository. The dataset has the attributes such as Soil pH, Temperature, Humidity, Rainfall, and NPK values, which are essential for accurate predictions.

For crop recommendation, we have used the Crop Recommendation Dataset available on Kaggle at https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset.This dataset contains relevant information such as soil characteristics, climate conditions, and recommended crops based on those factors.

Similarly, for fertilizer recommendation, we have utilized the Fertilizer Prediction Dataset from Kaggle https://www.kaggle.com/gdabhishek/fertilizer-prediction. This dataset provides information on soil properties, crop types, and recommended fertilizers based on soil nutrient levels.

By utilizing these datasets for training and testing our crop prediction and fertilizer recommendation models, we ensure that our system is built on accurate and comprehensive data, leading to more reliable predictions and recommendations for farmers[13].

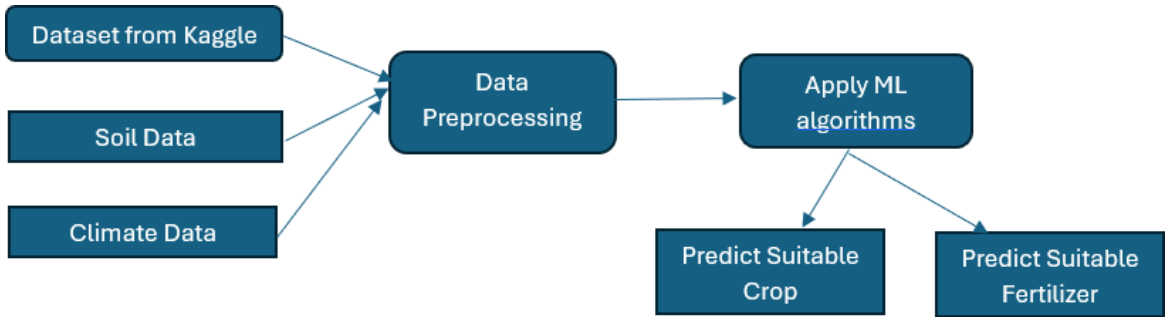

Figure 1: The proposed Model

3.2

Data Preprocessing: Once the datasets are collected from various sources, preprocessing is essential before training the model. This process typically begins with reading the collected datasets and continues with data cleaning. During data cleaning, redundant attributes are identified and removed as they are not relevant for crop prediction. Additionally, any datasets containing missing values need to be addressed by either dropping these values or filling them with appropriate placeholders to ensure better accuracy.

Next, the target variable for the model is defined, which is crucial for training and evaluating the model's performance. Once data cleaning is completed, the dataset is split into training and test sets using the sklearn library. This step ensures that the model is trained on a portion of the data and evaluated on another portion, helping to assess its generalization performance on unseen data.

3.3 Training and Testing strategies: It entails determining the correlation or link between predictor and responder variables. The prediction or estimation is based on a dependent variable from the training dataset. The training dataset contains historical data, including known dependent variables, which are utilized to train the model. The testing dataset, on the other hand, is made up of future data with unknown dependent variables that will be used for scoring or assessment [14]. We use training data to produce predictions or guesses about the test dataset using machine learning training. In our model,

(a)Dataset Splitting:

80% of the dataset is allocated for training, and the remaining 20% for testing to assess model performance.

(b)Feature and Target Label Separation:

Independent variables used for prediction are separated as features, while dependent variables to be forecasted are labeled as the target.

(c)Algorithm Selection:

We use Logistic Regression, SVM, Random Forest and XGBoost algorithm for our predictive analysis.

(d)Model Training:

The dataset is divided into training and testing sets, and models are trained based on the training data.

Data normalization is performed using MinMaxScaler for SVM to ensure consistent scaling across features.

(e)Prediction and Scoring:

Predictions are made using testing data, and model performance is evaluated using standard performance metrics.

3.4 Building the proposed system: Model building is the act of creating a mathematical model to help anticipate or compute future results based on previously obtained information [15]. In our model, the model construction processes are:

(a) Algorithm selection:

Logistic Regression, SVM, Random Forest and XGBoost algorithms have proven results in the domain of crop and fertilizer prediction.

