Data Driven and Soil Analysis of Land Suitability For Crop Prediction Using Machine Approaches

Mr. Adil Raja
M.Tech (CSE)
Department of Computer Science And Engineering
Aliah University, Kolkata

Dr. Sayied Umer
Assistant Professor
Department of Computer Science And Engineering
Aliah University, Kolkata

HIGHLIGHTS

- Comparative analysis of Machine Learning techniques for Crop prediction.
- Performance evaluation based on only soil, only environmental characteristics and both.
- Performance analysis for classifiers using k-fold validation.
- Performance analysis using data splitting method.

Abstract: Agriculture, the backbone of every country, has been an emerging field of research, particularly inthe recent past. The soil type and environment are critical factors that drive agriculture, especially in terms ofcrop prediction. To determine which crops grow best in certain types of soil and environment, the characteristics of the latter are to be ascertained. In the past, farmers picked suitable crops for cultivation, based on first-hand experience. Today, however, identifying appropriate crops for particular areas has become a difficult proposition. The application of machine learning techniques to agriculture is an emerging field of research that helps predicts crops for easy cultivation and improved productivity. In this work, a comparative analysis is undertaken using several classifiers like the k-Nearest Neighbor (KNN), Naïve Bayes(NB), Decision Tree (DT), Support Vector Machines (SVM), Random Forests (RF) and Bagging to help suggest the most suitable cultivable crop(s), based on soil and environmental characteristics, for a specific piece of land. The algorithms are trained with the training data and subsequently tested with the soil and climate-based test dataset. The results of all the approaches are evaluated to identify the best classificationtechniques. Experimental results show that the bagging method outclasses others with respect to all performance metrics.

Keywords: agriculture; soil; environmental; crop; machine learning; classification.

INTRODUCTION

Agriculture is a unique business proposition, with crop production largely dependent on the climate and soil. Consequently, agribusiness forecasts, recognizing plant disease, and advancing pesticide use areexamined using a slew of information mining procedures prior to crop cultivation. Soil is a material asset thatimpacts land use. It is a natural resource, given the benefits it offers in terms of agricultural productivity.Minerals such as nitrogen, potassium, and phosphorus contribute to the organic composition of soil with theirspecific characteristics. Environmental factors such as the seasons, soil types, rainfall, and temperature alsogreatly impact crop cultivation. Notwithstanding the interaction between crop prediction, the environment andthe weather, semi-linear variables involve a considerable level of difficulty. Machine learning could offer aneffective alternative to crop cultivation predictions. Recommending suitable crops for a particular area is amajor concern in agriculture, and is something that can be addressed through machine learning techniques.Machine learning offers multiple methods to recognize rules and trends in large datasets, and has demonstrated a well-known predictive ability. A predictive model can be developed on its own. Unfortunately, however, machine learning approaches have so far not been applied on a large scale in the country, chieflybecause numerical and plant simulation methods are still in vogue. In machine learning, classificationtechniques [1-3] are used to predict the classes for each record in a dataset. Besides the use of advancedclassification methods in remote sensing through the use of support vector machines, random forests, androtational forests, scientists and researchers have worked to improve classification accuracy for analyzingpredictions and helping make appropriate decisions. In general, there are three types of classificationtechniques used in prediction: supervised, unsupervised and reinforcement learning. Supervised learningtrains the models with labelled soil and environmental characteristics as inputs and different crops as outputpairs. Hence, it correctly predicts the suitable crop for unknown samples which contains soil andenvironmental characteristics from testing set. Supervised learning is used to classify the category of cropswhile unsupervised learning is used to groups the similar crop called clustering. Supervised learning predictsthe target class based on current input whereas reinforcement learning sequential decision is happens; thenext input depends on the outcome of learner. Hence, compared to other two learning techniques supervisedlearning is most suitable for crop cultivate prediction. This work uses supervised learning classificationtechniques for prediction, and shows their validity and quality, alongside those of graded crop mappingmethods, following a comparative analysis. The primary contribution of this work is its attempt to find the best classification method to predict suitable crops for cultivation, based on the soil and environment.

Related work

Belson [4] described DT as models of classification and regression, developed in a tree-like architecture.A decision tree organizes a dataset in small homogeneous subsets (sub-populations), while simultaneously creating a corresponding tree map. Kohonen [5] described instance-based models (IBM) as memory-based models that learn from the learning set by contrasting new examples with instances. Bayesian models (BM)are a family of probabilistic graphical models that help research Bayesian inference. They belong to the category of supervised learning models, and are used to solve classification or regression problems. Pearl
[7] discussed the Bayesian network, and Quinlan [8] the Iterative Dichotomizer as the most common learningalgorithm in this class. Russell and Peter [9] elaborated on the NB, Gaussian Naive Bayes, and multinomial Naive Bayes.

Ensemble learning (EL) models are designed to improve the predictive quality of a given statistical learning approach or model fitting technique by constructing a linear combination of simple base learners. Breiman [10] discussed the bootstrap aggregating or bagging algorithm, while Freund and Schapire [11] proposed the Adaboost to reduce the errors of learning algorithms, and Schapire [12] implemented the boosting algorithm. Smola and coauthors [13] described the most widely used SVM algorithms, including support vector regression. By turning the original feature space into a feature space of a higher dimension, the classification capabilities of conventional SVMs are significantly enhanced using the "kernel trick". Breiman [14] described RF as a combination of tree predictors, with each tree dependent on the values of aseparately sampled random vector with the same distribution for all the trees in the forest. As the number oftrees in the forest grows, the forest generalization error converges to a limit. Cultivable crops are predicted, primarily on the basis of climatic features, giving the C4.5 algorithm an accuracy score of approximately 95% [15]. The environmental factors affecting crop yield, regions under cultivation, annual rainfall and food price indices, and defined the relationship between them. Environmental factors, coupled with algorithms like regression analysis (RA) and linear regression (LR), are used to analyze crop yields [16]. Priya and coauthors

[17] used real-time Tamil Nadu data to predict crop yields using the RF method. Jahan [18], predict the soil types based on their characteristics and fertility by using NB classifier. Galvão and coauthors [19], proposed Multiple Linear Regression (MLR) method by using corn dataset it contains soil parameters as a input. To improve the performance variable elimination method is carried out. Prasad Babu and coauthors [20], proposed a tomato crop advisory system based on soil and climate factors. This process done by ID3 algorithm and some optimization rule is applied to improve the performance. Jeong and coauthors [21], predict the crop yield using wheat and maize datasets which contains environmental factors. This predictionprocess done by using RF and MLR techniques and from the results, it is evident that the RF technique is efficient for crop yield analysis.

Motivation and justification

Several parameters impact agricultural production, and include those to do with climate (temperature, humidity and moisture), precipitation (irrigation, rainfall, and region-wise precipitation), and soil (potential of Hydrogen (pH), nutrients, organic carbon, and minerals like phosphorus, among others). Farmers stillfollow the standard practices adhered to by their ancestors. Soil characteristics in a particular region make itmost appropriate for certain crops. Repeated planting of the same crops, however, decreases soil fertility andresults in chemical accumulations that alter soil pH. The radical climatic changes characteristic of recent times can be effectively countered by the cultivation of alternative crops. The manual prediction and data collection involved in identifying suitable crops are drawbacks in agriculture. Manual prediction is affected byclimatic changes. With advances in technology, the size of the data produced is enormous, and can be usedto collect interesting trends in miscellaneous fields. The use of machine learning in agriculture helps farmersimmensely. Thus justified, we use machine learning techniques to predict suitable crops for specific areas ofland. These techniques work best when all soil types and environmental conditions are taken into consideration.

