ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2025.0514

OPTIMAL FEATURE SELECTION AND CLASSIFICATION IN CROP PREDICTION

P. Nithya, A.M. Kalpana and P. Tharani
Department of Computer Science and Engineering, Government College of Engineering, Salem, India

Abstract

Agriculture is a very important factor in Indian economy. The major problem faced by farmers is that they are not selecting the right crop based on parameters such as soil nutrients, humidity, water level, moisture, and seasonal weather. As a result, they are experiencing a significant loss in productivity. Machine learning algorithms are used in modern farming practices, which examine soil types as well as other factors such as weather and climatic conditions for recommending the most suitable crops. Accurate crop prediction before cultivation helps the farmer to maximize the productivity. In this work, a model is developed to predict suitable crops based on soil nutrients and other environmental factors. A machine learning framework for crop recommendation is presented using Recursive Feature Elimination (RFE) and classification. The XGBoost (Extreme Gradient Boosting) classifier is employed in this proposed system, which provides better results than other methods such as K-Nearest Neighbors, Decision Tree, Naive Bayes and Support Vector Machine. Also, the performance of RFE method is compared with Boruta and Forward Feature Selection (FFS) method. The result shows that the model with RFE based XGBoost classifier achieves high accuracy of 94%.

Keywords:

Machine Learning, Recommendation, Classification, Prediction, Feature Selection

1. INTRODUCTION

Agriculture plays an important role in the economy of our nation. Agriculture provides food and raw materials and provides employment opportunities for a very large percentage of the population. The major issue faced by farmers is the selection of right crops based on parameters such as soil nutrients, humidity, water level, moisture, seasonal weather and yield estimate. Crop selection is influenced by a variety of factors such as climate, soil, markets, government initiatives, and producer preferences.

The crops that are naturally adapted will depend on the topographic aspects of the land, such as elevation, slope and topography, as well as the physical and chemical characteristics of the soil, such as texture, organic matter content, color, pH and fertility level. Since soil nutrients and physical characteristics can have a direct impact on yields, soil quality has a considerable impact on crop output in cultivated fields. The type of plant being grown is directly dependent on the soil type. Each piece of land will have a unique combination of minerals, living things, and inorganic substances, which determines the type of plant which grows successfully. Cultivating the crop that best fits the soil characteristics is an interesting alternative to decrease the need for soil treatment, reducing the costs and potential environmental damages. By choosing a suitable crop for the available soil and environment, yields can be maximized, and irrigation needs can be reduced.

The most significant factor determining whether a crop is suitable for a certain place is the climate. The crop’s potential output is primarily influenced by the climate. The most significant climatic variables that affect the growth, development and yield of crops are solar radiation, temperature and rainfall.

Crop rotation is the practice of growing several crops in succession in the same field to obtain the greatest return with the least amount of investment while maintaining the fertility of the soil. Mono cropping results in collapse of soil nutrients and depletes soil fertility. Market demand is another factor of consideration in crop selection.

Farmers typically choose crops for cultivation based on their traditional knowledge and previous agricultural experience, but natural disasters can make their predictions inaccurate. By implementing emerging technologies, conventional farming can be replaced by precision farming which increases productivity.

As a step towards precision agriculture, a crop recommendation model is developed in this work by considering the factors such as soil, climate and crops which would help the farmers in making decisions on crop cultivation. By cultivating the most effective crops the farmers increase their productivity and competence without wasting any resources with the help of machine learning technology. Choosing the right crops and rotations will help to enhance both economic and environmental sustainability.

The data mining approach in machine learning contributes to predicting the best recommendation through an analysis model. The major problems of Machine Learning are overfitting and the curse of dimensionality. These problems are addressed by performing feature selection. As it results in a small subset of the relevant features from the original one according to certain relevance evaluation criteria, this approach generally improves learning performance, lowers computational costs, and improves interpretability. Finally, a classifier is induced for the prediction phase with the selected features. The efficient framework of feature selection and classification of crop recommendations have been implemented to enhance crop recommendation with high accuracy. Feature selection mainly searches through the subsets of features and attempts to find the best among the candidate subsets based on certain evaluation functions in the agricultural data set. The Fig.1 shows the process of crop recommendation.

