Proceedings of the Second International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS-2024)
IEEE Xplore Part Number: CFP24DN7-ART; ISBN: 979-8-3503-6841-3

2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-6841-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSAS64001.2024.10760301

Enhancing Efficiency of Crop Recommendation using Incremental Rank-based Feature Selection Method

Umamaheswari.R, Kannan.E
[1] Research Scholar, [2] Professor
Department of Computer Science and Engineering
Vel Tech Rangarajan Dr.Sagunthala R & D Institute of science and Technology,Chennai
uma2007ap83@gmail.com,ek081966@gmail.com

Abstract

Traditional crop cultivation depends on the farmer's choice, but nowadays, for suggesting suitable crops, a crop recommendation system is required for facilitating optimal crop yield predictions. This study investigates the effectiveness of an innovative approach, named as Incremental rank-based feature selection method, which can improvise the accuracy and efficiency of recommendation systems. Existing feature selection methods often stuck with the high-dimensional dataset as it is more complex. In contrast, proposed method systematically ranks and selects features based on their relevance to crop yield. Through rigorous experimentation and diverse agricultural datasets, it is demonstrated that the incremental approach consistently outperform than traditional methods in terms of predictive accuracy metrics such as Mean absolute error (MAE), Root mean squared error (RMSE), and R-squared (R^2).Findings reveal that the incremental rank-based feature selection technique not only enhances predictive accuracy (MAE = 0.245, RMSE = 0.313, R^2 = 0.878) but also significantly reduces computational complexity. The proposed method exhibits a runtime efficiency of 8.0 seconds, making it suitable for real-time decision-making. By iteratively selecting the most informative features, the method improves model interpretability and generalization capabilities.

Keywords: Crop recommendation system, feature selection, incremental rank-based method, predictive accuracy, computational efficiency.

1. Introduction

The growth of the Indian economy greatly depends on the agriculture sector. Suggesting the right crop enhances agriculture activity [2].So accurate crop recommendation system is essential for optimizing agricultural productivity and sustainability. However, the high dimensionality of agricultural data poses significant challenges in the recommendation process [1]. An incremental rank-based feature selection technique is capable of overcoming the difficulties of existing feature selection techniques. Existing crop recommendation system typically employs machine learning techniques such as decision trees, support vector machines, and artificial neural networks [3, 6]. These systems often rely on traditional feature selection methods, including filter methods like Pearson correlation, wrapper methods like recursive feature elimination, and embedded methods such as LASSO (Least Absolute Shrinkage and Selection Operator). While these approaches have shown varying degrees of success, they often struggle with the scalability and adaptability required to handle the complex and high-dimensional nature of agricultural data effectively.

The proposed incremental rank-based feature selection (IRB-FS) method addresses these limitations by focusing on the ranking of features based on their relevance to crop yield predictions based on crop recommendation. By incrementally selecting features according to their ranks, the methods dynamically adjust to the dataset's

characteristics and assure that only informative features are enclosed in the model. Enclosing the vital features improves model predictive accuracy at the same time; it reduces computational overhead, making it more suitable for large-scale agricultural datasets. There exists a relationship between each feature and target variable. In order to find the relation, in the ranking process, dependency between each feature and the target variable is computed. The incremental selection process iteratively adds features to the model and evaluates the performance of the model at each step using cross-validation. The complications of this step exclusively depend on the number of features and the computational cost of the model, but overall, it remains manageable due to the reduction of feature space and efficient incremental updates. The main objective of this work is to use the IRB-FS method to select the most appropriate features to enhance crop yield through crop recommendation and to evaluate the performance of the proposed IRB-FS using various classifiers and various traditional feature selection methods.

2. Related Work

Feature selection method is used both in classification as well as prediction process in various works [11]. Main indention of selecting features is to reduce number of features, which is done by selecting the important features alone [12]. In literature review part existing methods like filter, wrapper and some hybrid feature selection methods are discussed.