(b) Training Model:

- Feature and target labels are separated.
- To construct training and testing sets, divide the dataset.
- SVM uses MinMaxScaler to normalize data.
- Models are trained based on training data.

(c) Prediction and Scoring:

- Predictions are based on testing data.
- Model performance is evaluated by accuracy scores and classification reports.
- Cross-validation scores are used to determine model robustness.

3.5 Methodologies Used: Each algorithm is configured with appropriate parameters to perform its function in the crop and fertilizer forecast system. SVM is set up to handle nonlinear connections, Logistic Regression for binary classification, and Random Forest for robustness and complexity management. XGBoost is an ensemble learner. These algorithms work together to improve the crop forecast model's accuracy and efficacy.

Logistic Regression

Logistic regression is a statistical and machine learning model commonly used for binary and multi-class classification tasks. Its primary objective is to predict the probability of an observation belonging to a particular class. The logistic function, also known as the sigmoid function, is employed in this model to map input features, resulting in values within the range of 0 to 1, which effectively represent probabilities[16]. This method works by estimating the logarithm of the odds of event occurrence, establishing a linear relationship between input features. During the training process, logistic regression calculates parameters that optimize the likelihood of observing the given data. In binary classification, class membership is determined by setting a decision boundary at a specified threshold. Instances with predicted probabilities above this threshold are assigned to one class, while those below it are assigned to the other class[17].

Logistic Regression is used in the model for its convenience and effectiveness in binary classification issues, estimating the chance of a specific crop being acceptable based on input data[18].

$$f(x)=\frac{1}{1+e^{-x}}$$

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a technique used in supervised machine learning for classification and regression applications. While SVM is often used for classification, it may also be utilized in regression tasks [19]. Its major purpose is to locate an ideal hyperplane in an N-dimensional space. The hyperplane successfully separates data points into discrete classes inside the feature space by maximizing the distance between the nearest points belonging to different categories. The number of features in the dataset has a direct impact on the hyperplane's dimensionality. When there are just two input characteristics, the hyperplane appears as a line; with three input features, it turns into a 2-D plane [20]. When there are more than three characteristics, it becomes difficult to visualize hyperplanes.

The model uses SVM to learn non-linear relationships between crop kinds and characteristics, allowing for accurate classification in complicated datasets.

Random Forest

The Random Forest Algorithm is well-known for its use in machine learning tasks like as classification and regression. It is a widely adopted supervised learning method. Like a forest with many trees for improved resilience, the Random Forest Algorithm's accuracy and problem-solving skills improve as the number of constituent trees grows. A Random Forest classifier averages the outputs of many decision trees on different pieces of data acquired, hence improving the overall predictive accuracy of the dataset [21].The operational process flow of this algorithm develops over the following stages:

1. Randomly picking samples from a given data collection or training set.
2. Create decision trees for each training dataset.
3. Using decision tree mean for voting.
4. The forecast result with the most votes is chosen as the final prediction outcome.

Random Forest was chosen for the model because of its resilience and ability to capture complicated interactions between characteristics and target variables while minimizing overfitting.

XGBoost

Agricultural datasets often involve numerous variables and intricate relationships. XGBoost's ability to handle complex data structures makes it suitable for analyzing diverse factors such as soil properties, weather conditions, and historical crop performance[22]. XGBoost is known for its high predictive accuracy. Its ensemble learning approach, which combines the predictions of multiple weak learners, helps to minimize errors and enhance overall prediction accuracy. This is crucial in crop prediction, where accurate forecasts are essential for decision-making.

3.3Procedure: To predict the specific cropto be grown or a particular fertilizer to be used, we utilize input parameters such as NPK values, weather, moisture, and rain. The crop prediction process begins by loading external crop datasets. Once the dataset is read, various stages of preprocessing are applied, as discussed in the Data Preprocessing section. After preprocessing the data, models are trained using algorithms like Logistic Regression, XGBOOST, SVM and Random Forest classifier on the training dataset.

