Resilient Farming with Climate-Based Crop Guidance

Yaswanth Krishna Kumar Pothuri
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
vadlamudi,guntur,AP
221fa04128@gmail.com

Bhargavi Maridu
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
vadlamudi,guntur,AP
bhargaviformal@gmail.com

Abhirama Raju Nadimpalli
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
vadlamudi,guntur,AP
221fa04096@gmail.com

Venkata Naga Sai Kiran Kothuru
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
vadlamudi,guntur,AP
221fa04156@gmail.com

Prasanth Tuta
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
vadlamudi,guntur,AP
221fa04421@gmail.com

Prince Kumar
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
221fa04451@gmail.com

***Abstract*—Agriculture is the main source of income in rural India and has been positively contributing to the nation's GDP. However, the crop produced per hectare in this country is less than the global average, causing misery to farmers, who even take their own lives the marginal farmers are no exception.This research seeks to tackle these challenges by developing a climate-based crop recommendation system powered by machine learning.The system uses critical field inputs, including rainfall, moisture, temperature, nutrient levels (NPK), soil and pH.The system indicates potential crops to help farmers make data-based decisions about using the right crop choice Of the three machine learning algorithms taken- Decision Tree, Random Forest, and Logistic Regression, the highest accuracy was found in Navie Bayes, which reached a peak of 99.54%.Besides this, farm practice also has changed with technological advancement. Mechanization and precision agriculture ensure that crop production yields quality and increases yield. A further development of the model provides a machine learning model that relies upon climatic variability and recorded weather to generate recommendations specific to the location for crops. Intermingling weather patterns with crop performance leads to better yields and sustainable agriculture practices. Future iterations of this could include other variables like soil conditions and market**

***Index Terms*—Data pre-processing, Min-Max scaler,Grid search ,Machine Learning, Logistic Regression, Random Forest, XG-Boost Algorithm,KNN Algorithm, Accuracy.**

I. Introduction

In previous years, the farmers used to predict and choose crops mainly based on their knowledge and experience, while they used to select crops based on what had been grown locally or what was popular in the area. As a result, farmers found it very challenging to address the soil's nutrient supply. On the other hand, contemporary agriculture has brought several solutions for more efficient management of various agricultural inputs due to its high technology inputs. One of the critical challenges is increasing productivity by maintaining quality and keeping the cost reasonable for the consumers. The unreliable effects of climate change also make decision-making tougher as every crop requires particular climatic conditions to become optimal. Therefore, the need for precision farming is floated in this regard. [1]

For example, during the last year in India, the onion price went quite volatile. Next, several farmers planted the crop; some regions, however, had unfavorable conditions, and thus, the crop is spoilt there whereas surplus is produced in other regions. This results in scarcity, and in this scenario, it hurts the middle-class family members as well.This void can be filled by the trustworthy crop recommendation model, which will advise farmers on which crops to plant in relation to the most important soil and environmental parameters, including soil type, moisture content, rainfall, pH, temperature, region, and season. This will maximize crop yields and fertilizer application timing.

Building a model requires proper data preprocessing, which would include handling missing data and the application of Min-Max Scaler to feature normalization. It can be a logistic regression model, a random forest, a KNN model, or even a deep learning approach with XGBoost. In this case,Grid search is a technique for optimizing hyper parameters that increases the model's performance. Combining these methods improves the farmer's real-time model analysis.. The method of evaluation as such ensures proper recommendations by the model to farmers, ensuring them to be financially stable and poor crop choice minimal. The answer will actually help not just one but build up the agriculture sector itself, which is a major contributor to the GDP of India.

II. Literature Survey

Pachade, R. S. [1] further claims that these are now finding increased usage in agricultural yield predictions of crops and pest detection. These include methods like decision trees and neural networks that scan large databases with much higher accuracy than traditional methods. Growth of crops is dependent upon the weather conditions; therefore, weather-based models are developed. While earlier systems were general crop recommendation systems, the current ML-based system considers particular weather, soil, and historical information to provide a relevant recommendation, being dynamic and assisting farmers in making an informed decision.