Ghimle et al [14] designed an intelligent crop prediction system is a machine learning-based solution that helps farmers determine the most suitable crops to cultivate on their farms based on soil composition and environmental factors. The system uses the K-Nearest Neighbor (KNN) algorithm to provide accurate predictions, reducing farmers' losses and enhancing production. Rahul et al [6] gives a detailed description about various existing feature selection methods used in crop recommendation systems, by emphasizing the challenges and threats posed by high-dimensional data. To know the strength of feature selection in accurate forecasting of crop yield Dhivya et al [7] mingle the plus point of filter and wrapper methods and developed a hybrid approach and achieved improved predictive efficiency and reduce computational complexity. To carry out the work real-time data's are collected.Gregorutti et al. [4] analysis and evaluated the RFE and non-RFE (NRFE) method to identify, among this two techniques which is better. To rank the features permutation importance (PIMP) measure was used and tested on the Landsat satellite data. From the results, it is found that earlier technique is better than later one. Hall et al. [5] by taking the benchmark dataset studied about various feature selection techniques and revealed that wrapper method is best.Mariammal et al [8] to enhance the crop yield, recommended suitable crop using modified recursive feature elimination method (MRFE) and choose the uttermost prominent features that the crops depends on for their growth and obtained better accuracy.

By merging the theory and rough set approach Qian et al [9] applied mutual information-based feature selection method to remove the redundant features by selecting the candidate features. Shekofteh et al [10] designed a new feature extraction method which comprises of optimization algorithm and fuzzy system to find the subset of features, which in turn minimizes the dimensionality reduction.Bikramet al [12] finds the various kinds of crops that are suitable for a land with the help of neural network by utilising crop yield as an intermediary and predicts the crop suitability. Agronomic variable is taken as input and the model produces scores for each land continuously. The way data driven approach used in this study enhance the accuracy of finding the field suitability for the crop. The reduction of input variables to alleviate computational constraints may introduce bias into the modelling process, leading to potential limitations in the model's performance. Jaison] et al [13] developed an innovative model named as "Adaptive Lemuria " to predict crop that will give maximum yield. The model comprises of Deep Belief Network for pre-training-Means clustering with Optimization used to train data and

Naive bayes for testing. Data set utilized in this work was very small.

Existing crop recommendation systems often employ traditional feature selection methods, which may not adequately handle the complex and high dimensional agricultural data. Also these approaches often face limitations in scalability and adaptability.

3. Methodology

3.1 Dataset Description

Following Indian crop dataset are collected from kaggle website,

- Climatic Conditions: Temperature, rainfall, and other weather-related factors.
- Soil Properties: Soil type, pH level, and nutrient content (nitrogen, phosphorus, potassium).
- Historical Crop Yields: Yield data for different crops across multiple years.
- Management Practices: Types of irrigation, fertilizer usage, and pesticide application.

Key features of the dataset are:

- Region **(**Categorical)**:** Identifies the specific geographic area in India where the data was collected, allowing for analysis of regional differences.
- Year and Month (Integer): Data recorded year and month. It is a temporal data
- Temperature and Rainfall (Integer): Average and total rainfall received in the region during the month.
- Soil Type (Categorical) and pH Level (Float): The type of soil present in the region (e.g., clay, sandy, loamy) and pH level of the soil. It is essential soil properties that influence crop selection and productivity.
- Nutrient Content (NPK) (Categorical): Percentage of nitrogen, potassium and phosphorus present in the soil. Key indicators of soil fertility, directly impacting crop health and yield.
- Crop Type (Categorical) and Crop Yield (Float): Type of crop grown in particular region (e.g., wheat, rice) and yield of the crop. It is the target variable.
- Irrigation type (Categorical) and Fertilizer/Pesticide Usage (Float): The type of irrigation used in the region (e.g., drip, sprinkler) and the amount of pesticide used in the region during the growing season.

Collected data's are pre-processed to remove the irrelevant features**.** Categorical features are encoded using one-hot encoding and continuous features are normalized using min-max normalization to a standard scale to facilitate model training. Pre-processed data is spilt into two parts namely training and testing to maintain the distribution of the target variable (crop yield) across both sets. In this work 80% of the dataset, used for model training and 20% of the dataset is used to evaluate performance of the model.

3.2 Incremental Rank-Based Feature Selection (IRB-FS) Method

IRB-FS method employs an incremental approach for feature selection, wherein features are ranked based on their relevance and selected incrementally. Involves following three steps:

1. Initial Ranking: Rank features based on their correlation with the target variable (crop yield).
2. Incremental Selection: Features are incrementally added to the model, one at a time, based on their ranks. The process continues until addition of features improves model accuracy.
3. Evaluation: Model is evaluated at each step using cross-validation to ensure robustness and prevent over fitting.