In machine learning, classification is the key to predicting the crop/s to be cultivated in specific areas. This work attempts an overview of techniques that help pick suitable crops, based on the soil and environment, using supervised learning techniques like the kNN, NB, DT, SVM, RF and bagging for crop prediction. Each algorithm has its pros and cons. The kNN does not work well with imbalanced data but resolves multi-class problems. The NB is very fast and can be used in real-time predictions, though each feature makes independent assumptions about the outcome. Data normalization is not needed in the DT, which is most data-sensitive. A slight change in the data is enough to change the outcome entirely. The SVMdoes not work well on overlapping classes, but has little impact on outliers. The RF handles errors in imbalanced data, but results in high computational costs while training a large number of deep trees. Bagging works well on high-dimensional data, and its performance is not affected by missing values in the dataset. However, it introduces a certain level of difficulty in the form of a loss of interpretability with regard to the model used. Since each classifier carries out prediction in its own unique way, it is essential to find the mostaccurate classifier for crop prediction. Motivated by the facts above, this work focuses on finding the best classifier for crop prediction so as to maximize production.

Outline of the work

Figure 1 depicts the overall process of this work. First, input data is preprocessed to find missing values,eliminates redundant data, and standardize the dataset. Next, the preprocessed data are subject to several classification techniques to determine the most suitable crops for a particular stretch of land. Prior to applyingthe classification techniques, the dataset is split into training and testing phases. Samples from the training dataset are used to train the classification algorithm to find the crop/s ideally suited to cultivation in a specificarea. The unknown data from the testing dataset are given to the trained classifier to predict a suitable crop, following which the results are evaluated using different performance metrics. An analysis is undertaken to obtain the best classification method. Information on the predicted or recommended crop/s to be grown can be provided to the farming community, based on the results obtained.

Organization of the paper

The remaining part of the paper is organized as follows: Section II gives methodology for crop prediction.
Section III discusses the experimental results followed by conclusion.

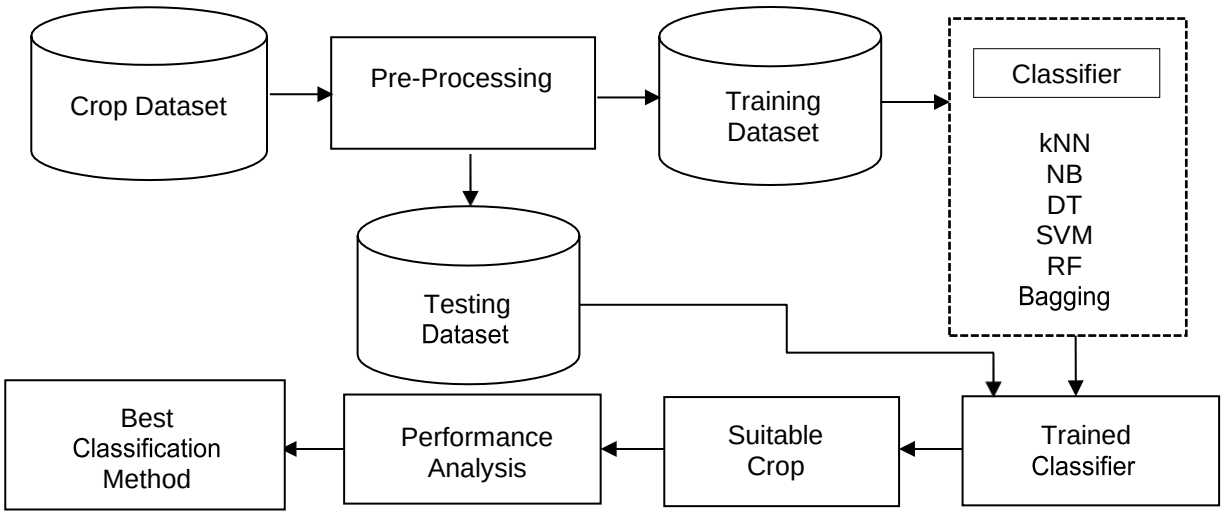

Figure 1. Outline of the Work.

MATERIAL AND METHODS

Background study

Predicting crops for cultivation enables agricultural departments to put in place strategies for improvement. Crop prediction is based on factors such as the climate, geography, genetics, politics and economics. Risks related to these variables can be quantified if the appropriate computational or quantitativemethodologies are implemented. Bootstrap aggregating (bagging) is a meta-algorithm for machine learning that enhances the consistency and precision of the algorithms used through statistical classification and regression. It also significantly reduces variability and prevents overfitting [22]. As noted earlier, preprocessing techniques can easily be incorporated into the learning algorithms that constitute the ensemble. Because of their simplicity and strong generalization potential, several methods have been developed using bagging ensembles to fix class disparity issues. This section compares different existing classification techniques and identifies the best for crop prediction.

K Nearest Neighbor

The kNN is a non-complex algorithm that predicts suitable crop based on certain similarity measures [23]. The closeness measure is calculated by distance measures like the Euclidean distance and Manhattandistance [24]. In this work, Euclidean distance is used to find the shortest distance between training and testing samples. Top nearest class is taken and that class is assign as suitable crop for cultivation. In the kNN, feature vectors are stored in the training phase of the algorithm. The class labels of the training samplesand target class of crops are classified by assigning the most frequent label of the nearest training samples.To validate the model, k cross fold (10 folds) validation is used. Fit the model using k-1 fold and validate themodel using k^{th} fold. Figure 2 depicts the work flow of kNN.

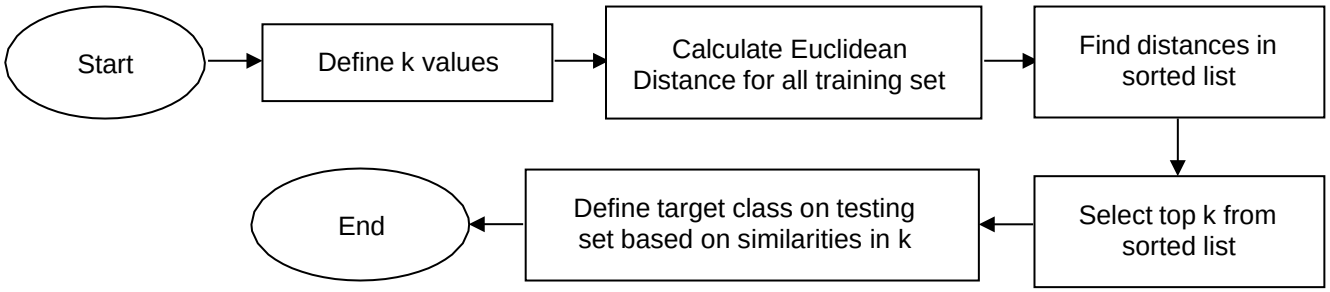

Figure 2. Flow diagram of KNN.

Naive Bayes

The Naïve Bayes technique assigns class labels to problem instances for constructing classifier models[23], and is based on Bayes' theorem [18]. The NB is not a single algorithm for training a classifier but a familyof algorithms based on common principles. It assumes that the value of a particular feature is independent of the value of any other quality, given the class variable [23]. It works based on probability theory and it choose the suitable crop from testing samples which has the maximum probability. The potential of NB classifier for crop prediction is evaluated using k cross fold validation method. The crop dataset is split into two subsets where k-1 fold is used to train the model and k^{th} model is used for validate the model. Figure 3 shows the NB flow diagram of crop prediction process.