Shams, M.Y., Gamel, S.A. & Talaat [2] Crop recommendation systems can include recent advancements in using machine learning to help farmers with decision-making. For example, Shams et al. (2024) proposed the XAI-CROP, where the algorithm is also transparent, primarily based on the use of the principles of eXplainable Artificial Intelligence (XAI). Trained on Indian crop data, XAI-CROP utilizes an integrated decision tree model that employs Local Interpretable Model-agnostic Explanations (LIME) to give understandable recommendations. It achieved a higher accuracy with a low MSE of 0.9412 and an R2 value of 0.94152, thus enhancing the transparency in the way AI-driven agricultural decisions are made. This research underscores the role of interpretation in machine learning models.

Mahale et al. [3] have also designed a crop recommendation and forecasting system for Maharashtra state by using machine learning techniques, such as the combination of LSTM models with a novel expectation-maximization approach. The system increases the accuracy of crop predictions with multifactorial approaches in agriculture, including climatic and soil settings. This research highlights the requirements for advanced machine learning methodology in crop yield advancement and effective farmers' decision-making. To learn more, you can read the full article.

III. Methodology

A. Algorithm

An algorithm can be described as a collection of instructions who describe how to finish a task or find a solution. There are various steps that have to be performed. in the right sequence so as to achieve an outcome in particular.

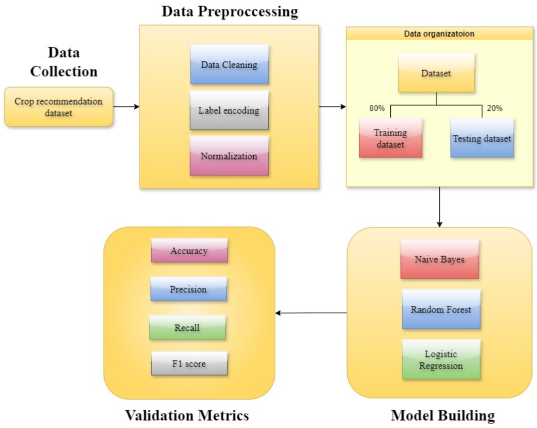


Fig. 1. Proposed model architecture

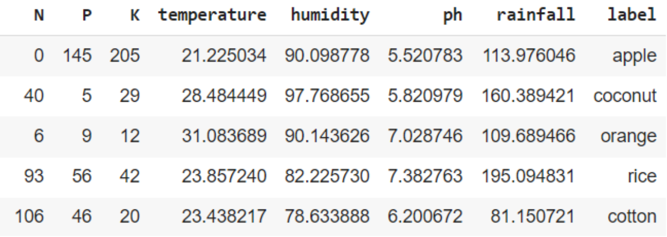
Algorithm 1 Resilient Farming with Climate-Based Crop Guidance process

```
Input: Dataset
$D = \{(u_i, v_i)\}_{i=1}^{n}$, where $u_i$ are features and $v_i \in \{0, 21\}$ is the label
Output: Best-performing model with metrics
Step 1: Pre-processing
- Remove duplicate entries from dataset $D$:
    $D = \{(u_i, v_i) \mid (u_i, v_i) \text{ is unique}\}$    (1)
- Split $D$ into train (80%) and test (20%) sets:
    $D_{\text{train}}, D_{\text{test}} = \text{split}(D, 0.8, 0.2)$    (2)
Step 2: MIN-MAX Normalization
- Apply MIN-MAX to linear transformation that rescales data values to a range of 0 to 1 or -1 to 1:
    $D_{\text{train}}^{\text{balanced}} = \text{MIN-MAX}(D_{\text{train}})$    (3)
Step 3: Apply machine learning models
- Train the following models on $F_k$:
    $M = \{\text{Naive Bayes},$
         $\text{Logistic Regression},$
         $\text{Random Forest},$
         $\text{Gradient Boosting},$
         $\text{Bagging},$
         $\text{K-Nearset Neighbours},$
         $\text{Support vector Machine}\}$
Step 4: Model Evaluation
- For each model $m \in M$,The evaluation criteria that are employed consist of F1-Score, Accuracy, Precision, and Recall.
- Naive Bayes provides the best performance for these metrics
Return Best model with metrics
```

The methodology for developing the climate-based crop recommendation system involves the use of machine learning algorithms trained on climate and soil data without the application of feature selection techniques. The system predicts the best crop for a specific region based on environmental factors.