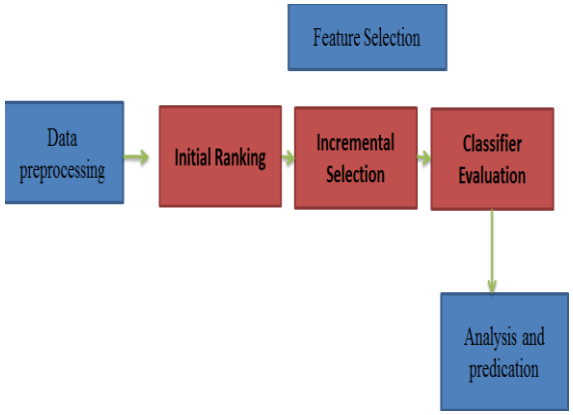

Figure 1: Proposed Architecture diagram

Figure 1 represents proposed architecture diagram. The wok flow begin by gathering a comprehensive dataset that includes various features related to climatic conditions, soil properties, historical crop yields, and other relevant factors and he correlation between each feature and the target variable (crop yield) are computed using pearson correlation coefficient for continuous features and spearman rank correlation for categorical features. Then features are ranked based on their correlation scores. Features with higher correlation values are ranked higher as they are more relevant to the target variable. Later list of features are created by ordering the features based on their relevance scores. This ranking helps to prioritize features. Finally analyse the importance of features and remove those that contribute little to model performance.

Work Flow of IRB-FS

Features are initially ranked based on their correlation with the target variable (crop yield).To compute correlation coefficient between each feature and the target variable: Pearson Correlation is used for continuous features and Spearman Rank Correlation is used for ordinal features. Then features are ranked based on the absolute values of their correlation coefficients in descending order. Features with higher correlation are considered as more relevant. Process starts by initial ranking, which gets started by creating empty set of features (EFSS) ,which contains no features EFSS={∅}.

Followed by this features are added repeatedly based on the rank of the features,

$EFSS_{(k+1)}=EFSS_{(k)}\cup\{\text{Upcoming highest-ranked feature}\}$.

To the empty set of feature set, each time the highest-ranked feature will be added if it not already in EFSS,therefore

$$ESSS_{(k+1)}=ESSS_{(k)}\cup\{X_{(ik)}\}.$$

Update the model to train (T) with selected features $S_{(k+1)}$

$$T_{(k+1)}=\text{train model}(X[S_{(k+1)}],Y)$$

Then to check whether the newly added features increases the accuracy of the model by calculating the score,

$$\Delta\,\text{score}=\text{score}_{(k+1)}-\text{score}_{(k)}$$

Suppose if Δ score exceeds a predefined threshold, retain the feature, else, remove it:

$ESSS_{(k+1)}=\{ESSS_{(k)}\cup\{X_{(ik)}\}$
if (Δ score > threshold then add $ESSS_{(k)}$
else remove feature}

It keeps on continuing, until there is no improvement is found in accuracy when features are added. Later the model is evaluated using cross-validation to ensure robustness and prevent over fitting. Cross validate the model with 5 fold cross validation,

$$\text{Score}(k)=(1/k)_{j=1}\ \text{evaluate}(fk,Xval_j[ESSS_{(k)}],y_{valj})$$

This is done to ensure only features contributing significantly to performance are retained by checking the score value with the threshold value.

Δ score> threshold

//Algorithm of IRB-FS//

Input: crop dataset of multiple features
Output: Feature selection by IRB-FS(X, Y, max_features)

// Initial Ranking//
Rank the features(X, Y)
Get the selected features = [ ]
Find the best score

//Incremental rank based selection //
Length of selected feature > max feature
THEN
Delete feature
ELSE
Select the features.by append (rank)
Train the model with (X [selected features], Y)
Get score to evaluate model
Add features (model, X [selected features], Y)
If (score <= best score) then
Break the loop
Else
Calculate score
Return selected features

//Model Evaluation//
Evaluate model efficiency by K fold (model, X [selected features], Y)
Score less than threshold value
Pop out the feature
Otherwise
Return selected features

4. Performance Metrics

Performance Metrics comparison

Performance of the proposed IRB-FS method is compared with few traditional baseline feature selection models like Filter Method (Pearson correlation),Wrapper Method(Recursive Feature Elimination (RFE), Embedded Method(LASSO (Least Absolute Shrinkage and Selection Operator) and Tree-based feature importance (Random Forest).Performance of the IRB-FS method was evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to measure the accuracy and robustness of crop yield predictions. The results are presented in the table below.