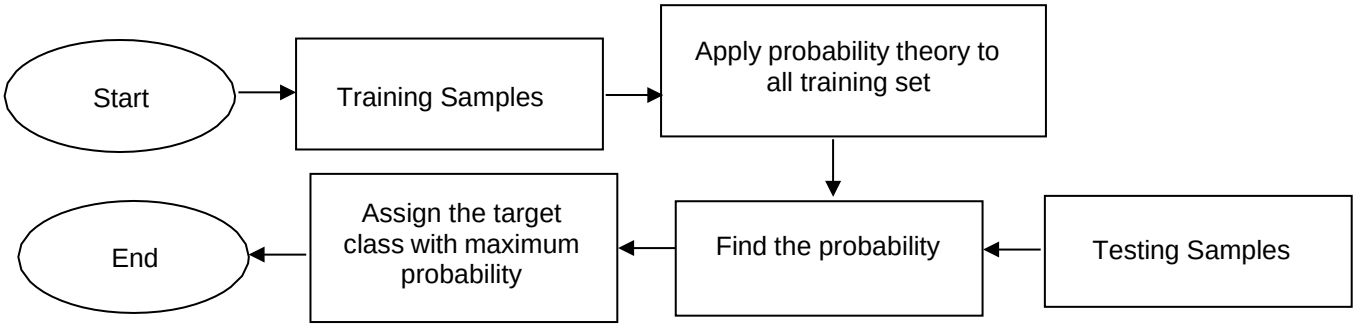

Figure 3. Flow diagram of NB.

Decision Tree

The decision tree is a single tree predictive model that is based on the data structure of the tree [18]. A tree consists of decision nodes and decision leaves [24]. Each split is labeled with an input feature and leaf as a target class that is crop. It executes a top-down approach by choosing a value for the variable at each stop that best splits a set of items [25], depending on the application and makes the decision to find the suitable crop for cultivation. The aptness of this technique for crop cultivates prediction examined by the k cross fold validation method. The samples are split into k and k-1 subsample. The sample k is used for testingthe model and k-1 samples are used to train the model. Figure 4 illustrates the DT work flow of prediction process.

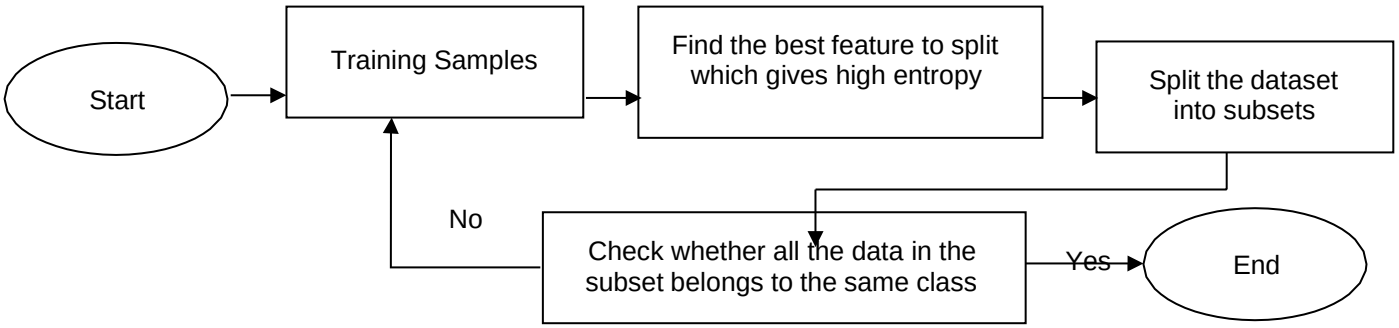

Figure 4. Flow diagram of DT.

Support Vector Machine

The SVM is a supervised machine learning algorithm which breaks data into decision surfaces. The decision surfaces further divide the data into two hyperplane groups [26]. The training points specify the vector which supports the hyperplane. This hyperplane is used for crop prediction process. The crop that liesnearby the surface is urging for cultivation. Further, the ability of this technique is examined using k- cross fold validation process. In this work, 10 fold is used for validation where k-1 folds are used to fit the model forcrop prediction and k^{th} fold is used to test the model. Figure 5 represents the SVM work flow for crop prediction.

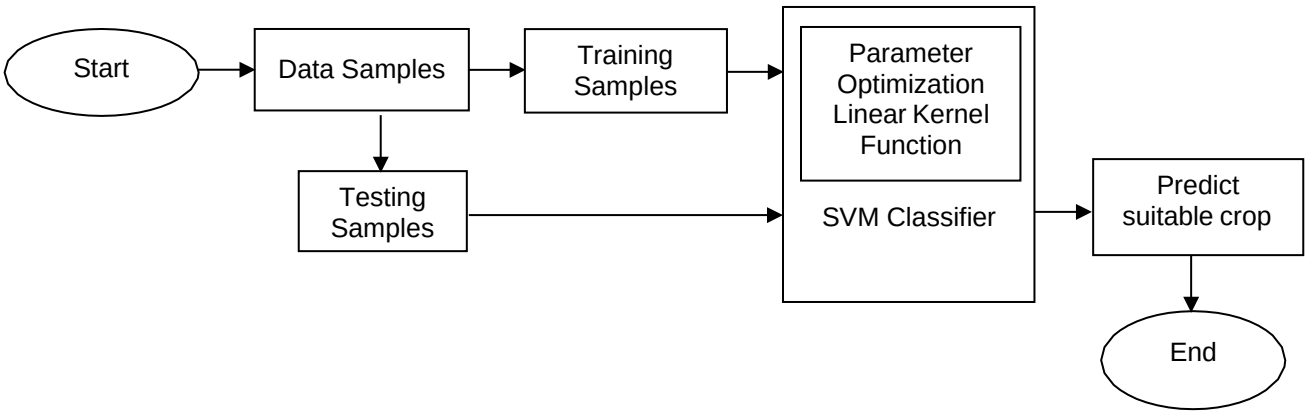

Figure 5. Flow diagram of SVM.

Random Forest

The random forest is a popular and powerful supervised machine learning algorithm that resolves both classification and regression problems [17]. The RF is a multiple tree which includes a large number of individual decision trees. To decide the suitable crop of a test samples, it aggregates votes from different decision trees and based on the results it recommend the suitable crop. Additional, this technique is evaluatedby k- cross fold method for predicting the suitable crop. The dataset samples are divided into two sub samplesthen fit the model using k-1 sample and test the model using k^{th} sample. Figure 6 depicts the work flow of RF.

Bagging

Bagging, also known as bootstrap aggregating, was introduced by Breiman [10], and is used to train andcombine multiple copies of a learning algorithm [23]. It improves the stability of the learning algorithm and enhances the results of the prediction algorithm [22]. Bagging splits the training samples as a sub samples to train the model for crop prediction. It takes the votes from each sub sample to predict the suitable crop from testing dataset. In this work, Adaptive Bagging (AdaBag) is used for prediction process. Since baggingdoes not permit weight recalculation, there is no need to change the weight update equation or modify the algorithm's calculations. To estimate the accuracy of bagging technique for crop prediction k- cross fold validation is used. For this process the dataset is split into two subsamples as k-1 and k samples. To train the model k-1 sub samples are used and k^{th} sample is used to validate the model. The work flow of baggingis given in Figure 7.