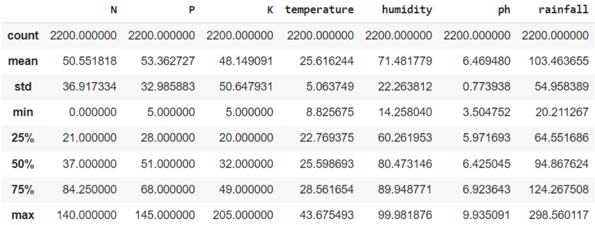
B. Data Collection and Initial Analysis

The dataset consists of important parameters such as:
Nutrient Levels - Nitrogen (N), Phosphorus (P), and Potassium (K).
Climate Variables - Temperature, Rainfall, and Humidity.
Soil Properties - pH levels.
The data was sourced from platforms like the Kaggle Platform, crop recommendation dataset [6] covering approximately 2200 instances for different crops and regions. To comprehend the dataset better, a number of exploratory data analysis (EDA) methods were used. The describe() method was used to generate summary statistics for each feature in the first step. In addition to important metrics like mean, median, and standard deviation, this gave insights into the range of values and made it easier to spot any potential abnormalities or outliers in the data.



N	P	K	temperature	humidity	ph	rainfall	label
0	145	205	21.225034	90.098778	5.520783	113.976046	apple
40	5	29	28.484449	97.768655	5.820979	160.389421	coconut
6	9	12	31.083689	90.143626	7.028746	109.689466	orange
93	56	42	23.857240	82.225730	7.382763	195.094831	rice
106	46	20	23.438217	78.633888	6.200672	81.150721	cotton

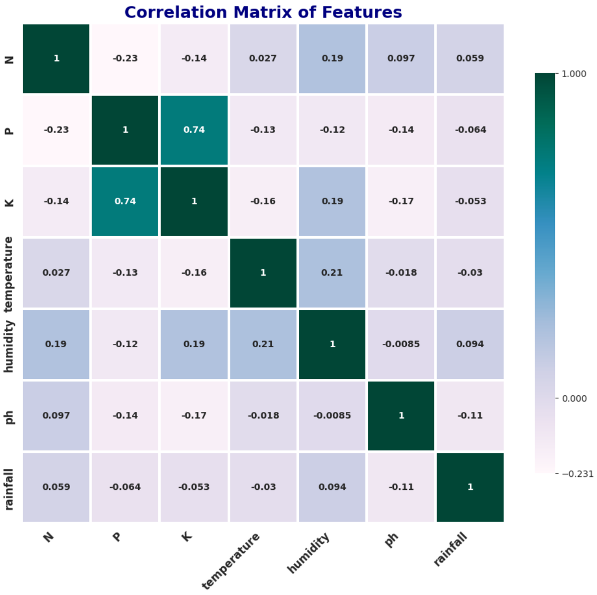
Fig. 2. Crop recommendation dataset



	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117

Fig. 3. Crop recommendation dataset

Subsequently, in order to investigate the connections among various traits, a correlation matrix was constructed. Positive and negative correlations were found with the use of the correlation coefficients, which provided insightful information on feature
In order to assess the connections between various features, the correlation matrix was finally generated. The correlation coefficients could be used to determine both of these.



	N	P	K	temperature	humidity	ph	rainfall
N	1	-0.23	-0.14	0.027	0.19	0.097	0.059
P	-0.23	1	0.74	-0.13	-0.12	-0.14	-0.064
K	-0.14	0.74	1	-0.16	0.19	-0.17	-0.053
temperature	0.027	-0.13	-0.16	1	0.21	-0.018	-0.03
humidity	0.19	-0.12	0.19	0.21	1	-0.0085	0.094
ph	0.097	-0.14	-0.17	-0.018	-0.0085	1	-0.11
rainfall	0.059	-0.064	-0.053	-0.03	0.094	-0.11	1

Fig. 4. Crop recommendation dataset

Ultimately, an analysis was conducted on the target label distribution to evaluate the class balance among various crops. By ensuring that there was no discernible class imbalance, our analysis prevented the model from favoring classes that appear more frequently.

