Journal of Scientific & Industrial Research
Vol. 82, May 2023, pp. 558-567
DOI: 10.56042/jsir.v82i05.1092

An Artificial Intelligence-based Crop Recommendation System using Machine Learning

Shraban Kumar Apat[1]*, Jyotirmaya Mishra[1], K Srujan Raju[2] & Neelamadhab Padhy[1]

[1]School of Engineering and Technology, Department of Computer Science and Engineering, GIET University, Gunupur 765 022, Odisha, India
[2]CMR Technical campus, Hyderabad 501 401, Telangana, India

Received 30 May 2022; revised 22 September 2022; accepted 15 October 2022

Agriculture is the backbone of the Indian economy and a source of employment for millions of people across the globe. The perennial problem faced by Indian farmers is that they do not select crops based on environmental conditions, resulting in significant productivity losses. This decision support system assists in resolving this issue. In our study, the AI system helps precision agriculture improve overall crop harvest quality and accuracy. This research feature selection, Industry 4.0, proposes one solution, such as a recommendation system, using AI and a family of machine learning algorithms. The data set used in this research work is downloaded from Kaggle, and labeled. It contains a total of 08 features with 07 independent variables, including N, P, K, Temperature, Humidity, pH, and rainfall. Then SMOTE data balancing technique is applied to achieve better results. Additionally, authors used optimization techniques to tune the performance further as smart factories. Cat Boosting (C-Boost) performed the best with an accuracy value of 99.5129, F-measure-0.9916, Precision-0.9918, and Kappa-0.8870. GNB, on the other hand, outperformed ROC-0.9569 and MCC-0.9569 in the classification, regression, and boosting family of machine learning algorithms.

Keywords: AI, Crop harvesting quality, Feature selection, Industry 4.0, SMOTE

Introduction

Indian economy is mainly dependent on agriculture. It is also the primary source of income for the vast majority of Indian Farmers. Agriculture is one of the most important economic sectors for a country's economic growth. Farming also provides a living for most people in a country like India.[1] However, farmers cannot select the best crop for cultivation, forecast market prices, and determine which crop is most suited to the environment and increase productivity. Many new agricultural technologies, such as ML and DL, are being implemented to help farmers grow more efficiently and profitably. In our research, we attempted to recommend the optimum crop and fertilizer for specific farmland. The user can enter soil data and the types of crops they are growing into the crop suggestion program. The application will forecast which crop the farmer should produce, anticipate what the soil lacks or has an abundance of and make recommendations for changes.

Soil is the most important natural resource for growing food, fiber, and firewood. Soil provides the life support system on which civilizations have flourished. It plays an essential role in human livelihoods. Soil serves many functions, such as productive environment, filtration, habitat, sources of raw materials, and ecological and genetic storage. The soil is the main component to provide plants with the necessary nutrients, and soil water in large quantities for crop growth and reproduction in the absence of toxic substances that can stymie crop growth.[2] The quality, cost, and ability to provide the basics that support the ecosystem are declining with deteriorating soil conditions. Therefore, information on soil types, their distribution, size, soil erosion, water installation, etc., is highly necessary for the development of administrative areas such as seed sorting, rain farming, water management, and degraded land reclamation.[3] This information also plays an important role in non-agricultural sectors such as road construction, railways, dams, etc., to ensure sustainable agricultural production and environmental protection. Crops such as wide varieties of rice and other crops result in severe nutrient depletion in the soil. Unbalanced and discriminatory use of chemical fertilizers has resulted in poor soil health. To restore soil health to ensure fertile nature, it is essential to improve the soil's nutritional status and determine the extent of the soil problem.[4] Therefore, soil management is essential. In addition, the

*Author for Correspondence
E-mail: shraban.apat@giet.edu

research study presents the importance of soil fertility management, integrated nutrition management, and Socio economy benefits through a crop recommendation system.

The major contributions of this paper involve applying data balancing techniques on ML Models for Crop recommendations and are studied to show how AI-machine learning classifiers perform for crop recommendations.[5–15] The family of machine learning algorithms techniques was used in the proposed model. The role of AI in crop recommendation systems is a significant phase of Industry 4.0. As a result of AI, agriculture has undergone a revolution. This technology has protected crop yields from various factors, such as unpredictable climate, the exponential growth of population, employment issues, and food security concerns. The primary goal of this paper is to assist farmers in selecting appropriate crops for a specific environment. This will aid in increasing productivity and, as a result, increase profitability.

The Industry 4.0 trend is a transformative force with significant industry implications. The current trend is based on various digital technologies, including Al, the IoT, big data & digital marketing.[13–30] Practices include collaboration, mobility, and open innovation. This article proposes automation and a decision support system. In the future, we will integrate with Industry 4.0, where farmers will benefit from many user-friendly mobile Apps, Chatbots, etc.