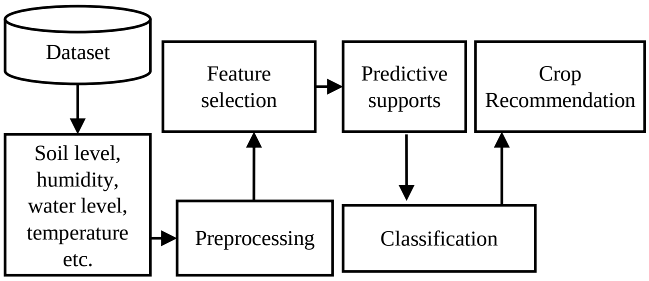

Fig.1. Process of crop recommendation

The proposed model helps farmers in making informed decisions about suitable crops for their farms. It plays an important role in improving the productivity and profitability of the farmers. The important relational features are factors such as crops, average rainfall, humidity, climate, weather, types of soil, soil structure, soil composition, and soil moisture.

The main goal of this work is outlined below:

- Efficient Feature selection is performed using RFE.
- Classification is performed on the selected features using XGBoost classifier.

2. RELATED WORK

The major contributor of India's GDP is agricultural production. Most of the Indian population is dependent on farming or livestock for their regular income [1]. Rather than planting repeated crops, farmers should go for suitable crops according to available environmental conditions. An Agriculture Advisory System is developed, taking soil and environmental parameters as input. It suggests whether the particular crop is suitable moderately, or marginally suitable. The C4.5 decision tree algorithm gave crop-specific suitability levels as output.

A machine learning approach is proposed, which focuses on evaluating crop-specific and potential soil suitability. It is a dynamic model that gives the privilege of fitting the model to any real-time scenario by choosing the real-time dataset [2]. The system is user-friendly and simple to understand. So, small-scale farmers can use it to determine the crop to be cultivated. It will contribute to an increase in yield without degrading the quality of the soil. Crop Selection Method (CSM) is developed, which predicts the yield rate based on various parameters such as weather, soil type, water density, and crop type [3]. It takes crops, planting time, plantation days, and predicted yield rate for the season as input and suggests a classification of suitable crops.

The data collected consist of 46 parameters such as minimum and maximum temperature, humidity, average rainfall, climate, weather and types of land, categories of chemical fertilizer, kinds of soil, soil composition, soil structure, soil consistency, soil texture and soil reaction for applying into the prediction process. Forecasts were made using the Deep Neural Network for crop selection and yield estimation. Majority of Indian farmers are comfortable following the ancestral farming patterns and norms without realizing that crop output is precise, depending heavily on the present-day weather and soil conditions [4]. AgroConsultant is a planned system that will aid Indian farmers in making informed decisions about which crop to plant based on the sowing season, geographical location, soil characteristics, and other environmental factors such as temperature and rainfall.

Farmers are not choosing the appropriate crop based on the nature of the soil, leading to serious losses in productivity. This problem of the farmers has been resolved through precision agriculture. Precision agriculture is a modern technology that uses soil types, soil characteristics and crop yield data collection. It suggests the farmers the suitable crop based on site-specific parameters. It reduces the wrong choice of crop and increases productivity [5]. Accurate crop yield prediction needs an essential understanding of the functional association between yield and the cooperative factors, and revealing such a relationship requires both wide-ranging datasets and powerful algorithms [6]. Feature selection is performed based on the trained DNN (Deep Neural Network) model, which successfully reduced the dimension of the input space without any significant drop in the prediction accuracy.

Nowadays, food production and prediction are inaccurate due to abnormal climatic changes, which will adversely affect the economy of farmers with a poor yield [7]. Machine Learning models such as the Naive Bayes algorithm help predict suitable crops. Soil, geographical and meteorological parameters mainly impact sustained crop production. Most farmers are not aware of the effects of these parameters on crop production. Farmers generally trust their traditional knowledge in selecting crops, which often leads to huge economic loss [8].

A scientific system that focuses on these site-specific parameters when combined with the farmers' traditional knowledge may give an effective solution. A fuzzy logic-based crop recommendation system is developed to support farmers. Different fuzzy rule bases are created to achieve faster parallel processing for individual crops. A rough set on fuzzy approximations identifies the more relevant features to reduce the computation time in the Neural Network [9]. A framework is needed for helping the farmer in making an informed decision about the suitable crop before cultivation [10]. A rough set-based feature selection method is adopted for classification [11].

An efficient filter feature selection algorithm is proposed based on the correlations analysis to select robust features, improving classification accuracy [12]. Machine Learning-based approach is developed to make appropriate crop cultivation choices in Mysore. In farming, crop selection is critical before production [13]. Machine learning techniques upgrade farming from conventional methods to the most cost-effective approach [14]. A hybrid Neuro -Fuzzy and Feature Reduction (NF-FR) model is developed for data analysis. By filtering out the insignificant features, the computational cost of the network will get reduced [15].