Crop prediction procedure

The algorithm for crop prediction is given below. The soil and environmental parameters are given as inputs, and a suitable crop is the output.

Algorithm

Step 1: Import a set of data.

Step 2: Preprocess the data to find the missing values and replicas for standardizing the data. Usingpreprocessing, it converts target variables into factor variables.

Step 3: Split the preprocessed data to be used in the training and testing datasets.

Training Phase

Step 4: Take 70% of the samples from the training dataset as training samples.Step 5:

Apply the classification algorithm to the training samples.

Step 6: Train the classification algorithm well with the training dataset to find a suitable crop.

Testing Phase

Step 7: Take 30% of the samples from the testing dataset as testing samples.

Step 8: Apply the trained classifier to all the testing samples used to identify a suitable crop for cultivationin a particular patch of land.

Step 9: The trained classifier finds the target label for new instances to predict a suitable crop.Step 10: Finally, the result recommends a suitable crop for cultivation.

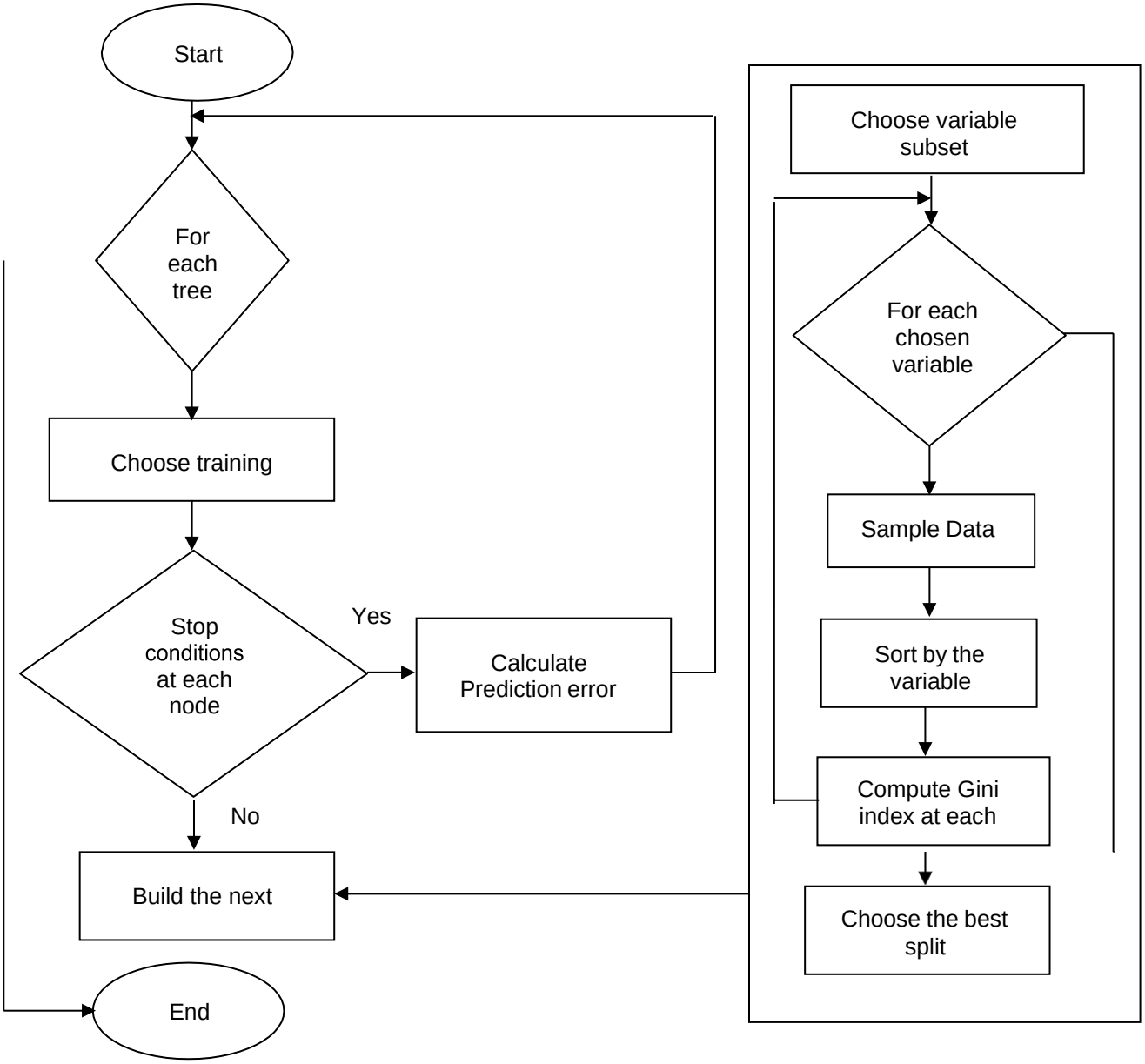

Figure 6. Flow diagram of RF.

RESULTS AND DISCUSSIONS

Dataset Description

This work utilizes an agricultural dataset that includes soil characteristics and environmental factors, collected from the Agricultural Department of Sankarankovil Taluk, Tenkasi District, Tamil Nadu, India. The dataset contains 1000 instances and 16 attributes, where 12 attributes are soil characteristics and the remaining 4 environmental. In this work, the 9 crops used for the prediction process include paddy, maize, black gram, green gram, 7rinja gram, 7rinjal, lady's finger, tomato and chickpea. The data are collected fromvarious villages in and around Sankarankovil.

Table 1 presents information on the soil type, a brief description of the soil, and the environmental attributes impacting crop prediction.