Literature Review and Gap Analysis

A study of existing ML techniques used to predict crop production or recommend suitable crops for specific environmental and soil fertility conditionsis discussed in this section. Furthermore, the key benefits of existing techniques and their limitations are described as follows. Abrougui *et al*. identified the organic yield of potato crops by developing an ANN and multiple linear regression (MLR) techniques.[9] This work used soil properties, tillage systems, and soil infiltration resistance to predict potato yield. The simulations were conducted on the alluvial-developed soil in the Agronomy of Chott Meriem (Tunisia) Higher Institute. The parameters such as accuracy, correlation coefficient, and RMSE were used to validate the efficiency of ANN and MLR. The ANN model determined the relationship between soil properties, tillage, and the production of potatoes. However, the ANN had a higher error percentage and minimum description length than MLR because ANN used only two hidden layers. Suchithra & Pai designed a neural network model for classifying pH and indices of soil fertility.[10] In this study, various ensemble learning method (ELM) activation functions are used to achieve better classification accuracy. The input data was taken from the Kerala Government for managing the deficiency of soil nutrients. The simulation results proved that a suitable model was created by optimizing ELM parameters for the index classification of soil fertility. However, this method didn't focus on the significant soil properties of the required crops. Kouadio *et al*. optimized coffee production by validating the soil data's fertility using ELM techniques in Vietnam.[11]

The ELM techniques used set of ten fertility data as predictor variables. Also, they considered the objective variable as coffee yield, where the ELM technique addressed ill-defined problems and complex issues. The parameters such as root mean square error (RMSE), means square error (MSE), Legates and McCabe's index, Willmott's index, and Nash-Sutcliffe efficiency coefficient were used to validate the efficiency of ELM with existing techniques such as random forest and Multiple Linear Regression (MLR). The validated results showed that ELM was more efficient in extracting the features between objective and predictor variables. The system is required to test the ELM method for larger farms with a wide range of conditions. Toseef & Khan proposed an intelligent approach to crop disease diagnosis that can be used on Android mobile devices. It employs a FIS as the primary decision-maker tool at the backend and helps them diagnose agricultural illnesses.[26] It could help agriculture professionals in the public sector diagnose and prevent crop diseases. Its IE uses crop symptoms and a vague input to produce the identified disease as an output. The prediction of the correct disease is up to 99% accurate in this study. This research could be extended to include larger datasets and more local languages. Muangprathub *et al*. used a WSN to develop a system for optimally watering crops.[27] Author used DM techniques to analyze the data to forecast the best temperature, humidity, and soil moisture for crop development in the future. The results revealed that the implementation was beneficial to agriculture. This work made an important contribution by utilizing data mining with association rules to obtain important knowledge to predict the future effects of environmental and climatic conditions. The creation of smart agro-food systems is critical to address global development challenges.[28]

All of the above research focused on a single parameter (either weather or soil) for predicting crop growth appropriateness and did not apply boosting techniques, which we found to be a flaw, and the scope in all of these critically published works is not established. However, we believe each element should be considered simultaneously for the best and most accurate prediction. This is because, while certain soil types may be ideal for supporting one type of crop, the yield will decline if the region's climatic circumstances are not conducive to that crop. The main goal of our study is to design a suitable AI-based decision support system to recommend good crops for the chosen environment, appropriate feature extraction techniques, relevant data balancing methods, and suitable feature processing strategies.

Proposed Methodology

Collection of data from reliable sources, applying pre-processing techniques and performing feature extraction, feature selection, and finally, classification techniques are used in this exploratory study to prognosticate crop recommendation grounded on N, P, K, Temperature, moisture, pH, and rainfall. Artificial intelligence methods are also used in the classification process.[14–16]

Data Collection: In this study, downloaded from Kaggle, the dataset incorporated environmental conditions. The next is to pre-process the data to improve the classifier's accuracy.

Pre-processing: Pre-processing techniques are carried out once the input data is obtained to improve the accuracy of the data collected. Two stages are presented in the pre-processing process, where noises in the input data are removed effectively in stage-I. In stage-II, tokenization and normalization have been carried out. The normalization further includes lemmatization and the stemming processes to complete the pre-processing stage.

Feature Extraction: Extraction of features is a data mining process that includes measures to reduce the amount of available data to explain large amounts of data. One of the major issues when analyzing complex texts is the large number of variables involved.

Feature Selection: This reduces the amount of data, improves classification accuracy, reduces the algorithm's running time, and helps improve the overall quality of the Classification algorithm during the learning process. Unrelated features become overfit, less understandable, and computationally complex, reducing learning accuracy.

Classification: The next and final step is classification, where each classifier includes two important modes: training and testing modes.

To train a custom model, we divided our article into two sections, each with its own set of sub-steps, as shown in Fig. 2 below.

1. Training: This will start with an N, P, and K value and dataset from the repository and train our proposed model with 80% of our trained data set.

2. Testing: Testing data (20%) used after the model is trained.

Phases and Individual Steps

The model is divided into two sections:

Phase#1: Data balancing techniques

Data resampling is one of the most widely recommended techniques for dealing with an imbalanced dataset. When we under-sample, we tend to exclude occurrences from data that may contain important information. This study uses various specific data augmentation oversampling techniques, such as SMOTE. This method of oversampling generates synthetic samples for the minority class. We then classified training (80%) & testing (20%) data sets. The data from the training set was then used for feature extraction, and the training and testing sets were separated for classification. During data pre-processing, our efforts come in the form of a dataset consisting of the rows that must be checked for missing values. The imbalanced nature of the dataset must be verified to determine the number of samples from the minority and majority classes, and the imbalance ratio has been discovered.

The pre-processing is done using the given dataset and sampling techniques (random sampling and oversampling). The findings are assessed after applying current algorithms to the skewed dataset and experimenting with other assessment measures. Deep insight into how the SMOTE algorithm works

Step 1: To Establish the minority class say set A. The closest neighborhood of x for each $ x $ in A, $ is determined by calculating the Euclidean distance between x and all other samples belongs to the same set A.

Step 2: The sampling rate N is determined by the unbalanced component. For each $ x $ in A $, N examples (i.e. x_1, x_2, x_n) are chosen at random from the closest neighbourhoods, resulting in the set $ A 1 $.