For crop prediction, factors such as weather, moisture, and predicted rainfall are considered. These parameters can be manually entered or obtained from sensors. The predicted rainfall and input parameter values are then appended to a list for further processing. Table 1 describes the fields of the dataset.

Table 1: A Snapshot of Dataset

Crop	N	P	K	Temperature	Rainfall	Humidity
rice	70	30	40	20.45	81.17	200
chickpea	60	50	50	23.12	82.22	224
mothbeans	20	40	20	23.67	82.32	263
pomegranate	20	40	10	25.91	85.34	265
watermelon	30	20	40	26.76	83.24	256

4. Results and Discussions

4.4.1 EvaluationMetrics:

1. **Precision:**Thisevaluatesaccuracyofpositiveforecasts,measuringproportionofcorrectlyforecastedpositivecasestotheoverallforecastedpositives.

$$Precision = \frac{TP}{TP + FP}$$

2. **Recall:**Theproportionofcorrectlyforecastedpositivestothetotalactualpositivesassessesthesystem's

abilitytocaptureallgenuinepositiveoccurrences.

$$Recall = \frac{TP}{TP + FN}$$

3. F1-score: : It combines recall and precision to account for both false negatives and false positives. It balancesthetwometricsbyprovidingasolitaryvaluere presentingtheharmonicmeanofrecallandprecision.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4. Accuracy: This evaluates total exactness of forecasts through determining the proportion of correctly forecasted cases to total cases, without distinguishing between different classes.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

