International Journal of Information Communication Technologies and Human Development
Volume 14 • Issue 1

Mining Agricultural Data to Predict Soil Fertility Using Ensemble Boosting Algorithm

Jayalakshmi R, Sri Vidya Mandir Arts and Science College, India
Savitha Devi M., Periyar University Constituent College of Arts and Science, India

ABSTRACT

Agriculture is the most important resource of livelihood and an emerging field the forms the backbone of India. Present challenges of the agriculture domain include uncertain climatic changes, poor irrigation facilities, weather uncertainty. Machine learning is one such technique that is employed to predict the fertility of the soil in agriculture. ensemble machine learning techniques aim to create meta-classifiers to produce better predictive performance. The primary focus of this paper is to analyze the soil data that is collected from the soil testing laboratory to predict fertility from a collected dataset by using multiple ensemble machine learning algorithms such as bagging, boosting, and stacking for better prediction, accuracy, and higher consistency. The soil fertility classes were evaluated using 10 selected attributes. Measurements of different soil parameters have been used for predicting soil fertility. The experimental result shows that the boosting method on the C5.0 algorithm achieved higher accuracy than other ensemble classifiers with 98.15%.

KEYWORDS

Agriculture, Bagging, Boosting, Ensemble Learning, Soil Fertility

1. INTRODUCTION

Machine learning in agriculture is an attractive novel research area. Agriculture data are extremely diversified in terms of environment, climate, weather condition, interdependency, and use of sources for farming. The main problem of using Machine Learning in agriculture is to solve issues based on the available data and its meaningful outcomes (Patel & Kathiriya, 2017). Agriculture plays a vital role in the Indian economy and the production of crops. Crops may be either commercial crops or food crops. Food crops include rice paddy, maize, wheat, grams, millets, etc., while commercial crops are cotton, sugarcane, groundnut, cashew, etc. The productivity of the crops is drastically influenced by weather conditions (Palanivel & Surianarayanan, 2019). It is necessary to cultivate the soil properly for maintaining fertility, achieve better yield and protect the environment. Present challenges of the

DOI: 10.4018/IJICTHD.299414

agriculture domain are uncertain climatic changes, poor irrigation facilities, weather uncertainty, water shortages, reduction in soil fertility, and uncontrolled cost due to demand-supply impose farmers to be equipped with smart farming. A soil test is the learning of a soil sample to ascertain an additional matter, its composition, and various attributes. Generally, soil tests are accomplished to determine the wealth (Bhuyar, 2014). Ensemble methods are meta-algorithms that combine several machine learning techniques into predictive models like bagging to decrease variance, boosting to reduce bias, and stacking for improving predictions. This article presents the evaluation of three ensemble classifiers on the soil datasets including bagging, boosting, and stacking. These classifiers are tested to achieve the best accuracy of the dataset. Bagging is an ensemble classifier that creates separate samples of the training datasets and creates a classifier for each sample. The results of these multiple classifiers are then combined such as averaged or majority voting. Boosting is an ensemble method that starts with a base classifier that is arranged on the training data. The second classifier is created behind the first classifier to focus on the instances in the training data. The process continues to append classifiers in anticipation of a limit is reached in the number of models or accuracy. The multiple different algorithms are prepared on the training data and a meta-classifier is prepared, which learns how to take the predictions of each classifier and make accurate predictions on unknown data as stacking. This research work utilizes the sample data that have been collected from the Vellore soil testing laboratory. In the beginning, the data has been preprocessed and then appropriate features are extracted using the feature extraction method then passed onto the various training ensemble classifiers of machine learning to acquire a better result.

1.1 Objectives

1. Investigating the important factors and selecting the high dominant features that affect the prediction of soil fertility.
2. Identifying the key factors that associate with Soil fertility.
3. Developing models to classify the fertility as Ideal or Not Ideal.