Step 3: To generate a new example for each $ x k $ of A 1 $ (k = 1, 2, 3... N), use the following formula. $ X' = x + rand (0, 1) * mid X x k mid $ where rand (0, 1) is a random number.

As shown in Fig. 1, this method is specifically used for data pre-processing, and data balancing (using the SMOTE method) As a result, a balanced data set is produced, which will be used as input for the CRB model, as shown in the Fig. 2 above.

***Phase-II*:** As shown in the above Fig. 2, CRB (Classification-Regression-Boosting) Model

As a result of Phase I, trained and balanced datasets, test datasets are applied and used in the phase-II. The process also includes family of ML algorithms such as linear regression, decision tree, Gaussian Naive Base, Multinomial Naive Bays, and Complementary Naive Bays. Bernoulli naive Baye's, SVM, ridge, RF, and boosting algorithms such as XG boost, CB boost, bagging, stochastic gradient descent, and so on are used.

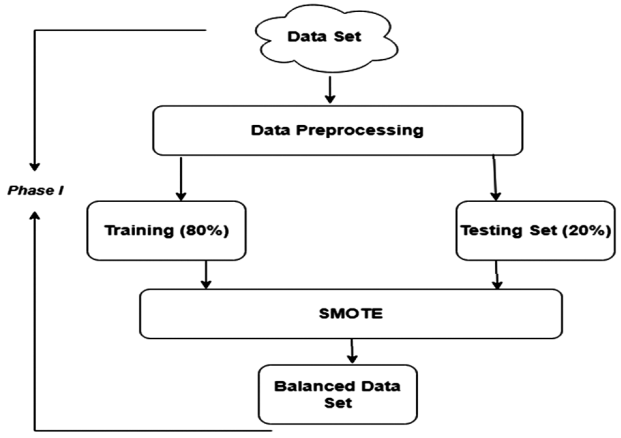

Fig. 1 — Data balancing using SMOTE

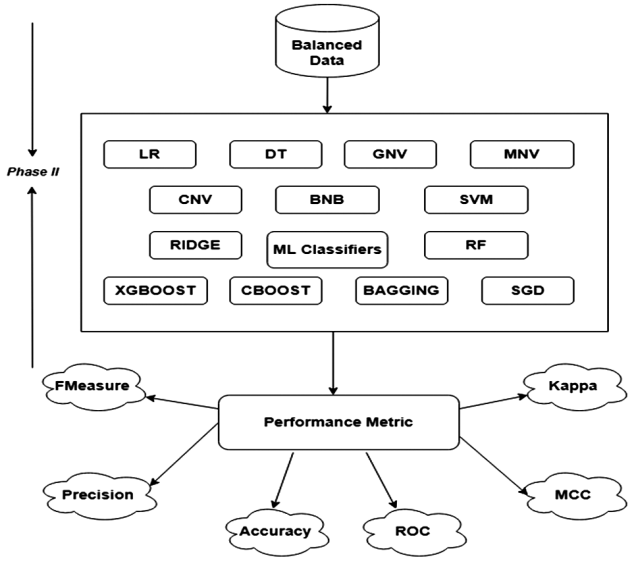

Fig. 2 — CRB Model

Experimental Results and Discussion

To determine the best performer in our study, we used 14 classifiers and evaluated six performance parameters.

Linear Regression (LR)

This connects various environmental variables such as temperature, rainfall, and crop yield. Estimating crop production rates are critical, and farmers will benefit from the outcome of this prediction. It is a statistical method for performing predictive analysis. The linear regression algorithm can represent a linear relationship between a dependent (Y) variable and one or more independent x1, x2, x3 variables.

$$Y = A + Bx + e \quad \text{... (1)}$$

where, e is the incorrect term.

As mentioned in Table 1, the linear regression classifier has an accuracy of 98.56%, F-measure = 98.59%, and precision = 98.74, but its Kappa and MCC remain low at 75.36 and 75.89%, respectively, with ROC at 83.88%.

Decision Tree (DT)

This DT is normally used in the classification and regression processes. In this experiment, we aim to construct a model that will predict the value of a target variable using simple decision rules derived from data features. A tree is an example of a piecewise constant approximation.

$$P(B) = P(A)*P\,(A)/P\,(B) \quad \text{...(2)}$$

The Decision Trees classifier has an accuracy of 98.12% and an F-measure of 97.5%, as shown in Table 1, but its Kappa and MCC are both 78.1% and 76.34%, respectively.

Prediction by the Gauss Naive Bayesian model

The Gauss probability density function is used to calculate the likelihood of a new x value. For predictions, these parameters can be inserted into Gauss PDF using the variable's new input, and Gauss PDF will provide an estimate of the probability of this new input value in this class. Change A to a spam event(y) and B to a communication conforming of a set of words(x_1, x_2.).