Table1: Performance metrics comparison of Feature selection method

Feature Selection Method	Accuracy	MAE	RMSE	R^2
Filter (Correlation)	78.5%	0.320	0.403	0.812
Wrapper (RFE)	79.2%	0.310	0.393	0.820
Embedded (LASSO)	80.0%	0.305	0.386	0.827
Embedded (Random Forest Importance)	81.0%	0.297	0.380	0.835
IRB-FS (Incremental Rank-Based Feature selection)	85.5%	0.245	0.313	0.878

IRB-FS method ensures that only the most relevant features are selected by continuously evaluating their impact on model performance. Targeted selection reduces noise and enhances model accuracy(table 1), as evidenced by the highest accuracy percentage (85.5%).Also IRB-FS method achieves the lowest MAE (0.245) and RMSE (0.313) values, indicating more precise predictions and highest R^2 value (0.878) shows that this method explains the most variance in crop yield, highlighting its robustness. The incremental nature of the proposed method allows it to adapt dynamically to the dataset, making it more flexible compared to static selection techniques like RFE or LASSO. By adding the feature iteratively, the method ensures continuous improvement in model performance. Figure 2 indicates metrics analysis overview of existing and proposed features selection method.

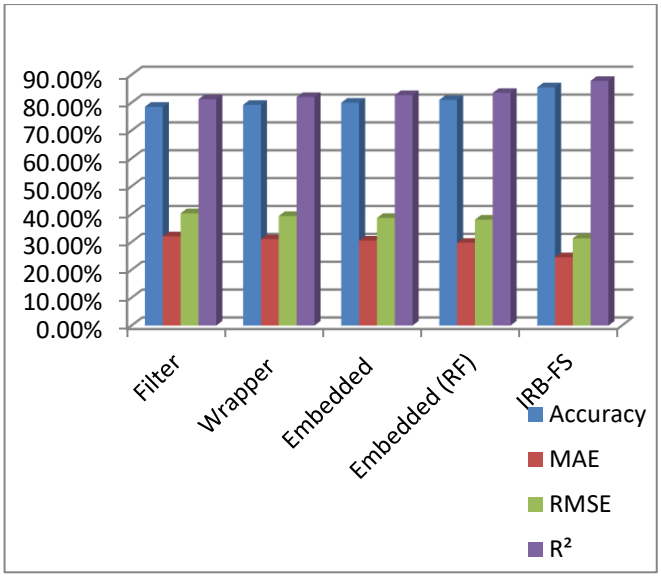

Figure 2: Performance metrics comparison

Table 2: Performance Metrics Comparison of existing algorithms with traditional method FS and IRB-FS method

Algorithm	Feature Selection Method	Acc	MAE	RMSE	R²
Decision Tree	Traditional method	84.3	0.358	0.457	0.782
Random Forest	Traditional method	85.4	0.276	0.348	0.854
Support Vector Machine	Traditional method	80.5	0.312	0.392	0.821
K-Nearest Neighbour	Traditional method	82.4	0.329	0.411	0.809
Decision Tree	IRB-FS	85.4	0.340	0.432	0.791
Random Forest	IRB-FS	85.7	0.245	0.313	0.878
Support Vector Machine	IRB-FS	81.8	0.291	0.370	0.832
K-Nearest Neighbour	IRB-FS	80.6	0.305	0.384	0.823

The method achieves higher performance metrics compared to traditional feature selection method than all existing algorithms. When features are selected using traditional method, performance metrics of the baseline model are low compared to IRB-FS method. At the same time features selected with IRB-FS method shows improved rate Decision Tree(MAE-0.340, RMSE-0.432, and R2-0.791), Random Forest(MAE-0.245, RMSE-0.313, and R2-0.878),Support Vector Machine (MAE-0.291, RMSE-0.370, and R2-0.832), K-Nearest Neighbour (MAE-0.305, RMSE-0.384, R2-0.823) (Table 2).