1.2 Outline of the Paper

This proposed work has been organized as follows: Section 2 describes the characteristics of soil and soil fertility. Related works on Machine learning and ensemble classifiers are explained in Section 3. Section 4 gives the Materials and Methods used. The Proposed methodology with the research framework is outlined in section 5. Section 6 describes Experimental results and discussion, and Section 7 concludes research work with future directions.

2. SOIL AND SOIL FERTILITY

Soil is crucial for plant life which consists of solids like minerals and organic matter, liquids like water, solutes, gases mainly oxygen, carbon dioxide and contains living organisms. All these elements afford their physical and chemical properties. Soil fertility refers to the capability of soil for sustaining plant growth and supply essential plant nutrients and water in a sufficient way to provide plant habitat and result in sustained and consistent yields of high quality. Soil fertility prediction plays a vital role in agriculture particularly food production and also important factors that have a direct impact on crop yield and quality. Soil fertility defines growth of plants when other environmental factors like light, water, temperature are favorable. Soil fertility is influence by numerous factors like irrigation (Soil water), Climate, Soil, Soil alkalinity, acidity, Nutrition in Soil. The quality of soil in India is decreased by globalization, urbanization, changing weather condition, higher use of pesticides. Lack in agricultural production and eventually higher cost of food products were due to deficient soil type. Different kinds of soil types are used to evaluate fertility of soil (Rahman et al., 2018).The main

objective of this work is to identify attributes for soil fertility and predict a particular crop using the soil dataset.

Some of the factors that influence soil fertility are,

- Acidity or soil PH
- Content of organic matter
- Water draining ability of the soil
- Soil structure
- Minerals present in the soil
- Nutrient release capability

3. RELATED WORK

Several techniques machine learning involved in predicting soil fertility are,

Patel, Amiksha A., Kathiriya, Dhaval R (2017) anticipated the techniques of predicting crop yield production are the core of focus. Yield forecasting is crucial in agriculture. The complexity of yield prediction can be addressed with the help of data mining techniques.

M.C.S.Geetha (2015a) discussed the significance of different data mining techniques in the agriculture ground and also confers about numerous techniques of machine learning and also discusses several fields of data mining for solving the different agricultural problems.

Riyaz Sikora (2014) presented a customized version of the typical stacking ensemble algorithm that uses the genetic algorithm for creating an ensemble and tested the GA-based stacking algorithm on ten data sets from the UCI Data Repository and showed the improvement in performance over the individual learning algorithms as well as over the standard stacking algorithm.

Kodimalar Palanivel (2019) performed an investigation on various machine learning algorithms that are useful in the prediction of crop yield. An approach has been proposed for the prediction of crop yield using machine learning techniques in the big data computing paradigm.

Narayanan Balakrishnan and Dr.Govindarajan Muthukumarasamy (2016) proposed ensemble models such as AdaSVM and AdaNaive are used to protect the crop productions that are compared to SVM and Naive Bayes methods. The two parameters were used separately for predictions of output are the accuracy and the classification error.

Geetha MCS (2015b) assessed the range of association techniques in Data Mining and applied them to the database of soil science for predicting meaningful relationships and provided association rules for different soil types in agriculture.

Tittonell P, Shephered KD, Vanlauwe B, Giller KE (2000) analyzed the virtual significance of soil fertility and the crop management factors for predicting the maize yields and in determining the yield inconsistency and the gap between farmers. Classification and regression tree analysis was used to predict the result.

Jayalakshmi R, Savitha Devi M (2021) used various supervised learning algorithms, such as SVM, KNN, and Decision tree that are helpful for identifying the type of fertility and conducting supervisory training on data sets collected from the agriculture domain to segregate information into multiple classes.

Feng L, Zhang Z, Ma Y, Du Q, Williams P. Jessica Drewry and Brian Luck (2020) developed an ensemble-based machine learning model for alfalfa yield prediction with UAV-based hyperspectral images. The designed model comprises random forest (RF), support vector regression (SVR), and K-nearest neighbors (KNN). Initially, a huge number of hyperspectral images were taken from the original data. Feature selection was used for retreating the data dimensionality. An ensemble learning model was engaged for improving prediction performance.