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \quad \text{... (3)}$$

$$P(x_1 \dots \dots x_n) = \frac{P(y)P(y)}{P(x_1 \dots\dots x_n)} \quad \text{... (4)}$$

Table 1 — Performance parameters of machine learning classifiers

Name of the classifier	Obtained accuracy (%)	Obtained F-Measure	Obtained Precision	Obtained Kappa	Obtained ROC	Obtained MCC
Linear Regression	98.5692	0.9859	0.9874	0.7536	0.8388	0.7589
DT(Decision Tree)	98.3396	0.9823	0.9832	0.7711	0.8798	0.7724
GNB (Gaussian naïve Bayes)	96.9572	0.9732	0.9810	0.6901	0.9569	0.9569
MNB (Multinomial naïve Bayes)	96.1548	0.9431	0.9366	0.1094	0.5029	0.0384
GB(Gradient Boosting)	96.1469	0.9426	0.9244	0.0000	0.5000	0.0000
BNB (Bernoulli naïve Bayes)	87.8923	0.9117	0.9694	0.3411	0.9232	0.4472
CNB (Complement Naive Bayes)	75.5304	0.8303	0.9614	0.1659	0.8252	0.2811
SVM	98.5620	0.9856	0.9858	0.8070	0.9042	0.8077
RF (Random Forest)	99.1261	0.9913	0.9913	0.8823	0.9426	0.8826
XGBoost	99.0863	0.9909	0.9909	0.8767	0.9404	0.8770
Ridge Regression	97.2750	0.9665	0.9711	0.4679	0.6621	0.5340
Bagging	98.9672	0.9896	0.9868	0.8565	0.9274	0.8591
SGD (Stochastic Gradient Descent)	98.5382	0.9846	0.9848	0.7822	0.8578	0.7833
CBOOST	99.1579	0.9916	0.9918	0.8870	0.9508	0.8858

The probabilities of all events in the (x1, x2...) set can be treated as independent according to the Naive Bayes theorem, so

$$P(x_1, \ldots . x_n) = P(y) \prod_{i=1}^{n} p(y) / P(x_1, \ldots . x_n) \quad \ldots(5)$$

$$P(x_1, \ldots . x_n) \propto P(y) \prod_{i=1}^{n} p(y) \quad \ldots (6)$$

It is purely based on the mean (μ) and Bessel corrected variance (σ) of each word's frequency in the message class.

$$P(y) = \frac{1}{\sqrt{2\pi\sigma 2_y}} \, exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad .. (7)$$

The Naive Bayes algorithm has a low accuracy of 91.12%, and F-measure of 92.31%, but its Kappa and MCC values are better than decision tree methods, which are 93.33% and 90.01%, respectively, as shown in Table 1.

Multinomial Naïve Bayes

This help in calculating the likelihood of an outcome occurring based on previous knowledge of the event's state. If predictor B exists, the probability of class A is calculated as follows:

$$P(\frac{A}{B}) = P(A) * P(\frac{B}{A}) / P(B) \quad \ldots (8)$$

The multinomial classification algorithm performs best with discrete values such as word counts. As a result, we anticipate that it will be the most accurate. Here the probability distribution for each case is computed as follows: Ny is the total number of features of the event belonging to y, Nyi is the count of each feature, and n is a smoothing agent. To eliminate the influence of non-vocabulary words, the Laplace parameter is used.

$$\theta_{yi}^{\vee} = \frac{N_{yi}+\alpha}{N_y+\alpha n} \quad \ldots (9)$$

The Multinomial naïve yes algorithms have an average performance with an accuracy of 95.34 % and F-measure of 93.94%, as shown in Table 1. Its Kappa and MCC values remain low at 49.29 % and 76.34%, respectively.

Complement-Naive Bayes

A closer examination of the equation reveals that the complement of Naive Bayes is the inverse of normal Naive Bayes. In Naive Bayes, the prediction class is the one with the highest value from an expression. As a result, Complement Naive Bayes is the inverse, and the prediction class is the class with the lowest CNB expression value.

$$P(x_1, x_2 \ldots\ldots x_n) = \left(\prod_{j=1}^{j=n} kp\,(c_i)\right) . \frac{p(c_i)}{P(x_1\, x_2 \ldots\ldots\ x_n)} \text{for } 1<i<k. \quad \ldots (10)$$

Here, $P(x_1, x_2, \ldots, x_n)$ is constant for all the classes. We also use the same smoothing parameters and begin working with the actual parameters after calculating the fundamental values.

The CN Bayes algorithms perform poorly with an accuracy of only 75.42% and a measure of 82.04% as shown in Table 1, however, its precision is higher at 95.24%, and its Kappa and MCC values are 49.29% and 76.34% respectively.

Bernoulli Naïve Bayes

This member of the Naive Bayes family accepts only binary values. The most common use is to see if each value matches a word in the document. Bernoulli Naive Baye's can provide better results if word frequency is not prioritized.

The Bernoulli distribution has two mutually exclusive consequences. P (X = 1) equals p, whereas P (X = 0) equals 1p. Bernoulli's theorem can be applied to multiple functions, each expected to be a binary variable or a Boolean value. As a result, in this class, the sample must be represented as a binary feature vector.

Bernoulli's Na Baye's' decision rule is based on

$$X = \{0,1\} \quad \text{... (11)}$$

Then we model P (X|Y) as a Bernoulli distribution. For each class and feature, we must model a Bernoulli distribution, so our terms are as follows:

$$\text{P (X|Y)} = \theta^{X} (1-\theta)^{1-X} \quad \text{... (12)}$$

$$\text{P } (X_j|Y=y_k) - \theta \frac{xj}{kj}(1-\theta_{kj})^{1-x_j} \quad \text{..... (13)}$$

We can estimate KPKP parameters based on this. So, the outcome can only be one of K possible outcomes. This generalization of the Bernoulli distribution is analogous to rolling a die. This is referred to as a categorical probability distribution. The formula must be binary according to the decision rule. Consider the formula in both the cases where, xi = 1 and xi = 0. So I is either the event where, xi = 1 or the event where, xi = 0. The BNB algorithms perform averagely, with accuracy and an F-measure of 89.78% and 92.35%, respectively, as shown in Table 1. However, its precision remains higher at 97.57%, and its Kappa and MCC values are as low as 35.22% and 45.26%, respectively.