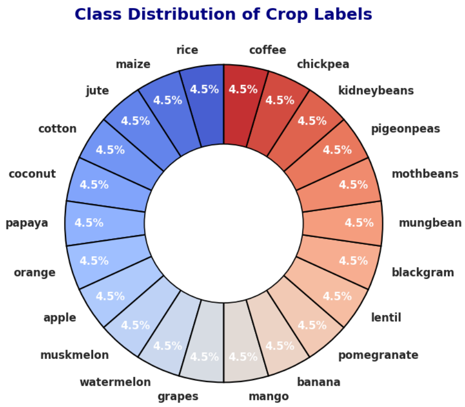


Fig. 5. Crop recommendation dataset

C. Data Preprocessing

Preprocessing was an essential step to make the dataset suitable for machine learning models:

1) Handling Missing Data: Missing values were handled through mean imputation to ensure a complete dataset for training.

2) Feature Scaling: The Min-Max Scaler was applied to normalize the data, scaling all feature values between 0 and

1. This ensures that no feature with larger numerical ranges disproportionately influences the model.

3) Label Encoding: By giving each category a unique number, label encoding transforms categorical data input into a format that machine learning models can understand. The algorithm is able to handle non-numerical data in this fashion. Because categories are handled independently, they do not inherently imply an ordinal relationship between them. In order to enable the model to operate with categorical features while preserving the true meaning of the data for accurate learning and prediction, the labels are converted to numeric values.

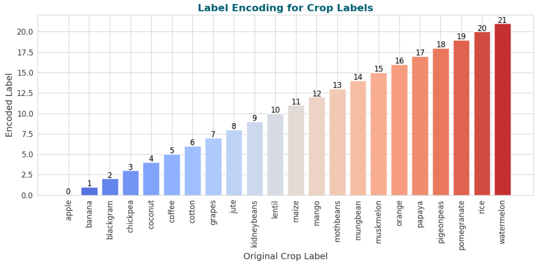


Fig. 6. Label Encoding

D. Model Selection and Training

1) Navie bayes: This method emphasizes the overall probability of the data for a given class rather than identifying what features are most distinguishing between classes.

$$P(Cls|a_1, \cdots, a_n) = \frac{p(cls) \cdot p(a_1|cls) \cdot p(a_2|cls) \cdot \cdot P(a_n|cls)}{P(a_1, a_2, \ldots, a_n)}$$

Where:
- $p(cls)$ is the possibility that a class cls. - $p(a_i|Cls)$ is the feature's feasibility a_i provided cls. - $p(a_1, a_2, \ldots, a_n)$ indicates the marginal probability.

3.4.2 Logistic Regression: In our approach, instances are divided according to the dataset's attributes using Logistic Regression. The logistic function, which is defined as:

$$P(\text{Label}|A) = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 a_1 + \cdots + \gamma_n a_n)}}$$

In this case, the posibility that an instance belongs to a class is represented by $P(\text{Label}|A)$, where the features $a_1, \ldots, a_n$ have coefficients $\gamma_1, \ldots, \gamma_n$ and the intercept is γ_0.

2) Random Forest: Random Forest generates many decision trees and predicts the mode of their classifications for detection purposes. The final prediction can be expressed as:

$$\hat{Z} = \text{mode}(rfc_1(t), rfc_2(t), \ldots, rfc_k(t))$$

where $rfc_i(t)$ represents the predictions from the separate trees and k is the total number of trees. This model is well-suited for detecting fraud because it can efficiently handle high-dimensional datasets and capture detailed relationships between features.

3) Gradient Boosting: Gradient Boosting was the technique we utilized to learn from misclassified examples to enhance the model's capacity for prediction.

The predictive function after the m^{th} iteration can be written as:

$$function_m(x) = function_{m-1}(x) + \eta h_m(x)$$

Where:

- $function_{m-1}(x)$ represents is the previous model's prediction,
- η the learning rate, regulating the input of each weak learner,
- $h_m(x)$ the weak learner established (typically a decision tree) added at iteration m.

In our model, Gradient Boosting iterative refined predictions by adding weak learners, with the learning rate adjusted to balance model performance and stability. This iterative learning process allowed the model to handle complex relationships in our data, making it highly effective in identifying subtle patterns for crop recommendation adjustments based on local climate variability.

4) Bagging: The method of training many models separately using various random subsets of the information is known as "bagging," or bootstrap aggregating. The predictions of these models are subsequently combined by voting or averaging. This can be stated in the following way:

$$func\hat{t}ion(x_val) = \frac{1}{B} \sum_{b=1}^{B} function_b(x)$$

Where: - B is the number of bootstrap samples.
- $function_b(x)$ is the prediction from the b-th model.