By choosing the most relevant features, IRB-FS method ensures reduced noise and improved accuracy. By evaluating feature iteratively most informative features are captured that contributes significantly to the model's predictive power. Than Random Forest IRB-FS method shows significant improvement in MAE, RMSE and R² values.

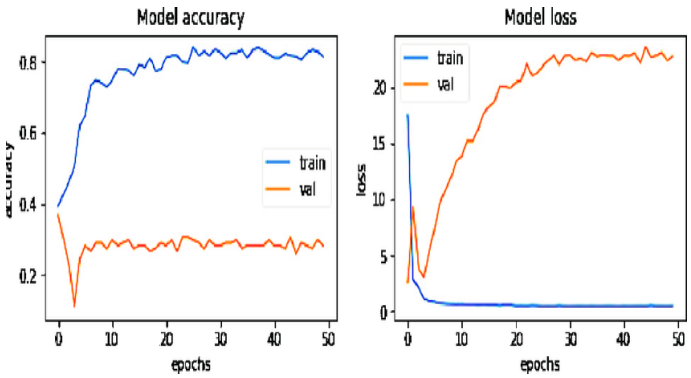

Figure 3: Accuracy and loss graph

Figure 3 represent the accuracy and loss of the model. The accuracy graph illustrates the performance of the model across training epochs. The training accuracy curve shows a consistent increase, indicating that the model is effectively learning from the training data. As expected, the training accuracy

improves with more epochs, reflecting the model's learning process. The loss graph presents the model's loss values over the training epochs. The training loss curve shows a clear downward trajectory, indicating that the model is successfully minimizing the error on the training data. This reduction in loss reflects the model's improved performance in predicting the target values.

Table 3: Runtime Comparison

Feature Selection Method	**Runtime (seconds)**
Filter (Correlation)	2.5
Wrapper (RFE)	15.0
Embedded (LASSO)	10.0
Embedded (Random Forest Importance)	12.0
IRB-FS	8.0

IRB-FS method. has a runtime of 8.0 seconds (Table 3) makes it significantly faster than wrapper methods like RFE (15.0 seconds) and embedded method like LASSO (10.0 seconds) and Random Forest Importance (12.0 seconds).Despite its efficient runtime, IRB-FS method achieves highest accuracy (85.5%) and the best performance metrics (MAE = 0.245, RMSE = 0.313, R^2 = 0.878), demonstrate its effectiveness in identifying the most relevant features while maintaining low computational overhead. So the shorter runtime of the IRB-FS method makes it suitable for practical deployment of its application in real-time crop recommendation systems, where both speed and accuracy are critical. Also this incremental approach is designed to handle large datasets efficiently, making it scalable for high-dimensional data without a substantial increase in computational time.

5. Results and Discussion

Comparison of incremental rank-based feature selection technique with traditional methods has consistently demonstrated its superiority in enhancing crop recommendation systems. Across various datasets and experimental setups, the incremental approach consistently achieves lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Moreover, IRB-FS method consistently outperformed well than traditional feature selection methods by achieving high R-square (R^2) value, which signifies its ability to better explain the variability in crop yield data. These findings highlight the method's effectiveness in identifying and prioritizing the most relevant features critical to crop yield prediction, thus optimizing agricultural productivity and resource allocation.

From the point of practical implications, this method enhances crop recommendation**,** by reducing dimensionality and focusing on relevant features, it provides more accurate and reliable crop recommendations. This is significant for practical applications in agriculture, where precision and efficiency are critical for optimizing crop production. By systematically evaluating and selecting features based on their impact on predictive performance, the method not only enhances model accuracy but also reduces the computational complexity associated with high-dimensional agricultural datasets. This efficiency is crucial for real-time decision-making in agriculture, where timely and accurate recommendations can significantly influence crop management practices, yield outcomes, and resource utilization.