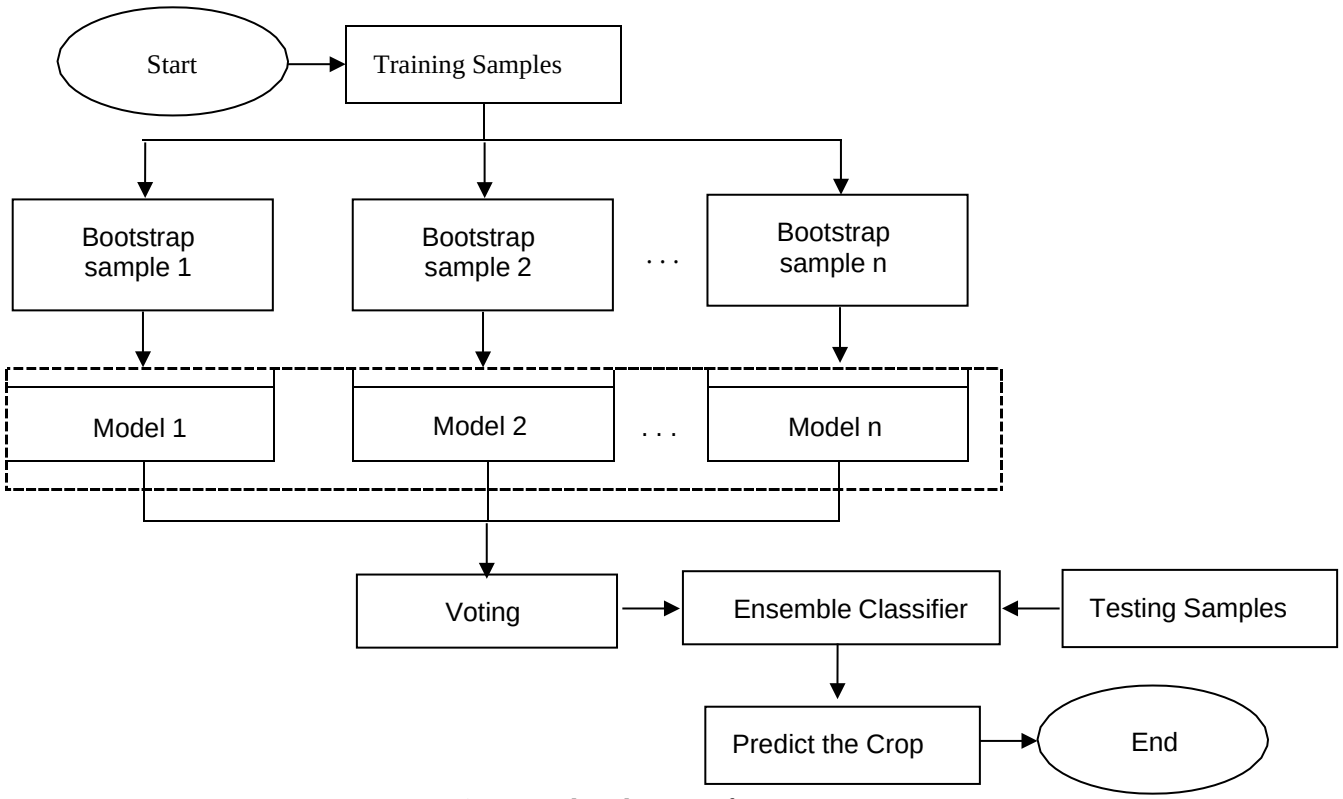

Figure 7. Flow diagram of Bagging.

Table 1. Dataset Description of Crop Dataset.

S.No.	Attributes	Type	Description
			Soil characteristics
1.	pH (potential of Hydrogen)	Numeric	pH is the main factor for farming
2.	EC (Electrical Conductivity)	Numeric	That affect crop productivity if EC is 0.01 that soil is good soil
3.	OC (Organic Carbon)	Numeric	OC enters the soil through the decomposition of plant and animal residues, rootexudates, living and dead microorganisms, and soil biota.
4.	N (Nitrogen)	Numeric	Nitrogen is a key element in plant growth.
5.	P (Phosphorus)	Numeric	Phosphorus helps transfer energy from sunlight to plants, stimulates early rootand plant growth, and hastens maturity.
6.	K(Potassium)	Numeric	Potassium increases vigour and disease resistance of plants, helps form andmove starches, sugars and oils in plants, and can improve fruit quality.
7.	S (Sulphur)	Numeric	Sulphur is a constituent of amino acids in plant proteins and is involved in energy-producing processes in plants.
8.	Z (Zinc)	Numeric	Zinc helps in the production of a plant hormone responsible for stem elongationand leaf expansion.
9.	B (Boron)	Numeric	Boron helps with the formation of cell walls in rapidly growing tissue. Deficiency reduces the uptake of calcium and inhibits the plant's ability to use it.
10.	Fe (Iron)	Numeric	Iron is a constituent of many compounds that regulate and promote growth
11.	Mn (Manganese)	Numeric	Manganese helps with photosynthesis.
12.	Cu (Copper)	Numeric	Copper is an essential constituent of enzymes in plants

Table 1- (cont.)

		Environmental Factors	
13.	Texture	Integer	It has major influence on crop growth. It influences aeration, water movement etc.
14.	Season	Integer	Season is the challenging factor for crop growth.
15.	Rainfall	Numeric	Rainfall has the great impact on crop growth. Excessive and insufficient rainfallaffects the yield.
16.	Average Temperature	Integer	It is important for growth and development

Table 1 presents information on the soil type, a brief description of the soil, and the environmental attributes impacting crop prediction.

Performance Metrics

The performance of crop prediction is measured using the following performance metrics. The formulae,and a description of each used in the experimental analysis, are given in Table 2.

Table 2. Performance Metrics Description

S. No	Metric	Formula	Range	Description
1.	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	From 0 to 1. Value close to 1 shows the better prediction	The accuracy rate has historically been the most commonly used statistical indicator.
2.	Kappa	$\frac{P_{agree} - P_{chance}1}{- P_{chance}}$	-1 to 1. 1 reflects classificationis significantly better than random; less than 1 reflects no better than random.	It is a measure of agreement betweentwo individuals.
3.	Precision	$\frac{TP}{TP + FP}$	0 to 1. Value nearby 1 denoteshas less false positiveprediction.	The number of true positive predictions divided by the total number of positive predictions is determined as precision.
4.	Recall	$\frac{TP}{TP + FN}$	0 to 1. Value close by 1 meanshas less false negative prediction	Analyzes of recall can help to examinethe main sources of uncertainty of model prediction..
5.	Specificity	$\frac{TN}{TN + FP}$	0 to 1. Value close to 1 indicates has less negative prediction	The proportion of negative results out of the number of samples which were actually negative.
6.	F1 Score	$2 * \frac{PPV.TPR}{PPV+TPR}$	0 to 1. Value close 1 implies has better precision and recall; Value 0 means worst precision and recall	The accuracy of the measurement test and is known as the weighted harmonic mean of the tests ' precisionand recall.

Where TP- True Positive; TN- True Negative; FP- False Positive; FN- False Negative; P_{agree}- Probability of agreement;P_{chance}- Probability of agreement due to chance; PPV- Positive Predicted Value; TPR- True Positive Rate.

RESULTS AND DISCUSSION

This section compares several classification techniques for crop prediction, based on the soil and environmental conditions of a particular land area, using the performance metrics of accuracy, kappa, precision, recall, specificity and F1 score.

Performance comparison of Classification techniques based on Soil Characteristics

Table 3 shows a performance evaluation of classification techniques, based only on the soil conditions discussed in Table 1.

Table 3. Performance comparison of classification methods based on soil conditions.

Classifiers	Performance Metrics (%)					
	Accuracy	**Kappa**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
kNN	0.5625	0.4939	0.6763	0.9438	0.6797	0.6783
NB	0.6812	0.6243	0.7494	0.9577	0.7467	0.7481
DT	0.7125	0.6637	0.7935	0.9622	0.7573	0.7643
SVM	0.7750	0.7362	0.8242	0.9706	0.7853	0.8042
RF	0.8375	0.8092	0.8434	0.9789	0.8420	0.8427
Bagging	**0.8875**	**0.8686**	**0.9108**	**0.9790**	**0.8597**	**0.8845**

Table 3 shows that bagging finds more accurate cultivable crops, based on soil characteristics, than other techniques. Further, bagging takes votes for each sample for improved performance, based on which it offers better crop prediction accuracy than other methods.

Performance comparison of Classification techniques based on Environmental Conditions

Table 4 represents a performance analysis of classification methods, based only on environmental factors such as texture, season, rainfall and average temperature.

Table 4. Performance comparison of classification methods based on environmental conditions

Classifiers	Performance Metrics (%)					
	Accuracy	**Kappa**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
kNN	0.3500	0.2151	0.8006	0.8804	0.3673	0.5036
NB	0.3512	0.2457	0.8198	0.9017	0.3875	0.5263
DT	0.4435	0.3089	0.8811	0.9154	0.4084	0.5581
SVM	0.4500	0.3433	0.8958	0.9166	0.4516	0.6005
RF	0.4625	0.3612	0.9138	0.9208	0.5017	0.6487
Bagging	**0.5400**	**0.3835**	**0.9176**	**0.9266**	**0.5184**	**0.6662**

Table 4 shows that bagging selects cultivable crops, based on environmental characteristics, more accurately than other techniques. In addition, bagging is a homogenous ensemble method; in this work decision tree is used for ensemble technique. It splits the whole dataset which contains soil and environmentalcharacteristics as sub samples. The different sub samples were separately trained with the single decision tree model and each model predicts the suitable crop. Finally, the outcomes of each model are combined byvoting techniques to produce single result for crop cultivation.