Crop forecast is made based on soil and environmental characteristics using feature selection and classification [16]. Crop Recommendation is done using the Random Forest method, and only soil factors, such as NPK (Nitrogen, Phosphorous, and Potassium) levels, are taken into consideration [17]. Crop recommendation is carried out using an ensemble technique that considers solely the soil type. As part of an ensemble technique, Naive Bayes, K-Nearest Neighbor, and random tree are merged [18]. Machine learning algorithms such as SVM and ANN (Artificial Neural Network) can be used to estimate crop yields while taking environmental factors into account [19]. The Random Forest Algorithm is used to create the Crop Recommendation system, which uses location and soil variables as input to forecast crop yield for the chosen crop [20]. The Apriori algorithm and Decision tree induction are used to perform the Crop Selection technique, which considers the factors such as water required, duration, soil type, budget, sowing season, profit and market price [21].

Banerjee, G et al. [22] have developed a fuzzy logic-based crop recommendation system in supporting farmers. Separate fuzzy rule bases were created for individual crops to achieve faster parallel processing. Bhimanpallewar, R.N and Narasingarao, R N [23] discussed machine learning methods for agriculture advisory system enhancement. An Agriculture Advisory System is developed which takes soil and environmental parameters as input. By using the C4.5 decision tree algorithm crop-specific suitability levels were given as output. It suggests whether the crop is suitable, moderately suitable, or marginally suitable. Khaki, S and Wang, L [24] have constructed a deep neural network (DNN) method for the crop yield prediction process, in which the benefit of state-of-the-art modeling and solution methods were taken. Feature selection was performed based on the trained DNN model, which successfully reduced the dimension of the input space without a significant drop in the prediction accuracy.

Doshi, Z et al. [25] have presented an intelligent system known as Agro Consultant, which intends to assist the Indian farmers in making an informed decision about which crop to cultivate depending on the sowing season, geographical location, soil characteristics as well as environmental factors such as temperature and rainfall. Mohan, P and Patil [26] presented an advanced technology called Weighted-Self Organizing Map (W-SOM) for accurate crop and weather prediction, which is the combination of both Self Organizing Map (SOM) and Learning Vector Quantization (LVQ). The prediction accuracy is enhanced by minimizing the Within Class Error (WCE) among the clusters. This approach gives a clear decision about suitable crop cultivation in Mysore. Pudumalar, S et al. [27] have presented a crop recommendation system through an ensemble model with a majority voting technique using Random tree, KNN (K-Nearest Neighbor), CHAID (Chi-square Automatic Interaction Detector), and Naive Bayes towards recommending a crop with high precision and efficiency. Kumar, R et al [28] have developed a Crop Selection Method (CSM) which predicts the yield rate based on various parameters such as weather, soil type, water density, crop type. It takes crops, their planting time, plantation days, and predicted yield rate for the season as input and suggests a classification of suitable crops. Mahabadi, T [29] suggested a method artificial neural network (ANN) which consists of numerous neurons with hidden layer and front/back propagation unit. The author made some modifications in this unit to produce good yield. XGBoost is a highly versatile, portable, and powerful enhanced distributed scaling enhancement library. It integrates machine learning methods through the usage of enhanced scaling. It is a parallel tree boost that provides fast and accurate solutions for a range of science data issues [30].

3. MATERIALS AND PROPOSED METHOD

The main intention of this work is to achieve optimal results in classification. The input data were collected which contains crop data along with soil and environmental factors. Next, preprocessing has been applied to the dataset to obtain clean data suitable for analysis. The Feature selection algorithm is then used to select significant features. By identifying significant features connected to a certain real-world problem, feature selection aids in producing the correct findings. The selected features are fed into XGBoost classifier for performing classification. The performance of XGBoost classifier is compared with the most extensively used algorithms such as K-Nearest Neighbors, Decision Tree, Naive Bayes and Support Vector Machine. The architecture of the proposed method is shown in Fig.2.

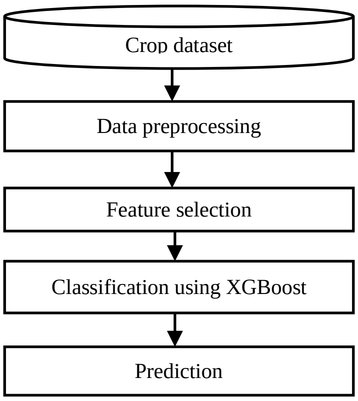

Fig.2. Architecture of the proposed method

3.1 DATA COLLECTION

The standard dataset with the required parameters is not readily available for this research. Hence a dataset is created from scratch, which is one of the significant contributions. Agro professionals, farmers, agricultural departments, government websites, and office records were the sources of information for data collection. The study area was Salem district in the state of Tamil Nadu, India. This district was selected because the agricultural crops like rice, ragi, green gram, black gram, red gram, sorghum, tapioca, ground nut and sesame are the major economic crops in these areas. The required data for this study were collected from various blocks in Salem district. Data related to weather, soil, water level and crops in this work are obtained from randomly selected farming sites in the study area. The data set consists of the following parameters related to climate, crops and soil.