4. MATERIALS AND METHODS USED

4.1 Dataset Description

The sample data for this research is a real-time dataset of soil fertility that has been obtained from the soil testing laboratories, Melalathur, Vellore District was used to predict the accuracy of soil data. Dataset has ten attributes and 1430 instances of soil samples and soil was classified into two class labels: Ideal and NotIdeal have been further used for decision making. The detailed description of the dataset is shown in Table 1.

The dataset consisted of ten attributes includes pH, EC, Fe, Zn, Mn, Cu, OC, P, K, FI. The dataset also consisted of a set of numerical and categorical features. The collected dataset has been organized in Excel Sheet with CSV file format. Table 2 shows attribute description.

Table 3 shows the sample dataset used for experimentation in our work that was stored in the form of a CSV (Comma Separated Values) file. As stated above, the soil dataset possesses 10 attributes wherein Class is the dependent variable while all the other variables are predictor variables that are utilized in predicting the soil fertility level.

4.2 Ensemble Machine Learning Algorithms

4.2.1 Bagging

Bagging stands for bootstrap aggregation combines ***B***ootstrapping associates in nursing ***Agg***regation to make an ensemble model. The most objective of bagging is to average multiple estimates along for reducing the variance of an estimate. With a precise sample of data, multiple bootstrapped subsamples are used to form a decision Tree. An algorithm is used to aggregate over the Decision Trees to form the most efficient predictor after each subsample Decision Tree has been formed. Bagging used

Table 1. Description of the Soil fertility dataset

Dataset	
Number of attributes	10
Number of instances	1430
Number of classes	2

Table 2. Attributes and its description

Attributes	**Description**
pH	pH value of soil
EC	Electrical conductivity
Fe	Iron
Zn	Zinc
Mn	Manganese
Cu	Copper
OC	Organic Carbon
P	Phosphorous
K	Potassium
FI	Class Label

Table 3. Sample Soil dataset

S. No	Sample	pH	EC	OC	N	P		Class
1	TN634912/2016-17/10171036	8.40000 Mal	0.20000 N	0.30000 L	256.00000 L	21.00000 L	…	Ideal
2	TN634912/2016-17/10171329	7.90000 Mal	1.00000 N	0.21000 VL	186.00000 L	14.00000 L	…	Not Ideal
…	…	…	…	…	…	…		
n	TN634920/2017-18/5419836	8.30000 Mal	0.10000 N	0.35000 L	49.00000 VL	19.00000 L	…..	Not Ideal

bootstrap sampling to get the data subsets for training the base learners, voting for classification, and averaging for regression to aggregate the outputs of base learners.

4.2.2 Boosting

Boosting refers to a family of algorithms which are capable of shifting weak learners into robust learners. The prime principle of boosting is to suit a sequence of weak learners solely better than random guess, like decision trees. The predictions are then combined through a weighted majority vote (classification) or a weighted sum (regression) to provide the final prediction. The key difference between boosting and the committee methods, like bagging in which, base learners are trained in cycle on a weight of the data. In this method, each one model boosts the performance of the ensemble. The most important goal of boosting is to reduce bias.

4.2.3 Stacking

Stacking is also called super learning. It is an ensemble learning technique that combines multiple classifications or regression models through a meta-classifier or a meta-regressor. The base-level models are trained based on a complete training set, and then the meta-model is trained on the outputs of the base-level model-like features. The goal of stacking is to ensemble strong, diverse sets of learners together for achieving better prediction.

4.3 R Tool

R is a software environment and programming language which can be used for data modeling, graphical representation statistical analysis, and reporting. R was foremost formed and developed by Ross Ihaka and Robert Gentleman at the University of Auckland New Zealand in 1993. The R Framework is open source and flexible. R consists of different packages that are useful in the analysis of data. R is an integrated suite of software facilities for data manipulation, calculation, and graphical facilities for data analysis and display (Jayalakshmi, 2018).