Performance comparison of Classification techniques based on Soil and Environmental Characteristics

Table 5 represents a performance analysis of classification techniques, based on factors such as the soiland environment, following a comparison of them all.

Table 5. Performance comparison of classification methods based on Soil and Environmental conditions.

Classifier	Performance Metrics (%)					
	Accuracy	**Kappa**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
kNN	0.6	0.5804	0.7023	0.9473	0.6909	0.6965
NB	0.7854	0.7749	0.8636	0.9584	0.8261	0.8444
DT	0.8188	0.8088	0.8687	0.9647	0.8528	0.8607
SVM	0.8312	0.8456	0.8937	0.9712	0.8743	0.8838
RF	0.8875	0.8681	0.919	0.9796	0.9133	0.9161
Bagging	**0.9062**	**0.8901**	**0.9257**	**0.9878**	**0.9255**	**0.9256**

Table 5 infers that bagging produces more accurate results than the others, based on both soil and environmental characteristics. The variance of an estimate is reduced considerably by the bagging technique,using its aggregation procedure. Consequently, it has better crop prediction accuracy than other methods.

Table 3, Table 4 and Table 5 infer that the bagging technique has the best crop prediction accuracy, based on both soil and environmental characteristics, compared to only on soil characteristics and only on environmental factors.

Performance evaluation of Classification techniques using k-fold validation

To validate the performance of classification techniques for crop prediction, the fold variation method is used. Table 6 shows a performance evaluation of classification techniques to find the most suitable crop fora particular land area, based on various cross-fold validations to obtain the best fold of all the classification methods. The fold ranges vary from 10 to 90.

Table 6. Performance of the Classification methods based on fold variation

Classifier	Folds	Performance Metrics (%)					
		Accuracy	**Kappa**	**Precision**	**Recall**	**Sensitivity**	**F1 Score**
Bagging	**10**	**90.62**	**89.01**	**92.55**	**92.57**	**98.78**	**92.56**
	20	89.3	87.69	91.23	91.25	97.46	91.24
	30	87	85.38	88.92	88.94	95.15	88.93
	40	88.32	86.4	89.94	89.96	96.17	89.95
	50	87.8	85.88	89.42	89.44	95.65	89.43
	60	86.5	84.58	88.12	88.14	94.35	88.13
	70	87.2	85.18	88.72	88.74	94.95	88.73
	80	87.77	85.75	89.29	89.31	95.52	89.3
	90	86.3	85.42	88.96	88.98	95.19	88.97
RF	**10**	**88.65**	**86.81**	**91.33**	**91.9**	**97.96**	**91.61**
	20	87.93	85.49	90.01	90.58	96.64	90.29
	30	85.63	83.18	87.7	88.27	94.33	87.98
	40	86.95	84.2	88.72	89.29	95.35	89
	50	86.43	83.68	88.2	88.77	94.83	88.48
	60	85.13	82.38	86.9	87.47	93.53	87.18
	70	85.83	82.98	87.5	88.07	94.13	87.78
	80	86.4	83.55	88.07	88.64	94.7	88.35
	90	84.93	83.22	87.74	88.31	94.37	88.02
SVM	**10**	**83.12**	**84.56**	**87.43**	**89.37**	**97.12**	**88.38**
	20	81.8	83.24	86.11	88.05	95.8	87.06
	30	79.5	80.93	83.8	85.74	93.49	84.75
	40	80.82	81.95	84.82	86.76	94.51	85.77
	50	80.3	81.43	84.3	86.24	93.99	85.25
	60	79	80.13	83	84.94	92.69	83.95
	70	79.7	80.73	83.6	85.54	93.29	84.55
	80	80.27	81.3	84.17	86.11	93.86	85.12
	90	78.8	80.97	83.84	85.78	93.53	84.79
DT	**10**	**81.88**	**80.88**	**85.28**	**86.87**	**96.47**	**86.06**
	20	80.56	79.56	83.96	85.55	95.15	84.74
	30	78.26	77.25	81.65	83.24	92.84	82.43
	40	79.58	78.27	82.67	84.26	93.86	83.45
	50	79.06	77.75	82.15	83.74	93.34	82.93
	60	77.76	76.45	80.85	82.44	92.04	81.63
	70	78.46	77.05	81.45	83.04	92.64	82.23
	80	79.03	77.62	82.02	83.61	93.21	82.8
	90	77.56	77.29	81.69	83.28	92.88	82.47
NB	**10**	**78.54**	**77.49**	**82.61**	**86.36**	**95.84**	**84.44**
	20	77.22	76.17	81.29	85.04	94.52	83.12
	30	74.92	73.86	78.98	82.73	92.21	80.81
	40	76.24	74.88	80	83.75	93.23	81.83
	50	75.72	74.36	79.48	83.23	92.71	81.31
	60	74.42	73.06	78.18	81.93	91.41	80.01
	70	75.12	73.66	78.78	82.53	92.01	80.61
	80	75.69	74.23	79.35	83.1	92.58	81.18
	90	74.22	73.9	79.02	82.77	92.25	80.85
kNN	**10**	**60**	**58.04**	**69.09**	**70.23**	**94.73**	**69.65**
	20	58.68	56.75	67.77	68.91	93.41	68.33
	30	56.38	54.44	65.46	66.6	91.1	66.02
	40	57.7	55.46	66.48	67.62	92.12	67.04
	50	57.18	54.94	65.96	67.1	91.6	66.52
	60	55.88	53.64	64.66	65.8	90.3	65.22
	70	56.58	54.24	65.26	66.4	90.9	65.82
	80	57.15	54.81	65.83	66.97	91.47	66.39
	90	55.68	54.48	65.5	66.64	91.14	66.06

Table 6 above infers that classification techniques perform best in 10-fold-based cross-fold validation interms of accuracy, kappa, precision, recall, sensitivity and f1 score. Table 6 clearly shows that the bagging classifier outperforms other 10-fold-based methods.

Performance evaluation of classification techniques using data splitting validation

To validate the crop prediction performance, a validation method termed data splitting is used. The following graphical representation below shows a performance evaluation of classification techniques for finding suitable crops for a particular land area, based on data splitting, to get the best training and testing splitting ranges. The ranges vary from between 25% - 75% and 75% - 25%. Performances are evaluated using the metrics of accuracy, kappa, precision, recall, specificity and F1 score.

The Figure 8 shows that the bagging classifier works better in the 70% - 30% splitting range than other splitting ranges, based on the metrics mentioned in Table 2.