3.1.1 Data Description:

- ***Climate Parameters:*** Precision farming aims to improve productivity by selecting appropriate crops based on climate factors such as temperature, ambient humidity, soil moisture, solar radiance, and rainfall. Evaluation of such parameters helps the farmers to select appropriate climate adapted crops and varieties and plan their agricultural activities.
- ***Site Specific Parameters:*** The soil site suitability evaluation helps in identifying the potential of the soils to produce different crops on a sustainable basis without degrading land. The most suitable crop selection based on site specific parameters not only improves yield but also aids in lowering the unnecessary application of fertilizers, which ultimately reduces soil quality and crop yield. Altitude, Drainage, Erosion and stoniness are the site-specific parameters collected for analysis.
- ***Soil Characteristics:*** The soil and site characteristics are used as parameters for assessing the suitability of land for crop selection. The parameters related to soil characteristics are soil type, soil texture, lime status, moisture retention, bulk density and soil depth

- ***Soil Fertility:*** Proper nutrition is necessary for satisfactory crop growth and production. Each piece of land will have a unique combination of minerals, living things, and inorganic substances, which determines the type of plant which grows successfully. Cultivating the crop that best fits the soil will decrease the need for soil treatment, reducing the costs and potential environmental damage. Parameters related to soil are potential of Hydrogen (pH), Electrical Conductivity (EC), Organic carbon (OC), Nitrogen (N), Phosphorus (P), Potassium (K), Sulphur (S), Zinc (Z), Boron (B), Iron (Fe), Manganese (Mn) and Copper (Cu).
- ***Crop data:*** In terms of nutrients, light, water, temperature, and air, different plant species may have distinct requirements. The plant will not grow correctly if one of these fundamental requirements is not supplied. Selection of the right crop and variety is a very important factor in obtaining maximum profit. Planting several crops in succession on the same piece of land to enhance soil health, maximize nutrient content, and reduce insect and weed. Crop type, Crop predecessor and season are parameters related to crops.

The dataset was created crop-wise in consultation with agriculture experts. Data such as zone name, district name, seasons, soil physicochemical properties, soil fertility, minimum, maximum, optimal temperature, crop name and crop predecessor were collected. An appropriate values/range of values was collected from experts. A total of 1250 instances with 31 attributes for 10 crops have been created. Among the data collected 70% datasets are used for training and 30% for testing respectively.

3.2 PREPROCESSING

In this research work, preprocessing includes various activities such as cleaning and integration, feature extraction, purity check, classification, transformation, and discretization. Data cleaning is the process of filling missing values, removing data noise, and fixing data inconsistencies. Data transformation is the process of normalizing data so that it can be used for ML. Raw data includes incomplete, noisy and inconsistent data. Since the data is collected from different sources and format, quality issues arise when processing with the ML algorithms.

In this work, data from various sources are collected and integrated into a single dataset in excel format. The data cleansing step is performed to remove irrelevant attributes and missing data. The Table.1 shows the variable, and its type used in this research

There are two types of variables in the dataset, depending on the type of value stored, such as numerical and categorical. Numeric variables represent the data in numbers, and nominal variables contain categorical data that correspond to label values instead of numbers. The Table.1 shows variable description and unit measures.

Many ML algorithms cannot handle categorical data directly. Therefore, data encoding is performed to convert categorical data to numeric values. There are many ways to encode data, such as label encoding, one-hot, leave-one-out, probability ratio, encoding, ordinal and M-estimator, etc. Among all these techniques, one-hot encoding and label encoding are used in this proposed research work. The categorical data is mapped to a vector containing 1 representing presence and 0 as the absence of the feature using one-hot encoding method. The number of vectors is mainly based on the number of categories of the feature, for example: the parameter “season” is categorized as kharif, rabi, and summer.