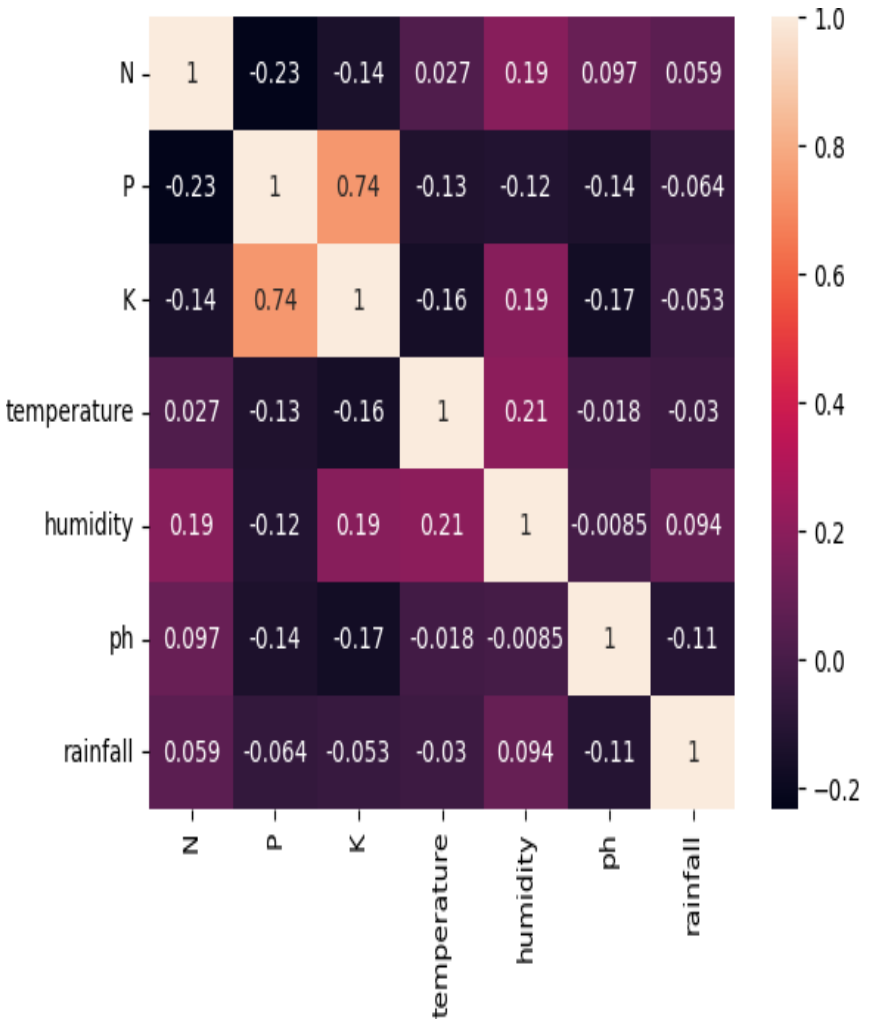

Figure 2:*Heatmap*

F

Figure 2 shows the heatmap specifying the correlations between all the variables.

Table2:*PerformanceMetricsComparison*

Algorithm	Precision	Recall	F1-score
LR	90.9%	91.9%	91.5%
SVM	91.09%	92.90%	91.04%
RF	93.23%	93.45%	94.23%
XGBOOST	99.27%	98.9%	99.04%

Table 2 describes the performance metrics with respect to all the attributes specified.

The above results show that the XGBOOST models outperform all the models, in terms of all the performance metrics.XGBoost incorporates regularization techniques such as L1 and L2 regularization, which help prevent overfitting and improve generalization performance.AlsoXGBoost is based on gradient boosting, which sequentially builds a series of weak learners to create a strong ensemble model. This technique helps improve predictive accuracy by minimizing errors at each stage of the boosting process.

5. Conclusion And Future Work

The study's goal is to achieve optimal crop and fertilizer predictions using data-driven learning help. Several machine learning algorithms are imposed or used for accuracy calculation. A variety of algorithms were used to datasets to get the best results, resulting in the most appropriate crop and fertilizer projections for specific geographies and soil conditions. Using the climatic and subsistence boundaries, this technique will help farmers visualize crop yield. If yield projections are incorrect, the farmer can use this knowledge to decide whether to plant the crop or consider other choices. We created a technique to help farmers discover crops that are most suited to their specific soil conditions. This approach delivers forecasts on the right crop needed for agriculture within the specific region, as well as advise fertilizer important for balancing soil pH levels, integrating expected yield and market price.

As an extension of this study, we want to use a variety of machine learning techniques, such as Neural Networks and various optimization strategies to improve the system's efficiency. To increase the quality of the training data, augment it with new data points or features. Furthermore, robust assessment methodologies, such as using cross-validation techniques to check model performance across many folds of the dataset, can reduce the danger of overfitting and provide more accurate estimates of prediction performance.

REFERENCES

[1] Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020, June). Supervisedmachinelearningapproachforcropyieldpredictioninagriculturesector.In20205thInternationalConferenceonCommunicationandElectronicsSystems(ICCES)(pp.736-741).IEEE.

[2] Govindwar, R., Jawale, S., Kalpande, T., Zade, S., Futane, P., & Williams, I. (2023). Crop and FertilizerRecommendationSystemUsingMachineLearning.InAI,IoT,BigDataandCloudComputingforIndustry 4.0(pp.139-149).Cham:SpringerInternationalPublishing.

[3] Akshatha, K. R., &Shreedhara, K. S. (2018). Implementation of machine learning algorithms for crop rec-ommendationusingprecisionagriculture.InternationalJournalofResearchinEngineering,ScienceandMan-agement(IJRESM),1(6),58-60.

[4] Manoj Kumar, D. P., Malyadri, N., & Srikanth, M. S. (2021). A Machine Learning model for Crop and Fer-tilizer recommendation. NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal— NVEO, 10531-10539.

[5] Suresh, A., Kumar, P. G., &Ramalatha, M. (2018, October). Prediction of major crop yields of Tamilnaduusing K-means and Modified KNN. In 2018 3rd International

conference on communication and electronicssystems(ICCES)(pp.88-93).IEEE.

[6] Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In 2017InternationalConferenceonElectrical,Electronics,Communication,Computer,andOptimizationTechniques(ICEECCOT)(pp.420-423).IEEE.