5. PROPOSED METHODOLOGY

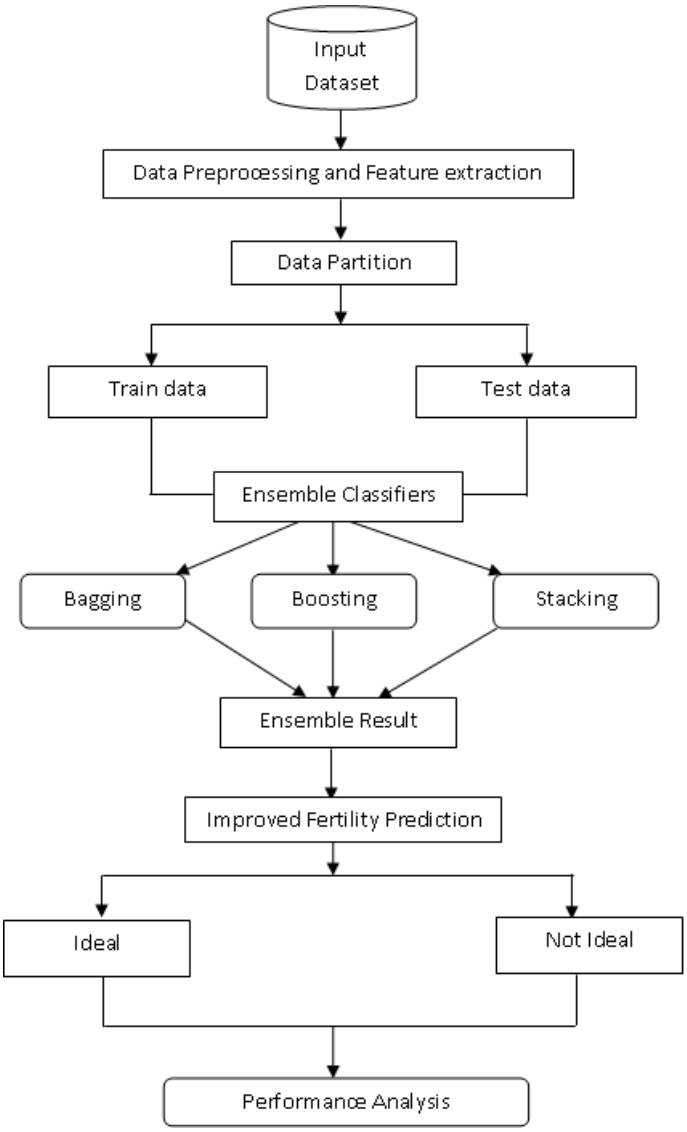
The main objective of the proposed work is to analyze the soil data using multiple ensemble learning algorithms. The methodology starts with soil data collection; then dataset preprocessing has been carried out for filtering missing attribute values, noisy data, and miss-match. The filtered dataset is partitioned into two sets such as a training set and a test set. The training set was used to build the classifier model and the test-set was used to test the built classifier model. Then multiple ensemble learning algorithms are applied to predict soil fertility in a better way for finding hidden information followed by the evaluation of results. Figure 1 represents the research Framework of soil fertility prediction and the steps involved as,

1. Preprocessing
2. Feature extraction
3. Ensemble classification with several voting techniques
4. Improved soil fertility prediction

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this work, multiple ensemble learning algorithms such as Bagging, Boosting, and Stacking were compared and evaluated based on accuracy to build a model. Two groups are separated from the dataset for training and testing the algorithms of classification. The Training data consists of 10 attributes

Figure 1. Research Framework of soil fertility prediction



along with an additional attribute as Label or Response attribute pre-defined by the Soil Testing Lab based on availability of Macro and Micro Nutrients present in the soil. R Tool is used for implementing ensemble learning algorithms. A number of experiments were performed to evaluate the proposed model. The performance of the proposed methodology was also evaluated using various error measures such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Square Error (MSE). In this section, performance analysis of various soil samples is carried out based on different