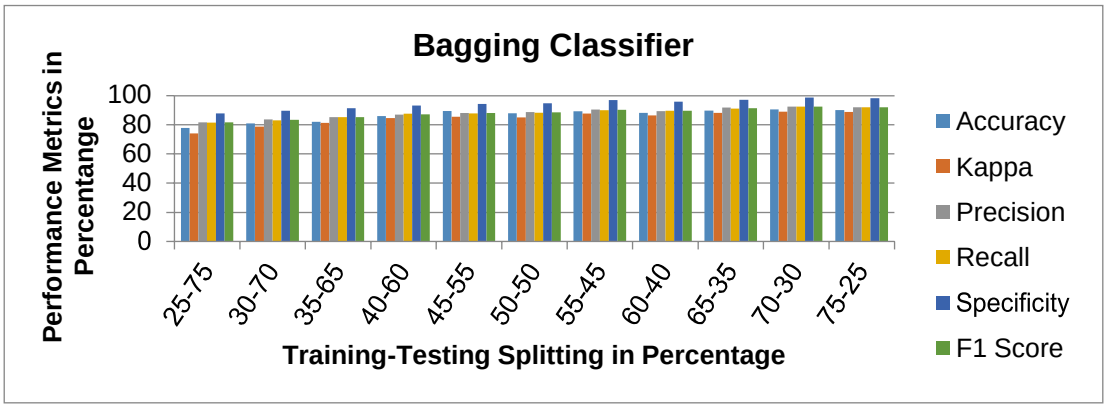

Figure 8. Performance evaluations of Bagging classifier using data splitting method

Figure 9 presents a performance evaluation of the RF classifier, based on several metrics. The RF classification technique works better in the 70% - 30% data splitting range than other splitting ranges. Figure10 shows a performance evaluation of the SVM classification method, with its prediction accuracy down fromthe RF and bagging. From the results, it is evident that the SVM classification technique performs better in the 70% - 30% data splitting range. Figure 11 clearly shows that the DT works better in the 70% - 30% rangethan other splitting ranges. The experimental results reveal that the decision tree classifier does not outperform the SVM, RF and bagging algorithms. Figure 12 depicts that the DT, SVM, RF and bagging algorithms outperform the NB classifier. The NB classifier work well in the 70%-30% range than other data splitting ranges. From figure 13, it is evident that the kNN classifier performs better with the 70% - 30% data splitting range than other splitting ranges. Further, the kNN technique has the least prediction accuracy of allthe techniques.