6. Conclusion

Based on the compelling results from comparing the incremental rank-based feature selection technique with traditional methods in crop recommendation systems, it is evident that the incremental approach offers substantial advantages. This shown that the IRB-FS method consistently outperforms traditional feature selection techniques by achieving lower Mean Absolute Error (MAE-0.245), Root Mean Squared Error (RMSE-0.313), and higher R-squared (R^2-0.878) values across diverse agricultural datasets. These outcomes

underscore its efficacy in improving the accuracy and robustness of crop yield predictions. Furthermore, the incremental rank-based feature selection method's ability to efficiently reduce the dimensionality of high-dimensional agricultural datasets enhances its practical utility. By focusing on the most relevant features through iterative selection based on their predictive power, the method not only enhances computational efficiency but also improves model interpretability and generalization. This capability is crucial for stakeholders in agriculture, enabling them to make more informed decisions on crop management strategies and sustainability practices. In conclusion, IRB-FS method represents a significant advancement in the field of crop recommendation systems.

References

[1].Min Li, Mingzhu Lou, Shaobo Deng, Lei Wang,"TRF-WGHC—Top-Ranking filter and wrapper-based greedy hill-climbing gene selection for microarray-based cancer classification",Biomedical Signal Processing and Control, Volume 86, Part C, September 2023, 105309,https://doi.org/10.1016/j.bspc.105309 ,Elsiver.

[2].Arabinda Dash,"Deep feature extraction based cascading model for the classification of Fusarium stalk rot and charcoal rot disease in maize plant", Informatics in Medicine Unlocked, 2023,Elsiver.

[3].Barnali Sahu, Satchidananda Dehuri,Alok Jagadev," A Study on the Relevance of Feature Selection Methods in Microarray Data" The open bioinformatics journal, 31 Jul 2018 ,Volume 11 ,Doi: 10.2174/1875036201811010117.

[4] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," Statist. Comput., vol. 27, no. 3,pp. 659–678, May 2017.

[5].M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," IEEE Trans. Knowl. Data Eng., vol. 15, no. 6, pp. 1437–1447, Nov. 2003.

[6] Rahul Gupta, Priya Sharma,"Feature Selection for Crop Recommendation Systems: A Review",Computers and Electronics in Agriculture, 2023.

[7] Dhivya Elavarasan, Durai Raj Vincent P M , Kathiravan Srinivasan,Chuan-Yu Chang,"A Hybrid CFS Filter and RF-RFE Wrapper-Based Feature Extraction for Enhanced Agricultural Crop Yield Prediction Modeling", Agriculture,doi:10.3390/agriculture10090400, www.mdpi.com/journal/agriculture,2020,Mdpi.

[8] G. Mariammal,A. Suruliandi ,S. P. Raja,E. Poongothai,"Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics Using Modified Recursive Feature Elimination Technique With Various Classifier",IEEE transactions on computational social systems, Vol. 8, NO. 5, October 2021.

[9] Qian, W, Shu W. Mutual information criterion for feature selection from incomplete data. Neurocomputing 2015, 168, 210–220. [CrossRef]

[10] Shekofteh, H, Ramazani, F, Shirani, H. Optimal feature selection for predicting soil CEC: Comparing the hybrid of ant colony organization algorithm and adaptive network-based fuzzy system with multiple linear regression. Geoderma 2017, 298, 27–34. [CrossRef]

[11] Ebrima Jaw,Xueming Wang ,"Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach" Symmetry 2021, 13, 1764.https://doi.org/10.3390/sym13101764 ,https://www.mdpi.com/journal/symmetry.

[12] Bikram Pratim Bhuyan ,Ravi Tomar , T. P. Singh and Amar Ramdane Cherif,"Crop Type Prediction: A Statistical and Machine Learning Approach", Sustainability, https://www.mdpi.com/journal/sustainability, https://doi.org/10.3390/su15010481,2023.

[13] Anjana, AishwaryaKedlaya K, Aysha Sana, BApoorva Bhat, Sharath Kumar, Nagaraj Bhat,"An efficient algorithm for predicting crop using historical data and Pattern matching technique, ,https://doi.org/10.1016/j.gltp.2021.08.060, 2021.

[14] A. M. Ghime1,"Crop Recommendation System Using Machine Learning", International Journal of Advanced Research in Computer and Communication EngineeringVol. 13, Issue 4, April 2024,doi: 10.17148/ijarcce.2024.13476.

[15].Mahmoud Y. Shams, Samah A. Gamel ,Fatma M. Talaat ,"Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making", Volume 36, pages 5695–5714, Neural Computing and Applications 36:5695–5714,https://doi.org/10.1007/s00521-023-09391-2,2024.