Figure 2. Ensemble Bagging result

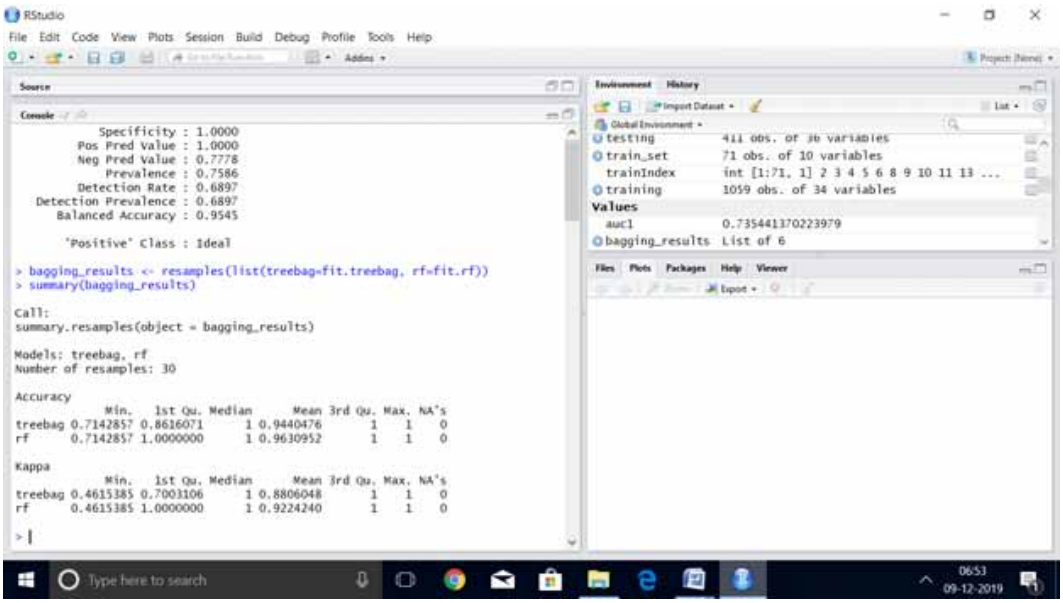
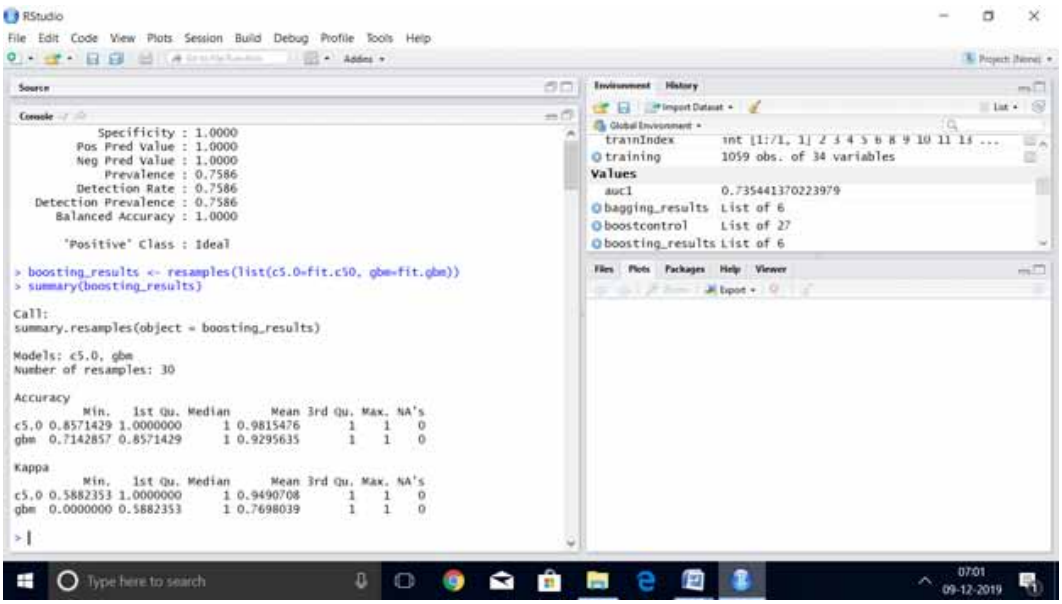