[7] Reddy, D. A., Dadore, B., &Watekar, A. (2019). Crop recommendation system to maximize crop yield inramtekregionusingmachinelearning.InternationalJournalofScientificResearchinScienceandTechnology,6(1),485-489.

[8] R. Kumar, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to maximize crop yield rate usingmachinelearningtechnique,"2015InternationalConferenceonSmartTechnologiesandManagementforComputing, Communication, Controls, Energy and Materials (ICSTM), 2015, pp. 138-145, doi: 10.1109/ICSTM.2015.7225403.

[9] Pudumalar,S.,Ramanujam,E.,Rajashree,R.H., Kavya,C.,Kiruthika,T.,&Nisha,J.(2017,January).Croprecommendationsystemforprecisionagriculture.In2016EighthInternationalConferenceonAdvancedComputing(ICoAC)(pp.32-36).IEEE.

[10] Mahule, A. A., & Agrawal, A. J. (2020). Hybrid Method for Improving Accuracy of Crop-Type DetectionusingMachineLearning.InternationalJournal,9(2).

[11] Jain, S., & Ramesh, D. (2020, February). Machine Learning convergence for weather based crop selection.In2020IEEEInternationalStudents' ConferenceonElectrical,ElectronicsandComputerScience(SCEECS)(pp.1-6).IEEE.

[12] N.Gandhi,L.J.Armstrong,O.PetkarandA.K.Tripathy,"RicecropyieldpredictioninIndiausingsupportvector machines" 2016 13th International Joint Conference on Computer Science and Software Engineering(JCSSE),KhonKaen,2016,pp.1-5.

[13] Garanayak,M.,Sahu,G.,Mohanty,S.N.,&Jagadev,A.K.(2021).Agriculturalrecommendationsystemforcropsusingdifferentmachinelearningregressionmethods.InternationalJournalofAgriculturalandEnvironmentalInformationSystems(IJAEIS),12(1),1-20.

[14] Chaudhary, K., & Kausar, F. (2020). Prediction of crop yield using machine learning. International JournalofEngineeringAppliedSciencesandTechnology,4(09),153-156.

[15] Doshi, Z., Nadkarni, S., Agrawal, R., & Shah, N. (2018, August). AgroConsultant: intelligent crop recommendationsystemusingmachinelearningalgorithms.In2018FourthInternationalConferenceonComputingCommunicationControlandAutomation(ICCUBEA)(pp.1-6).IEEE.

[16] Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. (2018, August). Crop yield predictionusing data analytics and hybrid approach. In 2018 Fourth International Conference on Computing CommunicationControlandAutomation(ICCUBEA)(pp.1-5).IEEE.

[17] Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019, November). Crop yield prediction using machinelearning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp.125-130).IEEE.

[18] Kumar,A.,Sarkar,S.,&Pradhan,C.(2019,April). Recommendationsystemforcropidentificationandpestcontrol technique in agriculture. In 2019 International Conference on Communication and Signal Processing(ICCSP)(pp.0185-0189).IEEE.

[19] JejurkarSiddhi,S.,Meghna,S.B.,&Wavhal,D.N. (2021).CropPredicationandDiseasesDetectionUsingMachineLearning.

[20] Kamatchi, S. B., & Parvathi, R. (2019). Improvement of crop production using recommender system byweatherforecasts.ProcediaComputerScience,165,724-732.

[21] Kulkarni, N. H., Srinivasan, G. N., Sagar, B. M., & Cauvery, N. K. (2018, December). Improving cropproductivity through a crop recommendation system using ensembling technique. In 2018 3rd InternationalConference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) (pp.114-119).IEEE.

[22] Medar, R., Rajpurohit, V. S., & Shweta, S. (2019, March). Crop yield prediction using machine learningtechniques.In2019IEEE5thinternationalconferenceforconvergenceintechnology(I2CT)(pp.1-5).IEEE.

[23] Bondre, D. A., &Mahagaonkar, S. (2019). Prediction of crop yield and fertilizer recommendation usingmachine learning algorithms. International Journal of Engineering Applied Sciences and Technology, 4(5),371-376.