5) K-Nearest Neighbour's: KNN, or k-Nearest Neighbor's, is one of the most intuitive and simple non-parametric, instance-based algorithms used both in classification and regression in machine learning. The main idea from which KNN sets off is that those data points, which should be similar to each other, lie close to each other in the feature space. Its objective function can thus be defined as:

$$d(u, v) = \sqrt{\sum(i = 1)^n (u_i - v_i)^2}$$

u$_i$ and v_i represent the values for data points u and v,
where n refers to the total count of elements in the dataset

6) Support Vector Machine: It is used in an N-dimensional space. It looks to find the optimal line or hyperplane that maximizes the distance to each class. Objective function for SVM:

$$\mathbf{w}^T \mathbf{x} + m = 0$$

Where:
- $\mathbf{W}$ is weight vector.
- $\mathbf{X}$ is feature vector.
- m is bias term.

E. Model Evaluation

The performance of our ensemble model is measured according to several essential metrics, which are spelled out in the section that follows:

1) Confusion Matrix: A confusion matrix is essential in machine learning as it evaluates the effectiveness of classification models. It compares the predicted classes with the actual classes, summarizing the model's effectiveness by presenting counts of true positive(TP), true negative(TN), false positive(FP), and false negatives (FN).

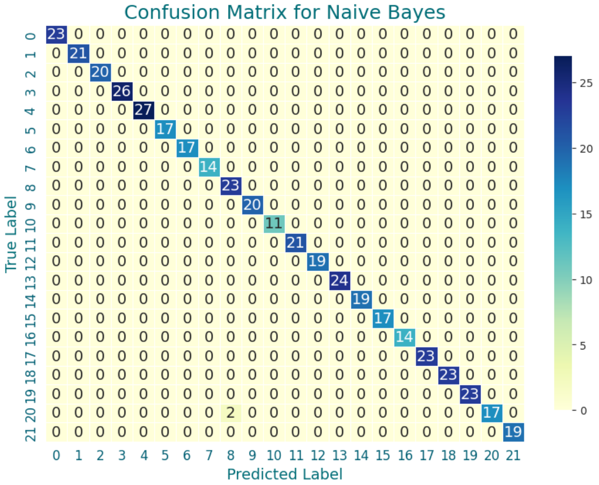


Fig. 7. Confusion matrix

2) Precision: Precision measures how often the correctly predicted positive cases actually are. Precision defines the ratio of true positive cases to the total number of cases that have been classified as such. Precision is formulaic notated as:

$$\text{Precision(U)} = \frac{P}{P+Q} \tag{4}$$

3) Recall: Recall is an evaluation of how good the model is in picking out all the actual positives from the total positive count. It represents how well it can classify frauds. The recall formula is stated to be:

$$\text{Recall (V)} = \frac{P}{Q+R} \tag{5}$$

4) F1-Score: These two parameters are measured in a balanced manner by the F1-score. When there are unequal class distributions, it is quite helpful. The following formula is used to get the F1-score:

$$\text{F\_1 score} = \frac{2 \cdot \text{U} \cdot \text{V}}{\text{U} + \text{V}} \tag{6}$$

5) Accuracy: The effectiveness of a classification model in identifying items is measured by its accuracy. It can be expressed as follows: It is defined as the ratio of successfully predicted instances to all instances in the dataset.

$$\text{Accuracy} = \frac{P+Q}{P+Q+R+S}$$

6) Variable Definitions: -

P = Correctly Predicted Positives (TP)
Q = Incorrectly Predicted Positives (FP)
R = Missed Positives (FN)
S = Correctly Predicted Negatives (TN)

IV. Results and Discussions

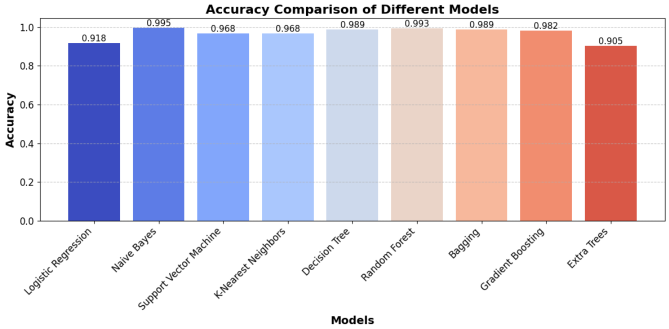


Fig. 8. Metrics performance

Figure 8 displays the performance of the several models used to select crops based on climate. With an accuracy of roughly 99.55% , Naive Bayes performed the best. This demonstrates its high performance in the classification test and how well it works in the resilient agricultural system.