Figure 3. C5.0 with Gradient boosting result



parameters i.e., soil fertility prediction time, accuracy, and false-positive rate. Comparative analyses are made with multiple ensemble classifiers (boosting, bagging, and stacking) on RandomForest, Gradient boosting, Logistic Regression, CART, k-Nearest Neighbors, and Support Vector Machine algorithms based on a 10-fold cross-validation as a test method through the feature selection.

7. CONCLUSION AND FUTURE WORK

Ensemble Machine Learning is a crucial approach of the present digital world for analyzing the mass of datasets to achieve effective solutions. The analysis of agriculture soil datasets with multiple

Figure 4. Ensemble Stacking result

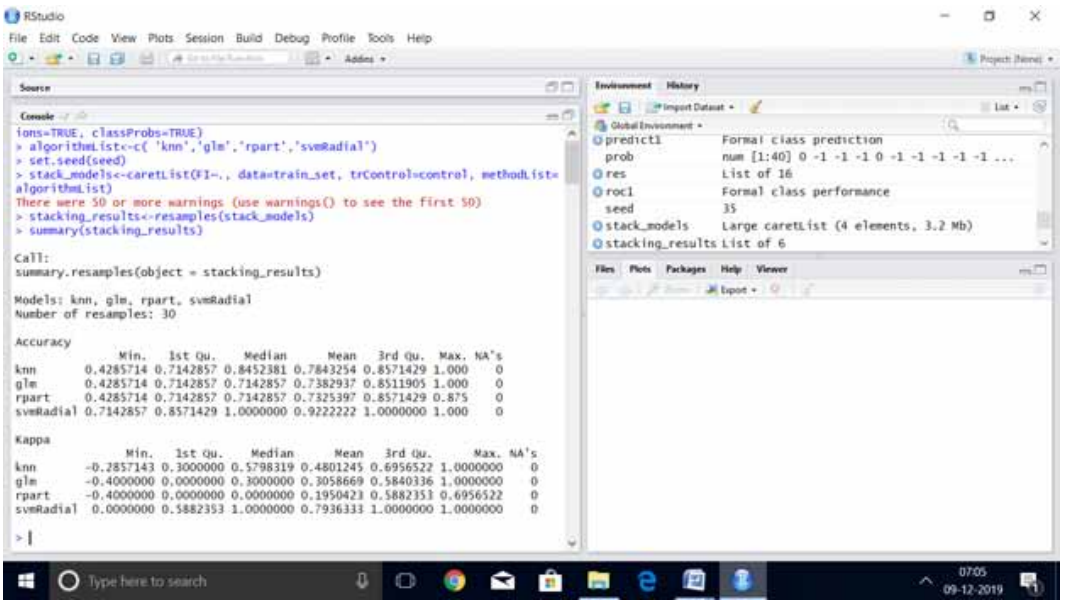
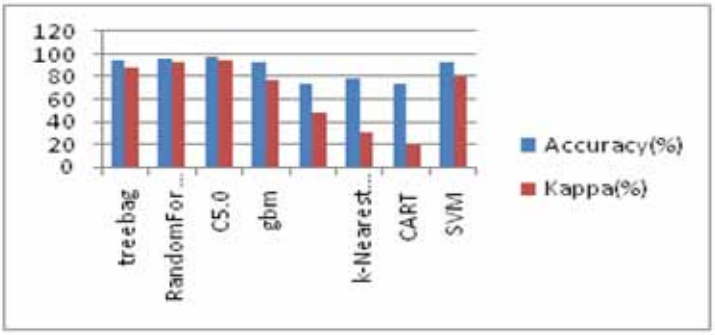