Support Vector Machine

The distance between any two lines can be defined as ax+by+c =0 from the given point as (x_0, y_0), where, d is the length of the line.

$$d = \frac{|ax_0+by_0+c|}{\sqrt{a^2+b^2}} \quad \text{... (14)}$$

Similarly, the distance of a given hyper plane for a given value $w^T\emptyset(x) + b = 0$ for the said vector $\emptyset(x_0)$ can be written as:

$$d_H(\Phi(x_0)) = \frac{|w^T+(\emptyset(x_0))+b|}{||w||_2} \quad \text{... (15)}$$

The SVM algorithms perform well with an accuracy of 98.68% and can-measure of 98.32%, as shown in Table 1 and its precision remains at 98.10%. Its Kappa and MCC values are 82.57% and 81.36%, respectively.

Ridge Regression: Ridge regression is used to estimate the coefficients of a multiple regression model when the independent variables are highly correlated. The ridge regression may solve the least square estimator's inaccuracy when the linear regression model has multi-co-linearity. Because the variance and mean squares estimators are frequently smaller than the previously derived least squares estimate, this provides a more accurate estimate of the ridge parameter. The ridge regression estimator is represented as:

$$Y = XB + e \quad \text{... (16)}$$

Here, X and Y represent the Independent and dependable variables, respectively, B represents the regression coefficients to be estimated, and e represents the residual errors. Ridge regression is also called L2 regression because it uses the L2 norm for regularization. In this technique, we minimize the below function w.r.t $'\beta'$ to find the$'\beta'$. To minimize we use the below function:

$$Min_\beta L_{2=}(y - x\beta)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \quad \text{... (17)}$$

As shown in Table 1, the Ridge regression performs moderately, with an accuracy of 92.54% and a measure of 92.56%. Its precision remains at 91.57%, with Kappa R and MCC values remaining as low as 48.79%, 67.12%, and 55.85% respectively.

Random Forest Algorithm

The random forest produces a more accurate prediction by merging multiple decision tree (DT) during training. The final prediction is either the mode of the classes or the mean prediction for regression, which is formed by combining predictions from all trees. Ensemble techniques get their name from making a final decision based on a collection of results.

$$ni_j = w_i C_j - w_{left(j)} C_{left(j)-} w_{right(j)} C_{right(j)} \quad \text{...(18)}$$

The significance of each feature on a decision tree is then calculated as follows:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k\ \epsilon\ all\ nodes} ni_k} \quad \text{... (19)}$$

As shown in Table 1, the RF performs very well, with an accuracy of 99.1%, F-measure 99.25%, precision 99.25%, and ROC value is 96.36%, but Kappa and MCC values remain average at 88.43% & 89.56% respectively.

XGBoost

The formula is as follows. $F = \{f_1, f_2, f_3, f_4 \dots \dots \dots . . f_m\}$ where, $f_1, f_2 \dots f_m$ are the given base learners.

The final prediction can be evaluated using XGBoost as:

Final Prediction = $\widehat{y_i} = \sum_{t=1}^{m} f_t(x_i)$ ……..... (20)

A function that minimizes the overall loss can be defined with the following equation:

$$0 = \{x, x_2, x_3, x_4 \dots \dots \dots . . x_n\} \quad \dots (21)$$

$$\iota^{<t>} = \sum_{t=1}^{n} l(y_i, y_\iota^{<t-1>} \widehat{+ f_t}(x_\iota)) + \Omega(f_t) \quad \dots (22)$$

As mentioned in Table 1, the XGBoost method performs better than other regression and classification algorithms, with an accuracy of 99.07percentt F-measure at 98.09%, precision at 98.09%, and ROC value of −96.36% but Kappa and MCC values remain average at 88.43% & 89.56% respectively.

CBoost

As per the experimented result shown in Table 1, the CBoost Method outperforms all other families of ML algorithms, with an accuracy of 99.51%, F-measure 99.16%, precision at 99.18%, Kappa 89.70%. However, GNV Outperforms CBoost in terms of ROC value and MCC values.

Bagging

It is mainly associated with decision tree methods, but any method can be used. The tagging method is a subset of the model averaging method.

For given training data set D

$$=\{(x_1, y_1), \dots \dots \dots (x_{1,} y_{1)}\}, \quad \dots (23)$$

where, T is a sample dataset with a set of n elements from D that have been replaced as follows

$$D_1, D_{2 \dots} D_T \rightarrow \quad \dots (24)$$

This trains the model on each $D_{i,} i = 1, \dots . . T$ and obtain a sequence of T outputs $f_1(x), \dots \dots \dots f_T(x)$

For classification the equation defined as:

$$f\ddot{(}x) = sign\ (\sum_{i=1}^{T} f_i(x) \quad \dots (25)$$

As per the experimented result shown in the Table 1, the bagging method shows an average result with an accuracy of 97.95%, F-measure = 98.89%, precision at 98.72%, Kappa = 86.54% ROC = 93.64 % and MCC = 86.91%.

SGD (Stochastic Gradient Descent): A "stochastic" system or process has a random probability distribution. As a result, rather than the entire data set, a few samples are chosen at random for each iteration in Stochastic Gradient Descent. A Stochastic Gradient Descent, the formula for normal gradient descent, is:

$$\omega \leftarrow \omega \leftarrow \eta \Delta Q(\omega) \quad \dots (26)$$

where, the error objective is written (with its gradient):

$$Q(\omega) = \frac{1}{n}\sum_i \ Q_i(w) \rightarrow \nabla Q(\omega) = \frac{1}{n}\sum_i \ \nabla Q_i(w) \quad \dots (27)$$

As per the experimented result shown in the Table 1, the SGD Method performs average result with an accuracy of 98.45% -measure 97.89%, precision at 98.58%, Kappa 79.52%, ROC = 85.88 but MCC remains low as 78.43.

GB (Gradient Boosting)

Like other boosting methods, the gradient-boosted trees model is built stage-by-stage and adds the ability to optimize any differentiable loss function.

As per the experimented result, as shown in Table 1, the GB Method shows average results with an accuracy of 96.14%, F-measure 94.26%, the precision of 92.44%, and ROC 50%. However, it offers very poor performance with Kappa and MCC Values. Values that maximise the average performance across all validation sets are chosen. Kappa values range between 1 to –1, with –1 is the min value and max value is 1. The interpretation of Cohen's kappa values, including the kappa result of all classifiers is given in Table 2 .

Similarly, ROC can also be measured. The Receiver Operating Characteristic (ROC) area is a useful tool for visualizing a classifier's performance. It's a graph which relates the classifier's sensitivity. The meaning of ROC values is shown in Table 3. Here classifier results better when value is close to 1.

Table 2 — Representation of Kappa value

Sl No.	Value of Kappa	Measurement parameter
1	If Kappa_value<0.20	Indicated as poor
2	If the Kappa values lies between 0.21 to 0.40	Then indicated As Weak
3	If the Kappa values lies between 0.41 to 0.60	Then indicated as Moderate
4	If the Kappa values lies between 0.61 to 0.80	Then indicated as Good
5	If the Kappa values lies between 0.81 to 1.00	Then indicated as Very Good

Accuracy, F-measure, precision, Kappa, ROC, and MCC performance results for each classifier on a balanced dataset are shown in Table 1.

Comparative Analysis among Different Machine Learning Classifiers

The average value of accuracy, f-measure, precision, Kappa, ROC, and MCC was computed. Among the family of algorithms such as classification, boosting, and regression, boosting algorithms outperformed with an accuracy of 98.3793%, kappa-0.6804, and MCC-0.6750, whereas regression performed with an accuracy of 97.9221%, kappa-0.61075, and MCC-0.6464. However, the f-measure and precision of the regression family performed better, with 0.9912 and 0.9792, respectively, followed by the boosting family (0.97986) and classification (0.9458). Although the accuracy, F-measure, and precision results for most classifiers are excellent, they are insufficient to conclusively demonstrate the benefit of using machine learning classifiers. As a result, we included the Kappa, ROC, and MCC tests that were previously mentioned. The kappa values are shown in the fifth column of Table 1. More than 60% of the classifiers (8 out of 14) achieved kappa values greater than 66%. According to the ROC interpretation shown in Table 1, 93% of the classifiers (26 out of 28) had ROC values greater than 80%, and 50% of this set (among 14 classifiers) had ROC values greater than 90%, indicating that the classifiers had good to excellent behavior. GNB had a higher ROC value of 96%, whereas GB had a lower value of 50%. In the Table's final column, the MCC values for each classifier are displayed.

Table 3 — Threshold value of ROC

Sl No.	Value of Kappa	Measurement parameter
1	If ROC values are between 0.5 to 0.6	Indicated as Fail
2	If ROC values between 0.6 to 0.7	Then indicated as Poor
3	If the ROC values lies between 0.7 to 0.8	Then indicated as Fair
4	If the ROC values lies between 0.8 to 0.9	Then indicated as Good
5	If the ROC values lies between 0.9 to 1	Then indicated as Excellent

A box plot for the accuracy, kappa value, F-measure, precision, ROC, and MCC performance results for each classifier is shown in Fig. 3.

Accuracy (Fig. 3(a)) of different classifiers is plotted and a comparative analysis is done. Based on