Table.1. Variable description and unit measure

Attributes	Type	Units
Average Rainfall	Numeric	mm
Maximum Temperature	Numeric	°C
Minimum Temperature	Numeric	°C
Optimum temperature	Numeric	°C
Altitude	Numeric	
Drainage	Nominal	-
Stoniness	Nominal	-
Soil Colour	Nominal	-
Soil Texture	Nominal	-
Soil Depth	Numeric	cm
CaCO3	Numeric	-
pH (potential of Hydrogen)	Numeric	-
EC (Electrical Conductivity)	Numeric	dS/M
OC (Organic Carbon)	Numeric	%
N (Nitrogen)	Numeric	Kg/Acer
P (Phosphorus)	Numeric	ppm
K(Potassium)	Numeric	Meq/100g
S (Sulphur)	Numeric	ppm
Z (Zinc)	Numeric	ppm
B (Boron)	Numeric	ppm
Fe (Iron)	Numeric	ppm
Mn (Manganese)	Numeric	ppm
Cu (Copper)	Numeric	ppm
Na(Sodium)	Numeric	ppm
Ca(calcium)	Numeric	ppm
Crop type	Nominal	-
Irrigation Type	Nominal	-
Base Saturation	Numeric	%
CEC	Numeric	cmol/kg
Crop type	Nominal	-
Crop predecessor	Nominal	-

Table.2.One-hot encoding representation of season attribute

Season		
Kharif	**Summer**	**Rabi**
1	0	0
0	1	0
0	0	1

The Table.2 shows the outcome of applying one hot encoding and Table.3 shows the outcome of label encoding methods.

Table.3. Label encoding representation of class label attributes

Label	Value
HR	0
MR	1
ModR	2
NR	3

Data discretization refers to converting continuous values into a range of attribute intervals. Table. 4 shows mapped values for variable Rainfall.

Table.4. Mapping of variable rainfall

	Values Range from	Values range to	Values mapped to
Rainfall (mm)	200	485	1
Rainfall (mm)	485	610	2
Rainfall (mm)	610	745	3
Rainfall (mm)	745	900	4
Rainfall (mm)	900	1200	5

3.3 FEATURE SELECTION

The dataset which contains redundant information may affect the classification performance. Feature selection technique is adopted to reduce the number of input attributes when developing a model. The advantages of feature selection are i) Take less time to train the model, ii) Reduces the complexity of the model, iii) Maximizes the model accuracy. The effectiveness of filter methods depends on performance assessment metrics including distance, consistency, dependency, and knowledge that are directly derived from training data. The wrapper approaches use performance as the evaluation measure and call for a specified learning algorithm. Although more computationally demanding, the wrapper method often outperforms the filter method. Filter and wrapper methods are combined in embedded methods, which also employ their own attribute selection. In order to choose the best attributes from the original dataset, this work focuses on wrapper feature selection method.

3.3.1 Boruta Algorithm:

The Boruta algorithm is a wrapper feature selection method works based on the random forest. It takes little time to analyze and assess how important the features are. The working principle of Boruta algorithm is as follows:

- Add duplicates of each variable to the information system to expand it
- Shamble the added features to remove their correlations with the response.
- Employ a random forest classifier, then collect the calculated Z scores.
- Discover the shadow features with the highest Z score (MZSA), and then give a hit to each feature that scored higher than MZSA.
- Perform a two-sided MZSA equality test for each feature with uncertain relevance.
- Assume the features which have importance significantly lower than MZSA as irrelevant and remove it.
- Consider the features which have considerably higher than MZSA.
- Eliminate all shadow features.
- Repeat the process until all features have been given a level of relevance or until the algorithm has used all the predetermined random forest runs.

3.3.2 Sequential Forward Selection Algorithm (SFSA):

The SFSA algorithm follows a greedy approach. It is a bottom-up search technique that starts with an empty set and increasingly adds features chosen by an evaluation function. SFS operate most effectively with a small optimal subset.

- On the basis of the objective function, the best single characteristic is first chosen.
- The best pair of features is then formed by combining this greatest feature with one of the remaining features.
- The best feature triplet is then chosen by combining these two greatest features with one of the remaining features.
- Repeat the process until a certain number of features have been chosen.

3.3.3 Recursive Feature Elimination (RFE):

RFE is a feature selection technique works well for small samples. Weak attributes are deleted to achieve the necessary qualities till they fit the model. The Gini coefficient is mostly used by the RFE to rank attributes according to their significance. A machine learning model and the precise number of parameters to be used must be provided in the RFE. The procedure of the RFE algorithm is as follows:

```
1. Primarily apply all predictors and train the model on the training set.
2. Determine model performance.
3. Assess the rankings or relevance of the variables.
4. For every subset of feature size $F_i$, $i$=1, 2,.. F do,
     Keep the $F_i$ most essential factors.
     Exploit $F_i$ predictors and train model on training set.
     Determine model performance.
     Update the predictions for each predictor's rankings.
     End
5. Determine the $F_i$'s performance profile.
6. Choose the right number of predictors.
7. Utilize the model that corresponds to the desired $F_i$.
```

3.4 EXTREME GRADIENT BOOSTING CLASSIFIER

XGBoost's fundamental assumption is to learn new features by including a tree structure, fitting the residuals of the final prediction, and calculating the sample score. The sample's final prediction score can be calculated by aggregating the scores of each tree. The formula for predicting scores with K addition functions for n samples with m features is given in the Eq.(1) and Eq.(2):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{1}$$

$$F = \left\{ f(x) = w_{kq}(x)\left(q : \mathrm{R}^m\right) \rightarrow T, w \in \mathrm{R}^T \right\} \tag{2}$$

where F is the regression tree's space, $f(x)$ is one of the regression trees, and $w_{kq}(x)$ represents each -leaf tree's independent structure score.