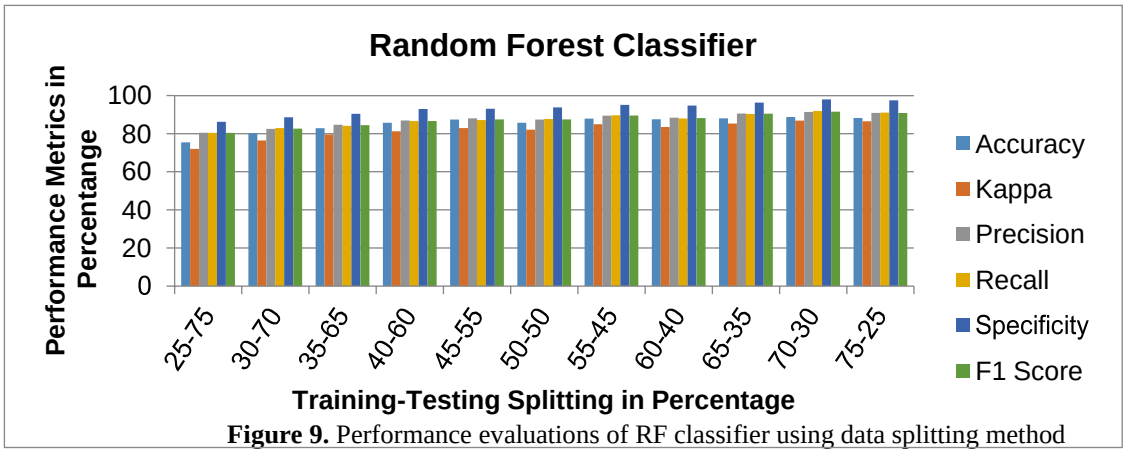

Figure 9. Performance evaluations of RF classifier using data splitting method

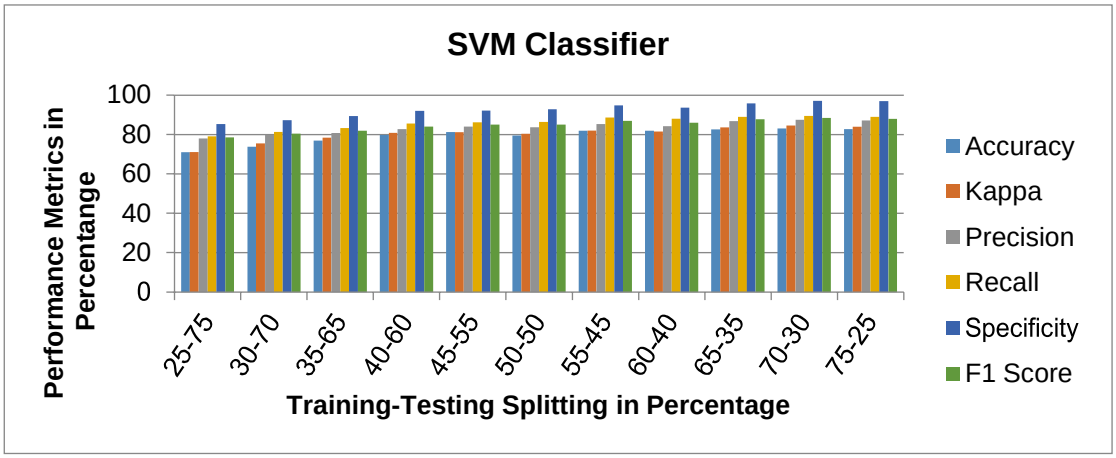

Figure 10. Performance evaluations of SVM classifier using data splitting method

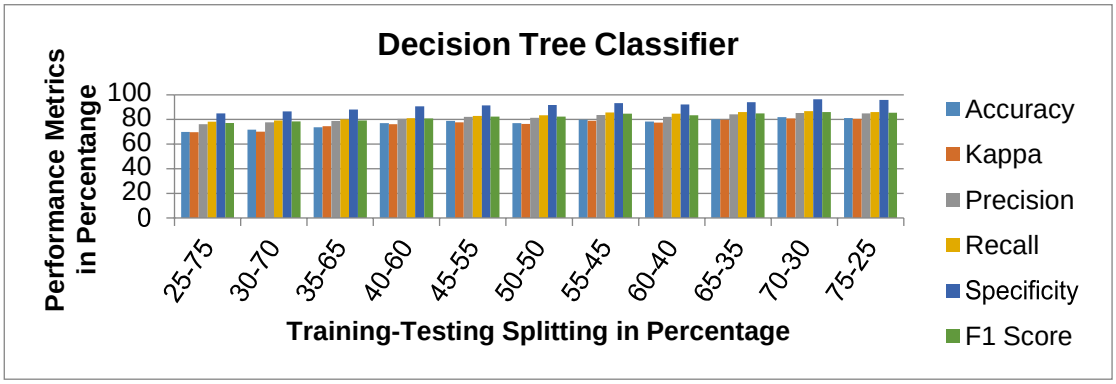

Figure 11. Performance evaluations of DT classifier using data splitting method

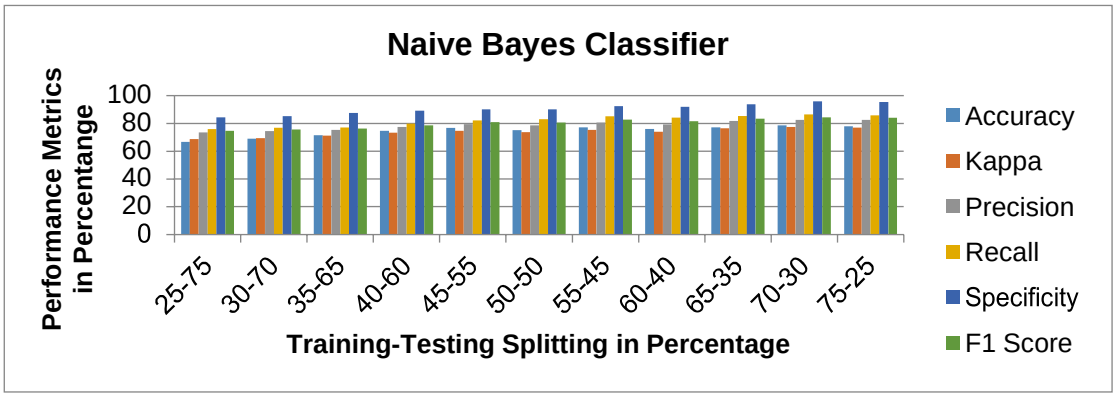

Figure 12. Performance evaluations of NB classifier using data splitting method

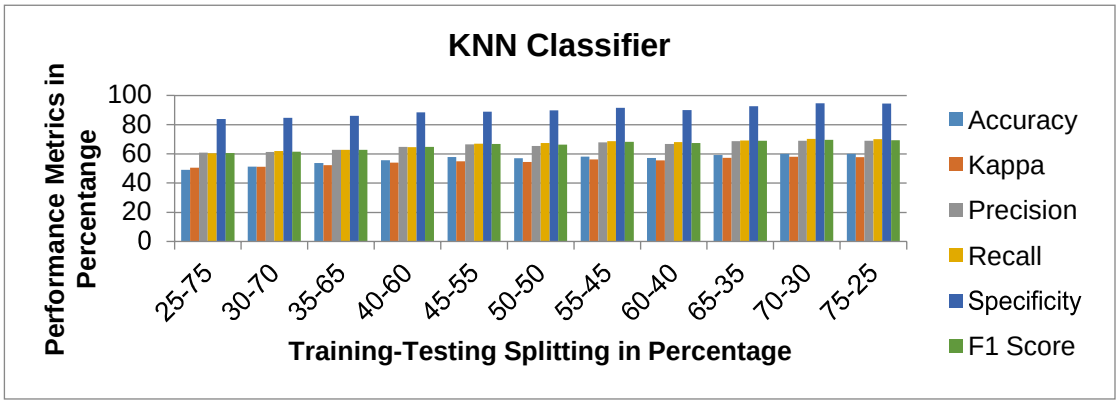

Figure 13. Performance evaluations of kNN classifier using data splitting method

The figures mentioned above reveal that all the classifiers perform much better with the 70%-30% data splitting range as training and testing ranges. The bagging classifier makes the best predictions, compared to other methods.

CONCLUSION

This work presents a comparative analysis of classification approaches such as the kNN, NB, DT, SVM,RF and bagging to predict suitable crop/s for particular land areas. The results are compared with respect toperformance metrics like accuracy, kappa, sensitivity, specificity, precision and F1-score. Owing to the use of multiple learning algorithms, the bagging algorithm offers better predictions than other algorithms, based on the soil and environmental conditions observed from the experimental results. The algorithms above onlyprovide guidelines for suitable crops for specific areas of land. Future directions include suggestions on fertilizer use for crops, as well as recommendations on alternative crops for arable land.

Acknowledgments: We would like to thank Department of Agriculture Government of Bihar, Gopalganj District, India for providing data for the analysis.

REFERENCES

1. Duda, Richard O, Hart, Peter E and Stork, David G. Pattern classification and scene analysis. New York: Wiley,1973; 3: 731-9.
2. Breiman L, Friedman J, Charles J S and Richard A. Olshen. Classification and regression trees. CRC press, 1984.
3. Richard E. Neapolitan, Models for reasoning under uncertainty. Applied Artifical Intelligence, 1987;1(4):337–66.
4. Belson WA. Matching and prediction on the principle of biological classification. J. R. Stat. Soc. Ser. C. Appl. Stat., 1959;8(2):65-75.
5. Kohonen T, Learning vector quantization. Neural Network, 1988;1:303.
6. Atkeson, Christopher G, Moore, Andrew W and Schaal S. Locally weighted learning. In Lazy learning, Springer, Dordrecht, 1997:11-73.
7. Pearl J, Probabilistic Reasoning in Intelligent Systems. Morgan Kauffmann Publishers Inc. San Francisco, CA,USA, 1988; 552.
8. Quinlan JR, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA,1993: 235-40.
9. Russell, Stuart J and Peter Norvig, Artificial Intelligence: A Modern Approach. Prentice Hall: Upper Saddle River,New Jersey, USA, 1995;9.
10. Breiman L, Bagging Predictors. Machine Learning, 1996;24:123–40.
11. Freund Y and Schapire RE. Experiments with a new boosting algorithm. In International Conference on Machine Learning, 1996;96:148-56.
12. Schapire RE., A brief introduction to boosting. Proceedings of sixteenth In International Joint Conference on Artificial Intelligence, 1999;99:1401-6.
13. Smola A, Burges C, Drucker H, Golowich S, Hemmen LV, Muller Klaus-Robert, Bernhard Scholkopt et al. Regression Estimation with Support Vector Learning Machines. Master's Thesis, The Technical University of Munchen, Germany, 1996:1–78.
14. Breiman L, Random forests. Machine learning, 2001;45(1):5-32.
15. Veenadhari S, Bharat Misra, and Singh C D. Machine learning approach for forecasting crop yield based on climaticparameters. In 2014 International Conference on Computer Communication and Informatics, IEEE, 2014: 1-5.
16. Sellam V and Poovammal E. Prediction of crop yield using regression analysis. Indian J. Sci. Technol., 2016;9(38):1-5.
17. Priya P, Muthaiah U and Balamurugan M. Predicting yield of the crop using machine learning algorithm. Int. J. Eng.Sci. Res. Technol., 2018;7(1):1-7.
18. Jahan R. Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land soils. Int.J. Eng. Sci. Res. Technol., 2018; 6 (5): 189-93.
19. Galvao RKH, Araujo MCU, Fragoso WD, Silva EC, Gledson EJ, Soares SFC et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. Chemometr. Intell. Lab. Syst.,2008; 92 (1): 83-91.
20. Prasad Babu MS, Ramana Murty, NV, and Narayana SVNL. A web based tomato crop expert information system based on artificial intelligence and machine learning algorithms. International Journal of Computer Science and Information Technologies, 2010; 1 (3): 6–15.

21. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, Timlin DJ et al. Random forests for global and regional crop yield predictions. PLoS One, 2016; 11 (6).
22. Zala, Dipika H, and Chaudhri, MB. Review on use of BAGGING technique in agriculture crop yield prediction, International Journal for Scientific Research & Development, 2018;6(8):675-7.
23. Pudumalar, S, Ramanujam E, Harine Rajashree R, Kavya C, Kiruthika T and Nisha J. Crop recommendation system for precision agriculture. In 2016 Eighth International Conference on Advanced Computing (ICoAC), IEEE,2017:32-6.
24. Anantha Reddy D, Bhagyashri D and Watekar A. Crop Recommendation System to Maximize Crop Yield in Ramtek region using Machine Learning. Int. J. Sci. Res. Sci. Technol. 2019;6(1):485-9.
25. Balducci F, Impedovo D and Pirlo G. Machine learning applications on agricultural datasets for smart farm enhancement. Machines, 2018;6(3):38-59.
26. Suykens, JAK and Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett,1999;9(3):293-300.