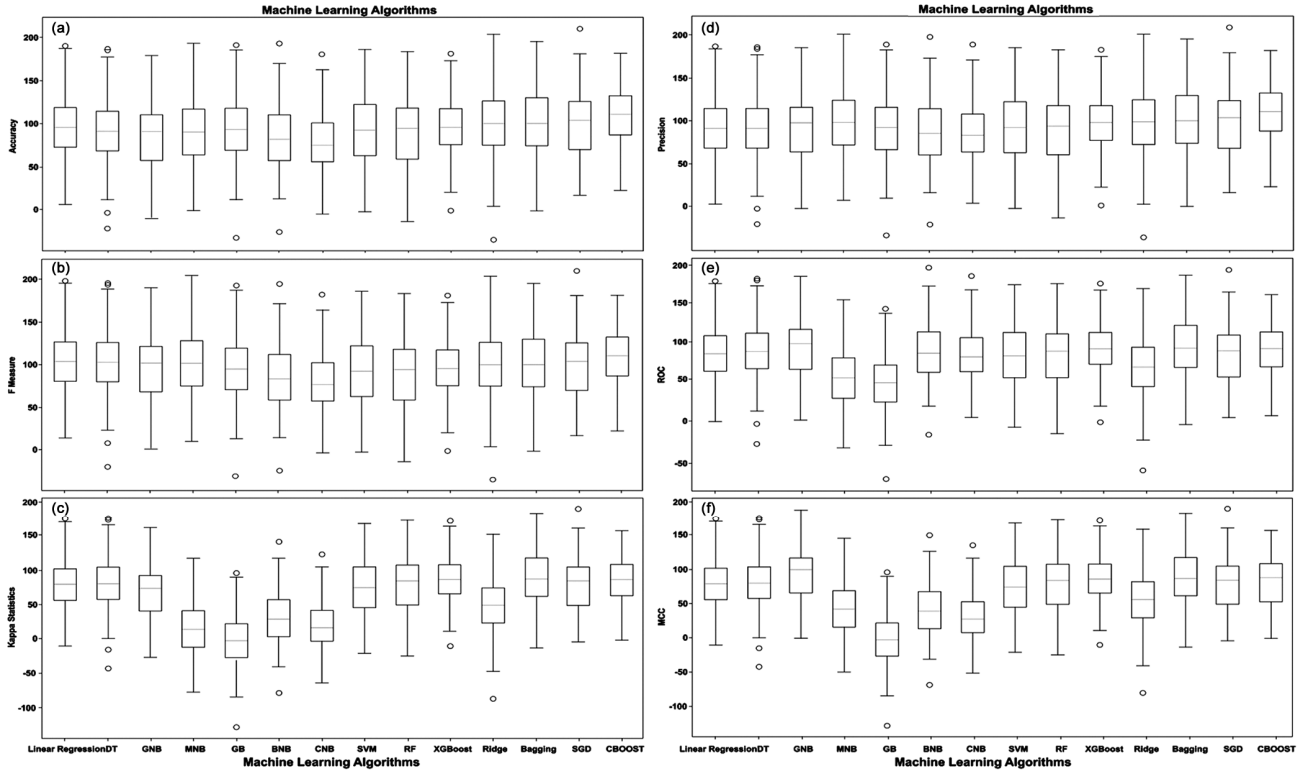

Fig. 3 — (a) Accuracy prediction of different ML, (b) F-measure prediction using different ML, (c) Kappa statistics of different classifiers, (d) Precision of different classifiers, (e) ROC values of different classifiers, (f) MCC of different classifiers

the comparative study, it is observed that CBoost gives the best results of accuracy of 99%. Similarly,the F-measure of different classifiers is plotted and a comparative analysis is done. As shown in Fig. 3(b) the f-score or f-measure is calculated, and it is observed that the F-measure of CBoost algorithms is 0.99.

In Fig. 3(c) kappa statistics of different classifiers are plotted. Based on the comparative study of different statistical parameters of different classifiers used,CBoost is approximately equal to 0.88. In the Fig. 3(d), different classifiers' precision values are plotted, and a comparative analysis is done by evaluating the true positive and true negative rate of different algorithms used for crop prediction. The precision value obtained is equal to 0.99.

The Receiver of Characteristics of different classifiers is plotted as box plot in Fig. 3(e). Based on the comparative study, it is observed that GNV has highest ROC value of 0.9569 against 0.9508 of CBoost which is the maximum crop prediction value. As shown in Fig. 3(f), the MCC value of different classifiers is plotted and a comparative analysis is done. Based on the comparative study, the Mathews Correlation Coefficient value of GNV is 0.9569 which outperformed CBoost which has 0.8858, is evaluated from true positive, false positive, true negative, true negative, and false-negative rates.

Conclusions

Crop yield forecasting is a critical issue in agriculture. A two-fold model, i.e. data balancing & classification model, which comprises a family of machine learning algorithms over the balanced data set were used taking natural factors such N, P, K, Temperature, Humidity, pH and rainfall, etc. into account. This study validates the potential efficacy of incorporating ML algorithms into a decision-making system that implements an intelligent crop recommendation system for recommending appropriate crops. Among the 14 classifiers and 6 performance metrics employed by the ML algorithm family, Boosting (Cboost) produces the best results with an accuracy value of 99.15, F-measure-0.9916, Precision-0.9918, and Kappa 0.8870.In contrast, GNB outperforms in terms of ROC-0.9569 and MCC-0.9569. The majority of ML classifiers also achieve high levels of accuracy. AI will positively supplement and challenge decision-making processes and improve farming practices. Such technological interventions will likely result in better agricultural practices, yields, and a qualitative improvement in farmers' lives. As a result, we conclude that AI approaches opt for an intelligent crop recommendation system and can be used effectively. As an extension to this model, there is enormous potential to make this research more users friendly along with the help of chatbots.

References

1 Adegbeye M J, Reddy P R, Obaisi A I, Elghandour M M, Oyebamiji K J, Salem A Z, Morakinyo-Fasipe O T, Cipriano-Salazar M & Camacho-Díaz L M, Sustainable agriculture options for production, greenhouse gasses and pollution alleviation, and nutrient recycling in emerging and transitional nations-An overview, *J Cleaner Produc*, **242** (2020) 18319.

2 Mutuku E A, Roobroeck D, Vanlauwe B, Boeckx P & Cornelis W M, Maize production under combined conservation agriculture and integrated soil fertility management in the sub-humid and semi-arid regions of Kenya, *Field Crops Res*, **254** (2020) 107833.

3 KiryushinV I, The management of soil fertility and productivity of agrocenoses in adaptive-landscape farming systems, *Eurasian Soil Sci*, **52** (2019) 1137–1145.

4 Nordjo R E & Adjasi C K, Integrated soil fertility management (ISFM) and productivity of smallholder farmers in the northern region of Ghana (2019).

5 Ogunlade M O & Orisajo S B, Integrated soil fertility management for small holder cocoa farms: Using combination of cocoa pod husk based compost and mineral fertilizers, *Int J Plant & Soil Sci*, **32(2)** (2020) 68–77.

6 Shukla S K, Solomon S, Sharma L, Jaiswal V P, Pathak A D & Singh P, Green technologies for improving cane sugar productivity and sustaining soil fertility in the sugarcane-based cropping system, *Sugar Technol*, **21(2)** (2019) 186–196.

7 Nwite J C, Nwafor S O, Nwangwu A O & Olejeme O C, Enhancing soil fertility, maize grain yield, and nutrients composition through different planting time and manure sources in farmers' fields of Southeastern Nigeria, *Asian Res J Agric,* (2018) 1–12.

8 Nabavi-Pelesaraei A, Rafiee S, Mohtasebi S S, Hosseinzadeh-Bandbafha H & Chau K W, Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production, *Sci Total Environ*, **631** (2018) 1279–1294.

9 Abrougui K, Gabsi K, Mercatoris B, Khemis C, Amami R & Chehaibi S, Prediction of organic potato yield using tillage systems and soil properties by an artificial neural network (ANN) and multiple linear regressions (MLR), *Soil and Tillage Res*, **190** (2019) 202–208.

10 Suchithra M S & Pai M L, Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters, *Inf Process Agric* **7(1)** (2020) 72–82.

11 Kouadio L, Deo R C, Byrareddy V, Adamowski J F & Mushtaq S, Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties, *Comput electron agric*, **155** (2018) 324–338.

12 Mohammadi T A, Ahmadi A, Gómez P A & Maghoumi M, Using the artificial neural network in determining postharvest LIFE of kiwifruit, *J Sci Food Agric*, **99(13)** (2019) 5918–5925.

13 Toriyama K, Development of precision agriculture and ICT application thereof to manage spatial variability of crop growth, *J Soil Sci Plant Nutr*, **66(6)** (2020) 811–819.

14 Bestelmeyer B T, Marcillo G, McCord S E, Mirsky S, Moglen G, Neven L G, Peters D, Sohoulande C &Wakie T, Scaling up agricultural research with artificial intelligence, *IT Prof*, **22(3)** (2020) 33–38.

15 Dharmaraj V & Vijayanand C, Artificial intelligence (AI) in agriculture, *Int J Curr Microbiol Appl Sci*, **7(12)** (2018) 2122–2128.

16 Palanivel K & Surianarayanan C, An approach for prediction of crop yield using machine learning and big data techniques, *Int J Comput Eng Technol*, **10(3)** (2019) 110–118.

17 Apat S K, Mishra J, Raju K S & Padhy N, A study on smart agriculture using various sensors and agrobot: A case study, *Smart Intel Comput Appl*, **1** (2022) 531–540.

18 Dasgupta A, Drineas P, Harb B, Josifovski V & Mahoney M W, Feature selection methods for text classification, *Proc 13th ACM SIGKDD Int conf Knowl Discov Data Mining*, 2007, 230–239.

19 Kou G, Yang P, Peng Y, Xiao F, Chen Y & Alsaadi F E, Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods, *Appl Soft Comput*, **86** (2020) 105836.

20 Pandith V, Kour H, Singh S, Manhas J S & Sharma V, Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis, *J Sci Res* **64(2)** (2020) 394–398.

21 Maya Gopal P S & Bhargavi R, Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms, *Appl Artif Intell*, **33(7)** (2019) 621–642.

22 Blackman N J & Koval J J, Interval estimation for Cohen's Kappa as a measure of agreement, *Stat Med*, **19(5)** (2000) 723–741.

23 Apat S K, Mishra J, Raju K S & Padhy N, The robust and efficient Machine learning model for smart farming decisions and allied intelligent agriculture decisions, *Int J Integr Sci Technol*, **10(2)** (2022) 139–155.

24 Argüeso D, Picon A, Irusta U, Medela A, San-Emeterio M G, Bereciartua A & Alvarez-Gila A, few-shot learning approach for plant disease classification using images taken in the field, *Comput Electron Agric*, **175** (2020) 105542.

25 Apat S Kumar, Mishra J, Raju K S & Padhy N, IoT-assisted crop monitoring using machine learning algorithms for smart farming, in *Next Generation of Internet of Things. Lecture Notes in Networks and Systems* (Springer, Singapore). edited by R Kumar, P K Pattnaik, R S, J M Tavares, **445**, https://doi.org/10.1007/978-981-19-1412-6_1

26 Toseef M & Khan M J, An intelligent mobile application for diagnosis of crop diseases in Pakistan using fuzzy inference system, *Comput Electron Agric*, **153** (2018) 1–11.

27 Muangprathub J, Boonnam N, Kajornkasirat S, Lekbangpong N, Wanichsombat A &Nillaor P, IoT and agriculture data analysis for smart farm, *Comput Electron Agric,* **156** (2019) 467–474.

28 Reynolds M, Kropff M, Crossa J, Koo J, Kruseman G, Molero Milan A, Rutkoski J, Schulthess U, Sonder K, Tonnang H & Vadez V, Role of modeling in international crop research: overview and some case studies, *Agronomy*, **8(12)** (2018) 291.

29 Apat S K, Mishra J, Srujan Raju K & Padhy N, State of the art of ensemble learning approach for crop prediction, *Next Gen IoT*, (2023) 675–685.

30 Picon A, Alvarez-Gila A, Seitz M, Ortiz-Barredo A, Echazarra J & Johannes A, Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild, *Comput Electron Agric*, **161** (2019) 280–290.