XGBoost converts the objective function optimization problem into a problem of finding the quadratic function's minimum value and then trains the tree model using the loss function's second derivative information. To avoid the problem of overfitting, the tree complexity is included as a regular term to the objective function. The XGBoost's goal function is given in Eq.(3) as follows:

$$J = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where y_i is the i^{th} target's actual value; $\hat{y}_i$ is the i^{th} target's predicted value; $L(y_i, \hat{y}_i)$ is the difference between y_i and $\hat{y}_i$; n is the sample size; $\Omega(f_k)$ is the tree complexity; The number of sample characteristics is denoted by K.

The objective function's iterative result in time is as follows in Eq.(4):

$$J^{(-t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + f_t(x_i) \right] + \Omega(f_t) + C \tag{4}$$

where $f_t(x_i)$ is the decision tree's complexity in the t^{th} iteration when the variable x_i is calculated; and c is a constant.

If the loss function is expanded to a second order Taylor expansion and the loss function is set to the mean square error, the objective function is given by Eq.(5).

$$J^{(t)} = \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + y_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

where g_i and h_i are the mean square loss function's first and second derivatives respectively.

XGBoost will have less accuracy in the first iteration because the model will be primitive. However, as the number of iterations increase, the model will use the Gradient Descent technique to optimize the loss function. This approach is repeated until the model reaches a point where it can no longer be optimized [22]. As a result, when the number of iterations increases, the model's accuracy is enhanced.

4. RESULT AND DISCUSSION

The proposed feature selection and classification is implemented and tested with numerous ranges of crop datasets. The XGBoost classifier provides more efficient results than other classifiers such as K-Nearest Neighbors, Decision Tree, Naive Bayes and Support Vector Machine. The Table.5 shows the details of parameters considered to test the performance of the proposed systems.

Table.5. Details of parameters in the execution environment

Dataset used	Crop data with soil and environmental factors
Simulation environment	Python framework Libraries
Number of attributes	31
Number of Class	4

The performance of the proposed model is measured based on evaluation metrics like classification accuracy, sensitivity, specificity and F-measure. Accuracy is a measure that denotes how well a model performs across all classes. It is calculated as the ratio between the number of right predictions and the total number of predictions used to compute it.

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + FN + TN} \tag{6}$$

Precision is the calculation for finding the ratio between the actual and positive scores predicted by classification algorithms.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

Recall is a measure of how well it can detect positive examples. The true positive rate (TPR) or sensitivity is another name for it. Sensitivity is used to assess model performance since it shows how many positive examples the model was able to identify correctly.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

F-measure denotes the harmonic mean of the two fractions

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

The Table.6 shows the performance metrics of various algorithms before and after feature selection. Among 31 attributes, the Boruta, FFS and RFE selects 15, 13 and 14 attributes respectively. The selected attributes are given to the classifier to find the most suitable crop based on the soil and environmental factors. Feature selection has improved the performance of various machine learning algorithms. The result shows that the XGBoost classifier with RFE achieves the highest accuracy among the other classifiers for recommending appropriate crops based on the soil and environmental factors.

Table.6. Performance evaluation of various feature selection and classification methods

Feature Method	Classification Method	Attribute Selected		Accuracy		Error Rate	
		Reduction		Reduction		Reduction	
		Before	After	Before	After	Before	After
Boruta	KNN	31	15	0.61	0.62	0.41	0.392
	Naïve Bayes			0.7841	0.7852	0.211	0.212
	Decision Tree			0.8284	0.82913	0.1724	0.1702
	SVM			0.8312	0.8331	0.1582	0.1551
	XGBoost			0.914	0.926	0.0931	0.0915

Forward Feature Selection (FFS)	KNN	31	13	0.63	0.652	0.4	0.3712
	Naïve Bayes			0.7853	0.7981	0.206	0.202
	Decision Tree			0.8293	0.8321	0.1723	0.1711
	SVM			0.843	0.8512	0.1543	0.1421
	XGBoost			0.9142	0.9273	0.0922	0.0814
Recursive Feature Elimination (RFE)	KNN	31	14	0.61	0.682	0.4	0.3671
	Naïve Bayes			0.7845	0.7952	0.202	0.201
	Decision Tree			0.8312	0.8421	0.1705	0.1702
	SVM			0.8492	0.8611	0.1567	0.1498
	XGBoost			0.9314	0.941	0.0928	0.0719

5. CONCLUSION

In this work, most extensively used algorithms, such as K-Nearest Neighbors, Decision Tree, Naive Bayes and Support Vector Machine are evaluated with various feature selection methods. Following the comparison, the XGBoost algorithm with RFE provides highest accuracy of 94 %.The proposed method also achieves higher performance in terms of sensitivity and specificity. The result also shows that the feature selection has improved the performance of various machine learning algorithms. In the future, it is planned to work with crop rotation prediction with an enhanced dataset with a large number of attributes.

REFERENCES

[1] R. Bhimanpallewar and M.R. Narasingarao, "Alternative Approaches of Machine Learning for Agriculture Advisory System", *Proceedings of International Conference on Cloud Computing, Data Science and Engineering*, pp. 27-31, 2020.

[2] R. Bhimanpallewar and M.R. Narasinagrao, "A Machine Learning Approach to Assess Crop Specific Suitability for Small/Marginal Scale Croplands", *International Journal of Applied Engineering Research*, Vol. 12, No. 23, pp. 13966-13973, 2017.

[3] R. Kumar, M.P. Singh, P. Kumar and J.P. Singh, "Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique", *Proceedings of International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials*, pp. 138-145, 2015.

[4] Z. Doshi, S. Nadkarni, R. Agrawal and N. Shah, "AgroConsultant: Intelligent Crop Recommendation System using Machine Learning Algorithms", *Proceedings of International Conference on Computing Communication Control and Automation*, pp. 1-6, 2018.

[5] S. Pudumalar, E. Ramanujam, R. Harine Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop Recommendation System for Precision Agriculture", *Proceedings of International Conference on Advanced Computing*, pp. 32-36, 2017.

[6] S. Khaki and L. Wang, "Crop Yield Prediction using Deep Neural Networks", *Frontiers in Plant Science*, Vol. 10, pp. 1-10, 2019.

[7] M. Kalimuthu, P. Vaishnavi and M. Kishore, "Crop Prediction using Machine Learning", *Proceedings of International Conference on Smart Systems and Inventive Technology*, pp. 926-932, 2020.

[8] G. Banerjee, U. Sarkar and I. Ghosh, "A Fuzzy Logic-based Crop Recommendation System", *Proceedings of International Conference on Frontiers in Computing and Systems*, pp. 57-69, 2020.

[9] A. Anitha and D.P. Acharjya, "Crop Suitability Prediction in Vellore District using Rough Set on Fuzzy Approximation Space and Neural Network", *Neural Computing and Applications*, Vol. 30, pp. 3633-3650, 2018.

[10] L. Bornn and J.V. Zidek, "Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies", *Agricultural and Forest Meteorology*, Vol. 152, pp. 223-232, 2010.

[11] S. Nancy and S. Appavu, "Optimal Feature Selection for Classification using Rough Set-based CGA-NN Classifier", *International Journal of Business Intelligence and Data Mining*, Vol. 11, No. 4, pp. 357-378, 2016.

[12] S. Jiang and L. Wang, "Efficient Feature Selection based on Correlation Measure between Continuous and Discrete Features", *Information Processing Letters*, Vol. 116, No. 2, pp. 203-215, 2016.

[13] N. Jain, A. Kumar, S. Garud, V. Pradhan and P. Kulkarni, "Crop Selection Method based on Various Environmental Factors using Machine Learning", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 8, No. 2, pp. 1530-1533, 2019.

[14] K. Kaur, "Machine Learning: Applications in Indian Agriculture", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, No. 4, pp. 1-6, 2016.

[15] H. Das, B. Naik and H.S. Behera, "A Hybrid Neuro-Fuzzy and Feature Reduction Model for Classification", *Advances in Fuzzy Systems*, Vol. 2020, pp. 1-15, 2020.

[16] A. Suruliandi, G. Mariammal and S.P. Raja, "Crop Prediction based on Soil and Environmental Characteristics using Feature Selection Techniques" ,*Mathematical and Computer Modelling of Dynamical Systems*, Vol. 27, No. 1, pp. 117-140, 2021.

[17] M. Shinde, K. Ekbote, S. Ghorpade, S. Pawar and S. Mone, "Crop Recommendation and Fertilizer Purchase System", *International Journal of Computer Science and Information Technologies*, Vol. 7, No. 2, pp. 665-667, 2016.

[18] T.K. Fegade and B.V. Pawar, "Crop Prediction using Artificial Neural Network and Support Vector Machine", *Data Management, Analytics and Innovation*, pp. 311-324, 2020.

[19] K. Anguraj, B. Thiyaneswaran, G. Megashree, J.G. Preetha Shri, S. Navya, "Crop Recommendation on Analyzing Soil using Machine Learning", *Turkish Journal of Computer and Mathematics Education*, Vol. 12, No. 6, pp. 1784-1791, 2021.

[20] T. Raguntha r, S. Selvakumar and G. Ilamurugan, "Counsel System for Effective Farming using Data Mining Algorithm", *International Journal of Pure and Applied Mathematics*, Vol. 117, No. 21, pp. 921-924, 2017.

[21] D. Mangesh Deshmukh, J. Amitkumar, J. Omkar and S. Rajashree , "Farming Assistance for Soil Fertility Improvement and Crop Prediction using XGBoost", *ITM Web of Conferences*, pp. 1-6, 2022.

[22] A.V. Kumar and S. Sujitha, "Multi-Modal Active Learning with Deep Reinforcement Learning for Target Feature Extraction in Multi-Media Image Processing Applications", *Multimedia Tools and Applications*, Vol. 82, No. 4, pp. 5343-5367, 2023.

[23] R. Shesayar, A. Agarwal and S. Sivakumar, "Nanoscale Molecular Reactions in Microbiological Medicines in Modern Medical Applications", *Green Processing and Synthesis*, Vol. 12, No. 1, pp. 1-15, 2023.

[24] K. Rajput and H. Gurjar, "Multi-Scale Object Detection and Classification using Machine Learning and Image Processing", *Proceedings of International Conference on Data Science and Information System*, pp. 1-6, 2024.

[25] M.D. Sreeramulu and A.S. Mohammed, "AI-Driven Dynamic Workload Balancing for Real-time Applications on Cloud Infrastructure", *Proceedings of International Conference on Contemporary Computing and Informatics*, Vol. 7, pp. 1660-1665, 2024.

[26] P. Mohan and K. Patil, "Weather and Crop Prediction using Modified Self Organizing Map for Mysore Region", *International Journal of Intelligent Engineering and Systems*, Vol. 11, No. 2, pp.192-199, 2018.

[27] S. Dhanasekaran and S.K. Singh, "Utilizing Cloud Computing for Distributed Training of Deep Learning Models", *Proceedings of International Conference on Data Science and Information System*, pp. 1-6, 2024.

[28] M.D. Sreeramulu and A.S. Mohammed, "Efficient Resource Management for Real-time AI Systems in the Cloud using Reinforcement Learning", *Proceedings of International Conference on Contemporary Computing and Informatics*, Vol. 7, pp. 1654-1659, 2024.

[29] T. Mahabadi, "Use of the Intelligent Models to Predict the Rice Potential Production", *International Academic Journal of Innovative Research*, Vol. 2, No. 10, pp. 20-31, 2015.

[30] S. Ramraj, U. Nishant, R. Sunil and S. Banerjee, "Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets", *International Journal of Control Theory and Applications,* Vol. 9, No. 40, pp. 651-662, 2016.

[31] A.S. Mohammed and V. Mallikarjunaradhya, "Optimizing Real-time Task Scheduling in Cloud-based AI Systems using Genetic Algorithms", *Proceedings of International Conference on Contemporary Computing and Informatics*, Vol. 7, pp. 1649-1653, 2024.

[32] P. Singh, B. Jagyasi, N. Rai and S. Gharge, "Decision Tree based Mobile Crowd Sourcing for Agriculture Advisory System", *Annual IEEE India Conference*, pp. 1-6, 2014.

[33] N.J. Pizzi and B. Park, "Spectral Classification using Fuzzy Feature Sampling", *Annual Meeting of the North American Fuzzy Information Processing Society*, pp. 1-6, 2011.

[34] N. Umamaheswari, and R. Renugadevi, "A Subset Feature Selection based DDoS Detection using Cascade Correlation Optimal Neural Network for Improving Network Resources in Virtualized Cloud Environment", *IOP Conference Series: Materials Science and Engineering*, Vol. 993, pp. 1-18, 2020.

[35] C.S. Thirumalai, K. Harsha, M. Deepak and K. Krishna, "Heuristic Prediction of Rainfall using Machine Learning Techniques", *Proceedings of International Conference on Trends in Electronics and Informatics*, pp. 1114-1117, 2017.