The remaining classifiers that performed well were Random Forest, which attained accuracy of 99.32% . For both bagging and gradient boosting, accuracy levels are approximately 98.64% and 98.18%, respectively.

TABLE I
Metrics of Performance for the Suggested Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.9955	0.9963	0.9955	0.9954
Random Forest	0.9932	0.9937	0.9932	0.9932
Bagging	0.9864	0.9867	0.9864	0.9864
Gradient Boosting	0.9818	0.9843	0.9818	0.9819
k-Nearest Neighbours	0.9682	0.9719	0.9682	0.9682
support vector machine	0.9682	0.9730	0.9682	0.9682
Logistic Regression	0.9182	0.9344	0.9182	0.9172

Unfortunately, neither of the two models—k-Nearest Neighbors, which had an accuracy of 96.72%, and Support Vector Machine, which had the same accuracy of 96.72%—performed quite as well when compared to ensemble approaches. With the lowest accuracy of 91.82% , logistic regression was trailing behind, suggesting that it would be less appropriate for this task

Table II shows the comparison of the suggested and

TABLE II
COMPARISON WITH PREVIOUS MODEL

Metrics	Existing model [1]	Proposed model
Accuracy	0.9932	0.9955
Precision	0.9950	0.9963
Recall	0.9963	0.9955
F1-score	0.9959	0.9954

current models. Both models were excellent; however, the accuracy and precision of the suggested model are improved to around 99.55% and 99.63%, respectively, whereas the accuracy and precision of the existing model are 99.50% and 99.32%, respectively. For both models, the F1-score and Recall values are nearly identical

V. Conclusion

The creation of a crop recommendation system based on machine learning has great potential to increase agricultural productivity by enabling farmers to make well-informed selections depending on particular soil and climate conditions. The usefulness of several machine learning algorithms is demonstrated in this work, with special attention paid to the Random Forest model, which attained an astounding accuracy rate of 99.54%. Such a high degree of accuracy suggests that the model can offer pertinent and trustworthy crop recommendations.

This task provides huge opportunities for increasing agricultural productivity to the extent that help farmers make their informed decisions based on specific soil and climate conditions. So, the research explores several machine learning algorithms focusing on the Random Forest model in detail, and the best accuracy rate reached is 99.54%. Such a great accuracy of the model means appropriate crop recommendations.

Looking ahead, there is potential for further development of this system by incorporating additional factors such as real-time soil health metrics, market dynamics, and pest management strategies. By continuously adapting to new data and technological advancements, the crop recommendation system can significantly contribute to enhancing farmers' decision-making processes, promoting sustainable agricultural practices, and fostering resilience in the agricultural sector as a whole.

VI. References

[1] Pachade, R. S., & Sharma, A. (2022). Machine Learning for Weather-Specific Crop Recommendation. In *International Journal of Health Sciences*, 6(S8), 4527–4537. doi: 10.53730/ijhs.v6nS8.13222.

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

[3] Mahale, Y., Khan, N., Kulkarni, K., et al. (2024). Crop recommendation and forecasting system for Maharashtra using machine learning with LSTM: a novel expectation-maximization technique. Discover Sustainability, 5(134). doi: 10.1007/s43621-024-00292-5.

[4] Gosai, D., Raval, C., Nayak, R., Jayswal, H.,& Patel, A. (2021). Crop recommendation system using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(3), 558-569.

[5] Pande, S. M., Ramesh, P. K., Anmol, A., Aishwarya, B. R., Rohilla, K., & Shaurya, K. (2021). Crop recommender system using machine learning approach. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 1066-1071. doi: 10.1109/ICCMC51019.2021.9418351.

[6] Ingle, A. (2020). Crop Recommendation Dataset. Retrieved from https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset.