An Empirical Evaluation of Machine Learning Techniques for Crop Prediction

G. Mariammal[1], A. Suruliandi[2], S.P. Raja[3], E. Poongothai[4] *

[1] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai - 600 062, Tamilnadu, (India)
[2] Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627012, Tamilnadu, (India)
[3] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, (India)
[4] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, (India)

Received 5 June 2021 | Accepted 11 April 2022 | Published 23 December 2022

Abstract

Agriculture is the primary source driving the economic growth of every country worldwide. Crop prediction, which is critical to agriculture, depends on the soil and environment. Nutrient levels differ from area to area and greatly influence in crop cultivation. Earlier, the tasks of crop forecast and cultivation were undertaken by farmers themselves. Today, however, crop prediction is determined by climatic variations. This is where machine learning algorithms step in to identify the most relevant crop for cultivation. This research undertakes an empirical analysis using the bagging, random forest, support vector machine, decision tree, Naïve Bayes and k-nearest neighbor classifiers to predict the most appropriate cultivable crop for certain areas, based on environment and soil traits. Further, the suitability of the classifiers is examined using a GitHub prisoners' dataset. The experimental results of all the classification techniques were assessed to show that the ensemble outclassed the rest with respect to every performance metric.

Keywords

Classification, Crop Prediction, Environmental Characteristics, Machine Learning, Soil Characteristics.

DOI: 10.9781/ijimai.2022.12.004

I. Introduction

AGRICULTURE is key to the development of human civilization, with farming playing a critical role in the process. Crop cultivation varies across areas, with each possessing unique soil, climatic and geographic characteristics. Soil is central to crop cultivation, and nutrients namely potassium, nitrogen, and phosphorus impact yield. Geography and climatic conditions, including the seasons, soil types, rainfall, and temperature also greatly influence in crop prediction. Based on these factors, the most suitable cultivable crop is predicted using several Machine Learning (ML) [1] techniques. Classification is fundamental to machine learning, for which it trains the system to obtain results using the given data. The supervised, unsupervised and reinforcement learning types of classification techniques are used in prediction. This research evaluates the performance of supervised learning techniques such as bagging, random forest (RF), support vector machine (SVM), decision tree (DT), Naïve Bayes (NB) and k-nearest neighbor (kNN) to predict a relevant crop for classification, using a GitHub prisoners' dataset. This work identifies the best classifier for the forecasting process.

* Corresponding author.
E-mail addresses: suba.g1212@mail.com (G. Mariammal), suruliandi@yahoo.com (A. Suruliandi), avemariaraja@gmail.com (S.P. Raja), poongothai.rp@gmail.com (E. Poongothai)

A. Related Work

Several papers that illustrate key features of common ML models are discussed in this section.

Soil characteristics alone are used to predict a suitable crop for cultivation [1]. Belson et al. [2] described the DT classification model as a tree structure, with leaf nodes representing the final decision made after the top-to-bottom path is established. The most efficient techniques used in the literature survey include the Gaussian mixture, the Chi-square Automatic Interaction Detector (CHAID), classification and regression trees, and the Bayesian network, presented by Duda et al. [3], Kass et al. [4], Breiman et al. [5], and Neapolitan [6], respectively. The NB classification technique, built on the Bayesian theorem, produces accurate forecast results that are easy to train and classify. Kohonen [7] and Atkeson et al. [8] discussed memory-based models and constructed hypotheses directly from the available data. However, information overload can increase their complexity. Data mining techniques with applications in agriculture include the K-means algorithm to forecast atmospheric emissions, the kNN to model daily precipitation and miscellaneous climatic variables, and the SVM to analyze possible adjustments to the weather. Bayesian models such as the NB, Gaussian NB and multinomial NB used in the prediction process were explained [9] - [11].

Ensemble learning (EL) models enhance the prediction process by constructing a prediction model using single base learners. Ensemble techniques such as bagging, boosting and the AdaBoost algorithms

were discussed and implemented [12] - [14]. The widely used SVM technique was improved through the use of the kernel trick for prediction [15], [16]. Breiman [17] proposed an RF technique, which is an ensemble model, and combined it with several DTs to constitute a single tree for a prediction model. The results are obtained after a comparison of all the trees in the forest and a final decision is made, based on the voting method. Suykens et al. [18] presented the minimum squares SVM, and Galvao et al. [19], the successive vector support algorithm. The proximity of data points is shown in the decision surface, that is, hyperplane support vectors. The data in the hyperplane are linearly separated by the total distance in the SVM. Babu et al. [20] discussed the application of artificial intelligence and ML algorithms in crop prediction. Designing an expert framework for crop cultivation calls for the services of computer engineers to model it, agricultural scientists to program it, and the know-how of experts in the field to back it up. Veenadhari et al. [21] described the role of data mining in agriculture. The most suitable crop for cultivation was predicted with 95% accuracy, based on climatic conditions as a major feature.

Monali et al. [22] posited a prediction system that categorizes soil types and predicts crop yields using the NB and kNN methods. Jeong et al. [23] explained the ability of the RF to predict crop yield responses to global and regional weather as well as biophysical variables in wheat, maize and potato. Sellam et al. [24] discussed crop yield prediction, which is primarily dependent on environmental characteristics, using regression analysis and linear regression. In their work, Pudumalar et al. [25] proposed a new ensemble model using the random tree, CHAID, kNN and NB to recommend crops for specific zones.

Zala [26] described bagging as a meta-algorithm that complements the power and precision of the ML technique used in mathematical classification and regression. It also eliminates variations and averts overfitting. Balducci et al. [27] described the DT as a predictive model and tested it at every level requiring decisions to be made. The levels depend on the request and outcome of the decision-making process. Jahan [28] averred that the NB is vulnerable to insignificant characteristics. Given its solid foundation, it manages both confidential and streaming data with ease. Priya et al. [29] used real-time Tamil Nadu facts to predict crop yields the usage of the RF method. Suresh et al. [30] examined soil profiles in conjunction with Global Positioning System-based technologies. The K-means and modified kNN are implemented to predict crop yields in Tamil Nadu.

B. Motivation and Justification

Crop prediction, which is critical to agriculture, employs machine learning algorithms for the purpose. Classification is central to machine learning [40]-[42]. It helps to learn the system for forecasting a relevant cultivable crop. Classifiers are divided into two sub-categories, single learner and ensemble learners. Thus motivated, various supervised classifiers are examined for the prediction process. Though the literature analysis makes it evident that the ensemble model offers better predictions, much of the research has, however, tended to use single learners for crop prediction. An ensemble model, which helps improve the prediction rate, is constructed using single learners. Thus justified, the efficiency of the ensemble model is examined with a crop dataset and a GitHub prisoners' dataset, using different performance metrics. The performance of the ensemble bagging model is evaluated with existing classification algorithms such as the RF, SVM, DT, NB and kNN for the prediction process.

C. Contributions

The significance contributions of this research are given below:

- The literature survey shows that much of the earlier work has examined either soil or environmental factors to predict crop cultivation. This work, on the other hand, undertakes crop prediction by examining both.
- A real-time dataset composed from the Sankarankovil Agriculture Department of Tenkasi District in the state of Tamil Nadu in India is used for the prediction process.
- The primary goal of this work is to predict an appropriate classifier for all sort of datasets.
- Further, the classifier performance is examined by the various k - fold and data splitting methods.

D. Outline of the Work,

Fig. 1 illustrates the comprehensive process of this work. Input data is fed into pre-processing step. In pre-processing, missing values in the dataset are identified to eliminate the redundant values. This is used to handle the imbalanced data which was done by mean imputation method. It improves the prediction performance of classifiers and accuracy rate has been increased after pre-processing stage. After that, the dataset is broken down into training and testing. Classifiers are well trained to predict the target class with the help of all training samples. The learned classifier is validated with the unknown samples from the testing dataset. The learned classifier helps to forecast the target class of the given dataset. Finally, the predicted result is examined by various performance metrics.

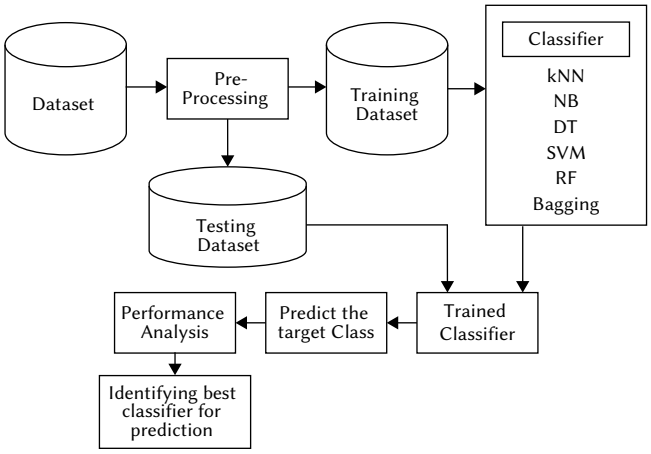

Fig. 1. Outline of the Work.

E. Organization of the Paper

The remainder of the article is organized in the following way: Section II describes the methodology for crop prediction. Section III illustrates the experimental results and final section concludes this work.

II. Methodology

A. Classification

Classification is indispensable to machine learning, given that it predicts the outcome of the process. This work evaluates the performance of the existing bagging, RF, SVM, DT, NB and kNN classifiers, using a crop dataset and a GitHub prisoners' dataset.

1. K Nearest Neighbor (KNN)

The kNN is not a complex algorithm that classifies new instances established on positive similarity measures [25]. The similarity degree is calculated by distance measures which include Euclidean distance, Manhattan distance, and many others [31]. In the kNN, feature vectors are stored in the training phase of the algorithm. The kNN technique

finds the similarity between unknown classes with known instances. The class labeling of training instances and unlabeled vectors are classified by way of assigning the most common label of the closest training samples. In the iterative classification, k is a parameter set by the user [27]. Algorithm 1 gives the pseudo code of the kNN classifier.

Algorithm 1. The pseudo code for the kNN

Input: s, C, d where s is unknown sample from dataset C;
d is the distance
Output: class of s
for $(s', c') \in D$ *do*
Compute the distance $d(s', s)$
end for
Order the $|C|$ *distances by increasing the sequence*
Calculate the number of hits for each class c_i *among the kNN*
Assign s to the highest class

2. Naive Bayes (NB)

In order to construct classifiers, the NB method offers class labels to problem instances [25] which is entrenched the theorem of Bayes' [28]. NB is one of the most effective classifiers and it predicts the outcome based on the probability of an instance. It deems the value of a separate variable to be independent of the value of any other given quality in the class variable. [25]. It is utilized for both binary and multi class classification problems. Algorithm 2 gives the pseudo code of the NB classifier [32].

Algorithm 2. The pseudo code for the NB

Input: $C \rightarrow$ *Dataset,* $T_1 \rightarrow$ *Training dataset,* $P = (p_1, p_2, ..., p_n)$ *//*
value of the predictor variable in testing dataset
Output: Predicted class

Step:

1. *Read all training data* T_1
2. *The mean and standard deviation of the predictive variable in each category are calculated*
3. *Repeat*
   Compute the likelihood of p_1 *using the gauss density equation in each class*
   Until the likelihood of all predictor variables $(p_1, p_2, ..., p_n)$ *has been computed*
4. *Compute the probability of each class*
5. *Obtain the highest likelihood*

3. Decision Tree (DT)

The DT is a single tree predictive model, and it is used for both classification and regression problems. DT is like a tree structure method [27]. Decision nodes and leaves make up a tree [31]. Each internal node shows an input variable, and each leaf node shows class prediction. It works supported a top-down approach by selecting a worth for the feature at every stop that best splits a collection of things [27], looking on the applying and decision-making outcome. DT algorithms contain the CART, C4.5 and ID3. In this work CART technique is used for implementing the DT algorithm which stands for classification and regression tree. Algorithm 3 gives the pseudo code of the DT classifier [33].

Algorithm 3. The pseudo code for the DT

Input: Dataset C,R number of Instances, P features
Output: Predicted class
ConstructTree (R):
if R contains instances of a single class then
return
else
The feature P which has the greatest information gain is selected to split on
Generate p leaf nodes of R,
whereR has $R_1, ..., R_p$ *and P has p possible values* $(P_1, ..., P_p)$
for i = 1 to p do
Define the content of R_i to C_i,where C_i is all the instances in R that match P_i
Get ConstructTree (R_i)
end for
end if

4. Support Vector Machine (SVM)

The SVM is a type of machine learning that information needed to determine into decision surfaces. It is used for both classification and regression problems. In this work, the SVM algorithm is used to categorize the result according to the input variables. The decision surfaces then break the data into two hyperplanar groups. [16]. The training data identify the vector that assists the hyperplane. Apparently, due to the larger margins, with a weak classifier generalization, a hyperplane that is farther away from the nearest training data point consistently has better margins and larger mistakes. Algorithm 4 gives the pseudo code of the SVM classifier [34].

Algorithm 4. The pseudo code for the SVM

Input: Dataset C
Output: Predicted Class
Require: X and y uploaded with training labeled data, $\alpha \leftarrow 0$ *or*
$\alpha \leftarrow$ *partially trained SVM*
$A \leftarrow$ *any value*
repeat
for all $\{x_i, y_i\}, \{x_j, y_j\}$ *do*
Optimize α_i *and* α_j
end for
until a change of α or other resource constraint criteria is not met
Ensure:Remember only the support vectors $(\alpha_i > 0)$
Test the model
Calculate Scores
Compute Confusion Matrix
Validate Model

5. Random Forest (RF)

The RF is a well-known and extensively used supervised machine learning approach to solve classification and regression issues [29]. The RF is an ensemble technique, and it combines several homogeneous learners as a single model. It uses decision tree algorithm for the prediction process, and it takes the final decision based on the average voting method. Algorithm 5 gives the pseudo code of the RF classifier [33].

Algorithm 5. The pseudo code for the RF

Input:Dataset C,R number of instances,P features
Output:Predicted class
To create L classifiers
for i = 1 to l do
Randomly select the training data C with substitution to produce C_i
Generate a parent node, R_i *containing* C_i
Get Construct Tree (R_i)
end for
ConstructTree (*R*)
if R contains instances of a single class then
return
else
The x% of possible splitting features in R are randomly selected
The feature P which has the greatest information gain is selected to split on
Generate p leaf nodes of R,R_1,...,R_p;where P has p possible values (P_1, ..., P_p)
for i = 1 to p do
Define the content of R_i *to* C_i, *where* C_i *is all the instances in R that match* P_i
Get ConstructTree (R_i)
end for
end if

6. Bagging

Bagging, also termed bootstrap aggregation, is a technique that was developed by Leo Breiman [26] to train and combine numerous homogenous learning algorithms. [13]. Bagging technique is used to reduce the problems related to overfit. Bagging is based on parallel method, and it uses data subsets for training the base learners. It optimizes the learning algorithm's robustness as well as the prediction algorithm's results [26]. It predicts the outcome with the help of voting method for classification. Since bagging does not allow recalculation of weight, changing the weight update equation is critical or reviews the algorithm's calculations. Algorithm 6 gives the pseudo code of the Bagging classifier [35].

Algorithm 6. The pseudo code for the Bagging

Input: T_1*: Training sample of C size dataset,*
s: count of bootstrap samples, Lc: Learning Classifier
Output: L bagging ensemble with s element classifiers*
Learning stage:
for i = 1→ s do
$K_i \leftarrow$ *bootstrap sample from C*
Create classifier $L_i \leftarrow Lc\ (K_i)$
end for
Predict the class label for a new sample
$L^*(x) = arg\ arg\ max_y \sum_{i=1}^{s} [L_i(x) = y]$

B. Characteristic Comparison of Each Classifier

This section discusses the pros and cons of each of the classifiers used for prediction. The kNN handles both classification and regression problems well but cannot deal with missing values. Though each feature makes unique assumptions about prediction outcomes, the NB is unaffected by irrelevant characteristics. While the DT provides feasible and adequate results for large data sources relatively rapidly, the algorithm must be trained over a long period of time and is also much more complex. Though the SVM is most effective at higher dimensions, it is vital to select a hyperparameter appropriately, and there is no probabilistic explanation for the classification. The RF handles missing data very well, but overfitting occurs with noisy data. On huge datasets, the bagging approach performs well; nonetheless, there is a loss of interpretability in the model.

III. Experimental Result Analysis & Discussions

A. Dataset Description

This research utilizes two different types of datasets such as Crop and Prisoner's respectively. The details of these dataset are given in Table I.

TABLE I. Dataset Description

Dataset	Number of Instances	Number of Attributes	Type
Crop	1000	16	Nominal
Prisoner's	463	31	Numeric
Iris	150	4	Nominal

The crop dataset comprises soil and environmental factors, is downloaded from www.tnau.ac.in. The crop dataset has 1000 instances with 16 attributes in which 12 attributes are soil characteristics such as macro nutrients (nitrogen, phosphorus, potassium, etc.), macronutrients (zinc, iron, copper, etc.) and the remaining 4 are environmental such as rainfall, soil texture, temperature, and season. Also, this work utilizes to validate the performance of classifiers with other two dataset such as prisoner's and iris. Crime Propensity Prediction dataset [36] that can be used to predict the crime of a prisoner which was taken from the website github.com. The prisoner's dataset contains behavior of the prisoners with 463 instances and 31 attributes. Iris dataset [37] helps to find the iris plant class, which was downloaded from the University of California, Irvine. The dataset includes types of iris plant with 150 instances and 4 attributes.

B. Performance Metrics

The performance metrics namely, Accuracy, Kappa, Precision, Specificity, F1 Score, Area Under the Curve (AUC), and Mean Absolute Error (MAE) are used to predict the performances of each classifier. The formulae, and a representation of each metric used in the result examination, are stated in [38, 39].

C. Results and Discussion

In this section, the prediction performances of the classifiers are examined by various above mentioned performance metrics.

1. Sample Input and Output

Table II demonstrates the sample input and output range of the crop dataset, which includes the 12 soil characteristics of the potential of Hydrogen (pH), electrical conductivity (EC), organic carbon (OC), nitrogen (N), phosphorus (P), potassium (K), sulphur (S), zinc (Zn), boron (B), iron (Fe), manganese (Mn), and copper (Cu), as well as the 4 environmental characteristics of soil texture, seasons, rainfall, and average temperature. The expected output for the given input data, collected from the particular region, is given.

2. An Empirical Assessment of Classifiers Based on Soil Characteristics

Table III compares of the classifiers and identifies the best for appropriate crop. The process is based solely on soil characteristics like pH, N, and P.

The ensemble bagging classification technique clearly beats the others, as evidenced by the findings. Bagging also receives votes for increased performance for each sample, and as a result, it has a higher crop prediction accuracy than other approaches.

TABLE II. Sample Input and Output

Input															Output
pH	EC	OC	N	P	K	S	Zn	B	Fe	Mn	Cu	Texture	Season	Rainfall	Avg. Temp	
7.8	0.72	0.26	160	252.5	400	16	0.56	0.95	10.64	6.46	0.98	1	Kharif	296.8	25	Black Grams
7.8	0.72	0.26	160	252.5	400	16	0.56	0.95	10.64	6.46	0.98	1	Kharif	296.8	25	Black Grams
7.4	0.28	0.26	157.5	95.0	720	23	0.76	0.86	15.20	11.60	0.82	1	Rabi	296.8	25	Chick pea
7.9	0.73	0.31	175	267.5	400	31	0.54	0.84	9.86	6.36	1.02	1	Kharif	296.8	25	Maize
7.9	0.73	0.31	175	267.5	400	31	0.54	0.84	9.86	6.36	1.02	1	Kharif	296.8	25	Maize
7.4	0.28	0.26	157.5	95.0	720	23	0.76	0.86	15.20	11.60	0.82	1	Rabi	296.8	25	Chick pea
7.4	0.28	0.26	157.5	95.0	720	23	0.76	0.86	15.20	11.60	0.82	1	Rabi	296.8	25	Chick pea
8.2	0.14	0.20	140	97.5	972.5	13.3	0.78	0.49	9.50	5.54	0.74	1	Kharif	296.8	25	Maize
8.2	0.10	0.02	140	240	1000	2.34	0.82	0.34	10.40	5.54	1.08	1	Kharif	296.8	25	Maize

TABLE III. Empirical Evaluations of Classifiers Based on Soil Factors

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	79.92	76.97	79.63	82.78	80.24	79.93	80.21	0.45
NB	80.65	78.07	81.03	83.90	81.25	81.13	81.06	0.37
DT	83.37	81.17	83.15	85.56	83.60	83.37	86.00	0.33
SVM	84.82	83.08	85.70	86.00	86.30	85.99	85.56	0.28
RF	88.82	87.35	88.73	90.27	89.79	89.25	89.19	0.20
Bagging	**91.55**	**90.42**	**91.27**	**92.59**	**91.79**	**91.52**	**92.34**	**0.18**

TABLE IV. Empirical Evaluations of Classifiers Based on Environment Factors

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	46.6	43.65	46.31	49.46	46.92	46.61	47.04	0.88
NB	47.33	44.75	47.71	50.58	47.93	47.81	48.12	0.78
DT	50.05	47.85	49.83	52.24	50.28	50.05	50.77	0.71
SVM	52.50	50.76	53.38	53.68	53.98	53.67	53.34	0.63
RF	54.50	53.03	54.41	55.95	55.47	54.93	55.00	0.57
Bagging	**58.23**	**57.10**	**57.95**	**59.27**	**58.47**	**58.20**	**59.45**	**0.50**

TABLE V. Empirical Evaluations of Classifiers Based on Soil and Environment Factors

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	77.73	75.78	78.44	81.59	79.05	78.74	79.43	0.40
NB	81.46	78.88	81.84	84.71	82.06	81.94	82.00	0.33
DT	84.18	81.98	83.96	86.37	84.41	84.18	85.07	0.27
SVM	86.63	84.89	87.51	87.81	88.11	87.80	87.83	0.20
RF	91.63	90.16	91.54	93.08	92.60	92.06	92.12	0.19
Bagging	**93.36**	**92.23**	**93.08**	**92.32**	**93.12**	**93.10**	**94.89**	**0.12**

3. An Empirical Assessment of Classifiers Based on Environment Conditions

Table IV depicts a comparison of the classification techniques that are exclusively based on environment factors like season, texture, average temperature and rainfall.

Table IV depicts the prediction rate based only on environmental characteristics. However, the bagging classification technique performs better than others, based only on environmental factors. In addition, bagging offers improved crop prediction accuracy because it uses multiple learning algorithms.

4. An Empirical Assessment of Classifiers Based on Soil and Environmental Characteristics

Table V shows a performance estimation of the classifiers, based on both soil and environmental factors, to find the right crop for cultivation in a specific area.

The information in Table V suggests that the prediction rate is higher with the combined features of both the soil and the environment, rather than that based solely on either of the two. Combining soil and environmental data, bagging provides more accurate prediction results than other classifiers. With the bagging technique's aggregation operation, the variance of estimation is greatly minimized.

TABLE VI. Empirical Evaluations of Classifiers for Prisoner's Dataset

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	94.20	91.25	93.91	97.06	94.52	94.21	94.93	0.40
NB	94.93	92.35	94.59	97.46	94.81	94.69	95.60	0.30
DT	95.65	93.45	95.43	97.84	95.88	95.65	96.00	0.25
SVM	96.10	94.36	96.22	97.58	96.98	96.59	96.89	0.20
RF	97.10	95.63	97.01	98.55	97.18	97.09	98.07	0.13
Bagging	**97.83**	**96.70**	**97.55**	**98.87**	**98.07**	**97.80**	**98.50**	**0.10**

TABLE VII. Empirical Evaluations of Classifiers for Iris Dataset

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	89.32	87.13	87.72	91.61	88.01	87.86	90.54	0.40
NB	91.43	89.00	91.00	94.55	92.11	91.55	92.77	0.31
DT	90.93	89.42	88.88	92.63	89.23	89.05	92.04	0.26
SVM	92.74	90.39	92.68	94.19	93.42	93.04	93.87	0.21
RF	94.32	92.42	93.12	96.39	94.93	94.01	96.49	0.15
Bagging	**95.43**	**93.90**	**95.21**	**97.00**	**96.21**	**95.707**	**97.51**	**0.11**

The results are evaluated with the real-world dataset discussed in section 3.1, using various classifiers. The results show that the ensemble technique provides the most accurate results of all.

5. Performance Evaluation of Classification Techniques for Prisoners Dataset

Further, the classifiers are tested, and their performance is verified with the prisoners' dataset downloaded from the GitHub website. Table VI presents the performance-wise results of the classification techniques, using the metrics discussed in section 3.3.

It is inferred that the ensemble learner gives better prediction accuracy, at 97.83%, than single learners. The bagging technique outperforms other classifiers in prediction. The ensemble technique combines two or more single prediction models for the best prediction rate. Since the performance of the bagging technique is unaffected by missing values, it works better than other techniques.

6. Performance Evaluation of Classification Techniques for Iris Dataset

Consequently, the performance of the classifiers is evaluated with the iris dataset which was taken from the UCI website to predict the class of iris plant. Table VII presents the performance-wise results of the classification techniques, using the metrics discussed in section 3.3.

It depicts that the ensemble learner gives better prediction accuracy, at 95.43%, than single learners. The bagging technique works well than other classifiers in iris plant prediction.

7. Empirical Evaluation of the Bagging Technique Using K-fold Validation

The results presented in Tables III-VII show that the bagging technique performs better than the rest. The best fold for the bagging technique is determined using the fold variation method. The fold method is used to evaluate the potential of each classifier for prediction process. The given dataset is divided into two subgroups, with the first (k -1) being used to train the classifier and the second (k^{th}) being used to examine the classifier.

Table VIII depicts a performance evaluation of the bagging classification technique for the crop, prisoners' and iris datasets, examining outcome prediction using several folds.

Table VIII presents the bagging technique performance for fold variation, with folds ranging from 10 to 90. The bagging technique performs better with 10 folds than any other. Performance is evaluated using the metrics given in section 3.3. The results show that the bagging technique achieved accuracy of 93%, 98% and 95% for the crop and prisoners' datasets, respectively.

8. Empirical Evaluation of the Bagging Technique Using Data Splitting Validation

To determine the best data splitting range, the bagging classifier's efficiency for the prediction process is assessed using the data splitting validation method. The graphical representation below displays a performance assessment of the bagging technique for the crop and prisoners' datasets, based on data fractionation, with ranges varying from 25% - 75% and 75% - 25%. Performance is assessed according to the metrics described in section 3.3.

Fig. 2 depicts the performance of the bagging classifier in forecasting an appropriate crop, using a crop dataset based on the data splitting validation method. From the results, it is observed that the 70% - 30% splitting range outperforms others.

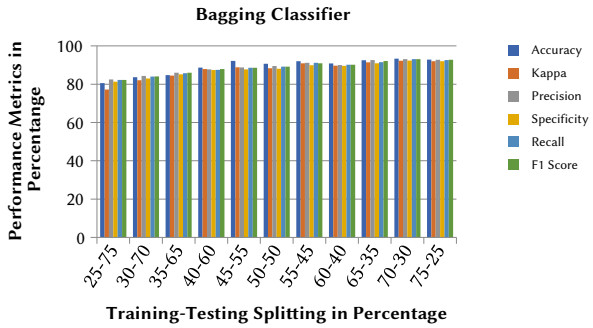

Fig. 2. Performance assessment of the Bagging classifier for Crop dataset utilizing data splitting method.

For the prisoners' prediction dataset, Fig. 3 illustrates a performance review of the bagging classification approach. The results show that the bagging technique outperforms the data splitting strategy in the 70% - 30% range.

TABLE VIII. Performance of the Bagging Technique Based on Fold Variation

Dataset	Folds	Performance Metrics					
		Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	Sensitivity (%)	F1 Score (%)
Crop	**10**	**93.36**	**92.23**	**93.12**	**93.08**	**92.32**	**93.10**
	20	92.04	90.91	91.80	91.76	91.00	91.78
	30	89.74	88.60	89.49	89.45	88.69	89.47
	40	91.06	89.62	90.51	90.47	89.71	90.49
	50	90.54	89.10	89.99	89.95	89.19	89.97
	60	89.24	87.80	88.69	88.65	87.89	88.67
	70	89.94	88.40	89.29	89.25	88.49	89.27
	80	90.51	88.97	89.86	89.82	89.06	89.84
	90	89.04	88.64	89.53	89.49	88.73	89.51
Prisoners	**10**	**97.83**	**96.70**	**98.07**	**97.55**	**98.87**	**97.80**
	20	96.51	95.38	96.75	96.23	97.55	96.48
	30	94.21	93.07	94.44	93.92	95.24	94.17
	40	95.53	94.09	95.46	94.94	96.26	95.19
	50	95.01	93.57	94.94	94.42	95.74	94.67
	60	93.71	92.27	93.64	93.12	94.44	93.37
	70	94.41	92.87	94.24	93.72	95.04	93.97
	80	94.98	93.44	94.81	94.29	95.61	94.54
	90	93.51	93.11	94.48	93.96	95.28	94.21
Iris	**10**	**95.43**	**93.90**	**96.21**	**95.21**	**97.00**	**95.70**
	20	94.87	93.56	95.65	94.97	95.78	95.30
	30	94.43	92.23	94.32	93.42	95.54	93.86
	40	95.11	93.45	95.48	94.51	96.18	94.99
	50	93.35	91.32	93.45	92.90	94.45	93.17
	60	93.66	92.45	94.12	92.39	94.43	93.24
	70	94.57	93.23	95.42	94.12	95.11	94.76
	80	92.14	91.45	93.04	92.04	93.34	92.53
	90	92.12	90.93	92.94	91.79	93.01	92.36

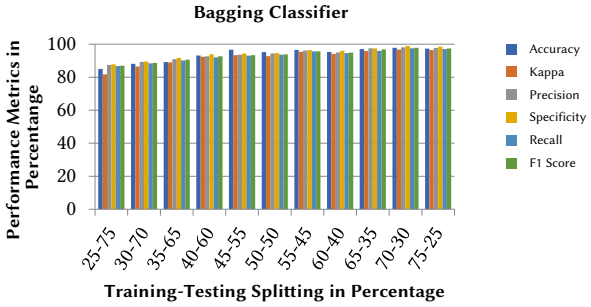

Fig. 3. Performance assessment of the Bagging classifier for Prisoners dataset utilizing data splitting method.

Fig. 4 shows a performance validation of the bagging classification method for the iris plant prediction dataset. The bagging technique performs better in the 70-30% data splitting range, according to the findings.

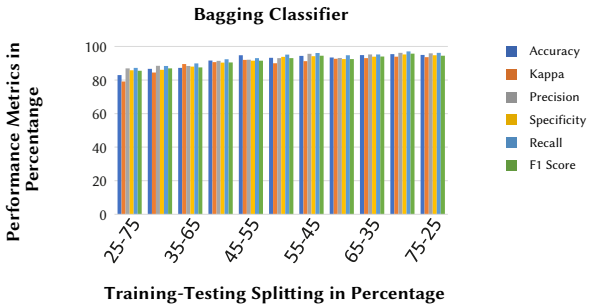

Fig. 4. Performance assessment of the Bagging classifier for Iris dataset utilizing data splitting method.

9. Empirical Evaluation of Each Classifiers Based on Time and Memory Occupation

Table IX illustrates the performance assessment of each classifier according to the execution time and memory of each.

TABLE IX. Comparison Table of Each Classifier Based on Time and Memory Occupation

Classifiers	Time Taken (secs)	Space Occupied
kNN	0.19	260.72
NB	0.30	274.38
DT	0.47	261.45
SVM	0.29	292.87
RF	0.68	257.32
Bagging	0.70	246.01

It is evident from the results that though the bagging classifier requires a longer execution time than other techniques, it also occupies little space. It is inferred from Table 9 that the kNN classifier takes the lowest execution time with 260.72 KB of occupied space, while the SVM occupies the highest space but takes the second-lowest execution time.

IV. Discussions and Conclusion

A great deal of research has been carried out on the forecasting process using classification techniques. This work has examined the performance of the bagging, random forest, support vector machine, decision tree, Naïve Bayes and k-nearest neighbor classifiers using a crop dataset, a prisoners' dataset and iris dataset. Using these algorithms, a relevant crop for cultivation was predicted from the crop dataset, the prisoners' outcome predicted from the prisoners' dataset, and the type of iris plant is predicted from the iris dataset.

The performance of the classifiers was examined using several performance metrics as accuracy, kappa, sensitivity, specificity, F1 score, area under the curve, precision, and mean absolute error. The results have shown that the bagging ensemble technique outperforms the rest. Then the bagging technique is examined by two validation methods namely fold and data splitting method. The obtained results show the bagging technique performs well on 10-fold and 70% - 30% data splitting range than others for predicting the target class of the given dataset.

References

[1] S. A. Z. Rahman, K. Chandra Mitra and S. M. Mohidul Islam, "Soil Classification Using Machine Learning Methods and Crop Suggestion Based on Soil Series," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2018, pp.1-4, doi: https://doi.org/10.1109/ICCITECHN.2018.8631943.

[2] William A. Belson, "Matching and prediction on the principle of biological classification," in *Journal of the Royal Statistical Society: Series C (Applied Statistics)* vol. 8, pp. 65-75, 1959, doi: https://doi.org/10.2307/2985543.

[3] Richard O. Duda, Peter E. Hart, and David G. Stork. "Pattern classification and scene analysis", in New York: Wiley, vol. 3, 1973, doi: https://doi.org/10.1086/620282.

[4] Gordon V. Kass, "An exploratory technique for investigating large quantities of categorical data", in *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 29, pp. 119-127, 1980, doi: https://doi.org/10.2307/2986296.

[5] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen, "Classification and regression trees", in *CRC press*, 1984, doi: https://doi.org/10.1201/9781315139470.

[6] R.E. Neapolitan, "Models for reasoning under uncertainty", in *Appl. Artif. Intell.* vol. 1, pp. 337–366, 1987, doi: https://doi.org/10.1080/08839518708927979.

[7] T. Kohonen, "Learning vector quantization", in *Neural Netw,* vol. 1, pp. 303, 1988, doi: https://doi.org/10.1007/978-3-642-97610-0_6.

[8] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal, "Locally weighted learning", *In Lazy learning*, Springer, Dordrecht, 1997, pp. 11-73, doi: https://doi.org/10.1023/A:1006559212014.

[9] J Pearl, "Probabilistic Reasoning in Intelligent Systems", in *Morgan Kauffmann San Mateo,* vol. 88, pp. 552, 1988, doi: https://doi.org/10.1016/C2009-0-27609-4.

[10] J.R Quinlan, "C4.5: Programs for Machine Learning", in *Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA*, vol. 1, 1992, doi: https://doi.org/10.1007/BF00993309.

[11] S.J. Russell, and P Norvig, "Artificial Intelligence: A Modern Approach", in Prentice Hall: Upper Saddle River, NJ, USA, vol. 9, 1995, doi: 10.1016/j.artint.2011.01.005.

[12] L. Breiman, "Bagging Predictors", in *Mach. Learn,* vol. 24, pp. 123–140, 1996, doi: https://doi.org/10.1007/BF00058655.

[13] Yoav Freund, and Robert E. Schapire, "Experiments with a new boosting algorithm", *In ICML* 96, pp. 148-156, 1996, doi: https://dl.acm.org/doi/10.5555/3091696.3091715.

[14] Robert E. Schapire, "A brief introduction to boosting", *In IJCAI* 99, pp. 1401-1406, 1999, doi: https://dl.acm.org/doi/10.5555/1624312.1624417.

[15] A. Smola, "Regression Estimation with Support Vector Learning Machines", in *Master's Thesis*, The Technical University of Munich, Munich, Germany, pp. 1–78, 1996.

[16] Johan A.K. Suykens, and Joos Vandewalle, "Least squares support vector machine classifiers", in *Neural processing letters,* vol. 9, pp. 293-300, 1999, doi: https://doi.org/10.1023/A:1018628609742.

[17] Leo Breiman, "Random forests", in *Machine learning,* vol. 45, pp. 5-32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[18] J.A.K. Suykens, Van Gestel, T, De Brabanter, J, De Moor, B and J Vandewalle, "Least Squares Support Vector Machines", in World Scientific: Singapore, 2002, doi: https://doi.org/10.1142/5089.

[19] Galvao, Roberto Kawakami Harrop, Mario Cesar Ugulino Araujo, Wallace Duarte Fragoso, Edvan Cirino Silva, Gledson Emidio Jose, Sofacles Figueredo Carreiro Soares, and Henrique Mohallem Paiva, "A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm", in *Chemometrics and intelligent laboratory systems,* vol. 92, pp. 83-91, 2008, doi: https://doi.org/10.1016/j.chemolab.2007.12.004.

[20] M.S.P. Babu, N.V. Ramana Murty and S.V.N.L Narayana, "A web based tomato crop expert information system based on artificial intelligence and machine learning algorithms", in *International Journal of Computer Science and Information Technologies,* vol. 1, pp. 1–5, 2010, doi: 10.1.1.206.2072 .

[21] S. Veenadhari, Bharat Misra, and C. D. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters", *In 2014 International Conference on Computer Communication and Informatics,* IEEE, pp. 1-5, 2014, doi: https://doi.org/10.1109/ICCCI.2014.6921718.

[22] Paul Monali, Santosh K. Vishwakarma, and Ashok Verma, "Analysis of soil behaviour and prediction of crop yield using data mining approach", *In 2015 International Conference on Computational Intelligence and Communication Networks (CICN),* IEEE, pp. 766-771, 2015, doi: https://doi.org/10.1109/CICN.2015.156.

[23] Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin et al. "Random forests for global and regional crop yield predictions", in *PLoS One*, vol. 11, 2016, doi: https://doi.org/10.1371/journal.pone.0156571.

[24] V. Sellam, and E. Poovammal, "Prediction of crop yield using regression analysis", in *Indian Journal of Science and Technology,* vol. 9, pp. 5, 2016, doi: 10.17485/ijst/2016/v9i38/91714.

[25] S. Pudumalar, E. Ramanujam, R. Harine Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, "Crop recommendation system for precision agriculture", *In 2016 Eighth International Conference on Advanced Computing (ICoAC),* IEEE, pp. 32-36, 2017, doi: https://doi.org/10.1109/ICoAC.2017.7951740.

[26] Dipika H Zala, "Review on use of BAGGING technique in agriculture crop yield prediction", in *International Journal for Scientific Research & Development,* vol. 6, 2018.

[27] Fabrizio Balducci, Donato Impedovo, and Giuseppe Pirlo, "Machine learning applications on agricultural datasets for smart farm enhancement", in *Machines,* vol. 6, pp. 38, 2018, doi: https://doi.org/10.3390/machines6030038.

[28] Jahan Raunak, "Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land soils", in *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* vol. 6, pp. 189-193, 2018, doi: https:// 10.22214/ijraset.2018.5030.

[29] P. Priya, U. Muthaiah, and M. Balamurugan, "Predicting yield of the crop using machine learning algorithm", in *International Journal of Engineering Sciences & Research Technology,* vol. 7, pp. 1-7, 2018, doi: 10.5281/zenodo.1212821.

[30] A. Suresh, P. Ganesh Kumar, and M. Ramalatha, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN", *In 2018 3rd International Conference on Communication and Electronics Systems (ICCES),* IEEE, pp. 88-93, 2018, doi: 10.1109/CESYS.2018.8723956.

[31] D. Anantha Reddy, Bhagyashri Dadore, and Aarti Watekar, "Crop Recommendation System to Maximize Crop Yield in Ramtek region using Machine Learning", in *International Journal of Scientific Research in Science and Technology,* vol. 6, pp. 485 - 489, 2019, doi: https://doi.org/10.32628/IJSRST196172.

[32] Ivan Kholod, Andrey Shorov, and Sergei Gorlatch, "Improving Parallel Data Mining for Different Data Distributions in IoT Systems", *In International Symposium on Intelligent and Distributed Computing*, Springer, Cham, pp. 75-85, 2019, doi: https://doi.org/10.1007/978-3-030-32258-8_9.

[33] Grant Anderson, "Random relational rules", in *PhD diss.*, The University of Waikato, 2008.

[34] Angelina Tzacheva, Jaishree Ranganathan, and Sai Yesawy Mylavarapu, "Actionable Pattern Discovery for Tweet Emotions", *In International Conference on Applied Human Factors and Ergonomics,* Springer, Cham, pp. 46-57, 2019, doi: 10.1007/978-3-030-20454-9_5.

[35] Mateusz Lango, and Jerzy Stefanowski, "Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data", in *Journal of Intelligent Information Systems,* vol. 50, pp. 97-127, 2018, doi: https://doi.org/10.1007/s10844-017-0446-7.

[36] H. Benjamin Fredrick David, A. Suruliandi, and S.P. Raja, "Preventing crimes ahead of time by predicting crime propensity in released prisoners using Data Mining techniques", in *International Journal*

of Applied Decision Sciences, vol. 12, pp. 307 – 336, 2019, doi: 10.1504/IJADS.2019.100433.

[37] Dua, D. and Graff, C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[38] Mariammal, G., A. Suruliandi, S. P. Raja, and E. Poongothai. "Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics Using Modified Recursive Feature Elimination Technique With Various Classifiers." in *IEEE Transactions on Computational Social Systems*, vol. 8, no. 5, 2021, pp. 1132-1142, doi: 10.1109/TCSS.2021.3074534.

[39] Ganesan, Mariammal, Suruliandi Andavar, and Raja Soosaimarian Peter Raj. "Prediction of Land Suitability for Crop Cultivation Using Classification Techniques." in Brazilian Archives of Biology and Technology, vol. 64, 2021.

[40] Bhaik, A., Singh, V., Gandotra, E., & Gupta, D. (InPress). Detection of Improperly Worn Face Masks using Deep Learning – A Preventive Measure Against the Spread of COVID-19. International Journal of Interactive Multimedia and Artificial Intelligence, In Press(In Press), 1-12. http://doi.org/10.9781/ijimai.2021.09.003

[41] Alvarez, P., García de Quirós, J., & Baldassarri, S. (InPress). RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of Spotify Songs. International Journal of Interactive Multimedia and Artificial Intelligence, In Press(In Press), 1-14. http://doi.org/10.9781/ijimai.2022.04.002

[42] Sánchez-Torres, F., González, I., & Dobrescu, C. C. (2022). Machine Learning in Business Intelligence 4.0: Cost Control in a Destination Hotel. International Journal of Interactive Multimedia and Artificial Intelligence, 7 (Special Issue on Artificial Intelligence in Economics, Finance and Business), 86-95. http://doi.org/10.9781/ijimai.2022.02.008

G. Mariammal

G. Mariammal completed her B.E. degree in Computer Science and Engineering from Francis Xavier Engineering College, Tirunelveli, India, in 2011. She completed her M.E. degree in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli, India, in 2017, also completed her Ph.D. degree in Computer Science and Engineering at Manonmaniam Sundaranar University, Tamilnadu, India in 2021. Her research areas are machine learning, data analytics and image processing.

A. Suruliandi

A. Suruliandi completed his B.E. in Electronics & Communication Engineering in the year 1987 from Coimbatore Institute of Technology, Coimbatore. He completed his M.E. in Computer Science & Engineering in the year 2000 from Government College of Engineering, Tirunelveli. He obtained his Ph.D. in the year 2009 from Manonmaniam Sundaranar University, Tirunelveli. He is working as a professor in the Department of Computer Science & Engineering in Manonmaniam Sundaranar University, Tirunelveli. He is having more than 29 years of teaching experience. He published 50 papers in International Journals, 23 in IEEE Xplore publications, 33 in National conferences and 13 in International conferences. His research areas are remote sensing, image processing and pattern recognition.

S. P. Raja

S. P. Raja was born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. His area of interest is image processing and cryptography. He is having more than 14 years of teaching experience in engineering colleges. Currently he is working as an Associate Professor in the school of Computer Science and Engineering in Vellore Institute of Technology, Vellore. He published 42 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Associate Editor of the International Journal of Interactive Multimedia and Artificial Intelligence, Brazilian archives of Biology and Technology, Journal of Circuits, Systems and Computers, Computing and Informatics, International Journal of Image and Graphics and International Journal of Bio-metrics.

E. Poongothai

E. Poongothai completed her B.E. in 2011 from Anna University. She completed her M.E. and Ph.D., in computer science and engineering from Manonmaniam Sundaranar University, Tirunelveli, India, in 2013 and 2020 respectively. At present she is working as an Assistant Professor in the Department of Computer Science and Engineering, SRM University, Kattankulathur, Chennai. Her research areas are Machine Learning and Computer Vision.