Table 3. Result of Ensemble Learning Algorithms

Ensemble Learning Algorithms	Accuracy(%)	Kappa(%)
Ensemble Bagging		
Treebag	94.40	88.06
RandomForest	96.30	92.24
Ensemble Boosting		
C5.0	98.15	94.90
Gbm	92.95	76.98
Ensemble Stacking		
Logistic Regression via GLM	73.82	48.01
k-Nearest Neighbour	78.43	30.58
CART	73.25	19.50
SVM	92.22	79.36

Figure 5. Accuracy and kappa statistics of different ensemble learning Algorithms



Ensemble Learning Algorithms may yield useful outcomes to farmers to make the right decision for achieving and maintaining the necessary level of soil fertility. This paper aims to predict and analysis of soil fertility parameters as inputs. The soil fertility classes were evaluated using 10 selected attributes. The authors implemented multiple ensemble learning algorithms like Bagging, Boosting, and Stacking with data analytics tool R for generating efficient predictive models. The experimental results show that boosting method with the C5.0 algorithm achieved the highest accuracy (98.15%) compared to other ensemble classifiers. In the future, our goal is to use the deep feature learning to accurately analyze the prediction of the fertility of the soil with an extended benchmark dataset under the different climatic conditions for obtaining the best prediction with high accuracy.

FUNDING AGENCY

The publisher has waived the Open Access Processing fee for this article.

ACKNOWLEDGMENT

Authors wish to express gratitude and respect for the Senior Agricultural Officer who provided the data, guidance, and support which greatly contributed towards the timely completion of this work.

REFERENCES

Balakrishnan, , & Muthukumarasamy, . (2016). Crop Production-Ensemble Machine Learning Model for Prediction. *International Journal of Computer Science and Software Engineering*, *5*(7), 148–153.

Bhuyar, V. (2014). Comparative analysis of classification techniques on soil data to predict fertility rate for Aurangabad District. *International Journal of Emerging Trends and Technology in Computer Science*, *3*(2), 200–203.

Bindraban, P. S., Stroorvofel, J. J., Jansen, D. M., Vlaming, J., & Groot, J. J. R. (2000). Land quality indicators for suitable land management: Proposed methods for yield gap and soil nutrient balance. *Agriculture, Ecosystems and Environment*, *81*(2), 103–112.

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., & Luck, B. (2020). Alfalfa Yield Prediction Using UAV-Based Hyperspectral Imagery and Ensemble Learning. *Remote Sensing*, *12*(12), 2028. doi:10.3390/rs12122028

Geetha, M. C. S. (2015a). A Survey on Data Mining Techniques in Agriculture. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(2), 287–290.

Geetha, M. C. S. (2015b). Implementation of association rule mining for different soil types in agriculture. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(4), 520–522. doi:10.17148/IJARCCE.2015.44119

Jayalakshmi, R. (2018). R: A Predictive Analytics and statistical Data Mining Tool. *International Journal of Science Technology and Humanities*, *5*(2), 26–35.

Jayalakshmi, R., & Savitha Devi, M. (2021). Predictive Model Construction for Prediction of Soil Fertility using Decision Tree Machine Learning Algorithm. *INFOCOMP*. *Journal of Computational Science*, *20*(1), 49–55.

Palanivel, K., & Surianarayanan, C. (2019). An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques. *International Journal of Computer Engineering and Technology*, *10*(3), 110–118. doi:10.34218/IJCET.10.3.2019.013

Patel, A. A., & Kathiriya, D. R. (2017). Data Mining Trends in Agriculture: A Review. *Agres- International Journal (Toronto, Ont.)*, *6*(4), 637–645.

Rahman, S. A. Z., Chandra Mitra, K., & Mohidul Islam, S. M. (2018). Soil Classification Using Machine Learning Methods and Crop Suggestion Based on Soil Series. *21st International Conference of Computer and Information Technology (ICCIT)*. doi:10.1109/ICCITECHN.2018.8631943

Sikora, R. (2014). A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms. *Journal of International Technology and Information Management*, *23*(1), 1–12.