IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 12, December 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

A Hybrid Machine Learning Approach for Identifying the Type of Fertilizer to Increase the Productivity in Agriculture

Nithya P[1], Dr.A.M.Kalpana[2]

Assistant Professor, Department of Computer Engineering, Government College of Engineering, Salem, India[1]

Professor and Head, Department of Computer Engineering, Government College of Engineering, Salem, India [2]

ABSTRACT: Soil testing is a valuable tool for evaluating the available nutrient status of soil and helps to determine the proper amount of nutrients to be added to a given soil based on its fertility and crop needs. In the current study, the soil test report values are used to classify several significant soil features like village wise soil fertility indices of Available Phosphorus (P), Available Potassium (K), Organic Carbon (OC) and Boron (B), as well as the parameter Soil Reaction (pH). The classification and prediction of the village wise soil parameters aids in reducing wasteful expenditure on fertilizer inputs, increase profitability, save the time of chemical soil analysis experts, improves soil health and environmental quality. In this paper, a hybrid machine learning method named as Hybrid probabilistic neural network with lasso regression is proposed along with optimization model to find the type of fertilizer for that particular soil to increase the productivity.. We have discussed all of the performance evaluation metrics such as accuracy, recall, precision, F-score. The performance of the proposed system has been validated on full features and on a reduced set of features. Thee features reduction has an impact on classifiers performance in terms of accuracy and execution time of classifiers.

KEYWORDS: Fertilizers, Productivity, machine learning, probability, optimization.

I. INTRODUCTION

India is the leading producer of some of the crops like it is second leading producer of pulses. When we compare per hectare yield of India with other countries, it is 3 tonnes which is lowest in the world [1]. The reasons are lack of irrigational facilities, improper pest management, soil erosion due to natural calamities, inadequate storage facilities, lack of transportation, scarcity of capital etc. Moreover most of the Indian farmers are illiterate and poor. They cannot afford for good quality seeds and modern fertilizers, and cannot take decisions that help them to increase the crop yield. There are some techniques to solve agricultural issues like rule based systems [2], data mining techniques [3] etc. The rule based systems help farmers to take decisions in choosing good quality seeds, proper quantity of pesticides to be used etc. They work well if and only if all the situations under which decisions can be made are known a head. Maintenance and extension of a rule base can be difficult for a relatively large rule base (beyond 100 rules). It also involves high operational costs.

In earlier days of ML techniques, the Levenberg-Marquardt based back-propagation method in the Artificial neural networks (ANNs) was used to predict the soil fertility [4]. Partial least squares regression was also used to forecast the soil fertility using the available water capacity, electrical conductivity (EC), clay loam, silt loam, sandy loam soils, soil OC and soil bulk density [5]. Numerous studies have been applied with ML techniques to identify and solve soil problems in agriculture like the prediction of soil fertility, the supply of the required nutrient levels and water etc. [6]. The wheat yield was classified and predicted by J48, K-nearest neighbors (KNN), One-R and Apriori classifier methods using phenotypic plant traits as inputs [7]. Using supervised Kohonen and counter-propagation neural networks, the wheat yield has been quantified as low, medium and high [8]. Naive Bayes classifier, Decision Tree (DT), Random Forests (RF), Support Vector Machine (SVM), AdaBoost and Logistic Regression (LR) were used to make decisions about insecticide application for leafroller pest monitoring on kiwifruit [9]. An unbiased linear predictor was used to predict the soil organic carbon [10]. The organic carbon on Sicilian soils was predicted by using the boosted regression trees [11]. The random forest has been combined with the feature selection method of genetic algorithms to predict organic carbon on soils of eastern Australia [12]. Organic carbon alongside with soil acidity (pH) and Cation Exchange Capacity (CEC) from mid-infrared spectra for a number of soils were predicted by partial least squares [13]. The Bayesian network was applied to perform soil fertility rating with the pH value of soil and the soil

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

nutrients like nitrogen, copper, iron, potassium, phosphorus, organic carbon and Zinc [14]. The geographic portability of machine learning algorithms has been evaluated for the forming of wind speed [15]. DT, RF, SVM, Bayesian Networks, and ANN methods were able to analyze soil and climate which are directly involved in precision farming and crop growth [16].

Different machine learning approaches were used to predict the soil nutrient content, soil type and soil moisture [17]. To categorize village wise soil fertility indices and soil nutrient levels, a collection of twenty classifiers including RF, AdaBoost, SVM, neural networks and bagging have been used and the class labels were calculated on a scale of low, medium and high based on their numeric values [18]. A wide collection of regression methods has been used to generate pedotransfer functions which straightly foretell the numeric values of village wise fertility indices [19]. The soil fertility information of India is summarized for district and block levels. This information is suitable for making decisions about proper quantity usage of fertilizers, the consumption in terms of variations in fertility levels and the procedure of fertilizer distribution. The key objective of this research work is to classify area wise soil fertility indices based on the village level soil fertility information. This classification can be used in making a village wise fertility index analysis report and it would be used to make fertilizer recommendations with the decision support system.

II. LITERATURE SURVEY

P. Jasmine Sheela. Sivaranjani [20] made a comparative study of various classification algorithms such as Random forest, J48, Naïve Bayes and concludes that accuracy level for J48 algorithm is 98.17%, for Random forest algorithm is 90.81%, for Naïve Bayes algorithm is 77.18%, for JRip algorithm is 92.53%, and after using boosting for J48 it increase the accuracy level as 96.73%. This algorithm is used to analyze a large number of datasets and it is used to create a soil fertility prediction in easy manner.

Jay Gholap [21] used Decision Tree algorithm for predicting soil fertility. He also used Selection and Boosting techniques to tune the performance of J48 algorithm. Ad boost is one of the boosting techniques used in this research. After using selection and boosting algorithm it increase the accuracy level to 96.73%.

S.S.Baskar, L.Arockiam, S.Charles[22] classifies the soil dataset using various Classification algorithms such as regression, J48, Naive Bayes. J48 gives the best result and accuracy level is 93.86%. For given soil samples this system will recommend the fertilizer appropriately.

VrushaliBhuyar [23] focuses on classification of soil sample of Aurangabad District to predict fertility rate. She use J48, Naïve Bayes, and Random forest algorithm to calculate fertility rate of soil data. Her Study shows that among the classifier J48 classifier perform best to predict fertility index. This will help to decision maker to recommend fertilizers accordingly.

Amrutha A, Lekha R, A Sreedevi [24] says that the time taken for soil testing by chemical methods in a laboratory takes a few days, while in the proposed system the results are obtained within 30 minutes. The results obtained from the soil test are fed to sensors and the results are analyzed using a microcontroller which in turn needs a few seconds. The system has been developed for the controlled addition of fertilizers in order to avoid excess/ deficient fertilizers in the soil.

Navneet, Nasib Singh Gill [25] propose a technique that produces a compact decision tree having increased classification accuracy and also Schwartz criterion is used to get the optimal number of clusters. The experimental results show that incorporating both clustering and classification algorithm is fast and more accurate for recommendation of fertilizers of real world soil dataset.

Veenadhari [26] S. Influence of climatic factors on major kharif and rabi crops production in Bhopal District of Madhya Pradesh State was considered. The findings of the study revealed that the decision tree analysis indicated that the productivity of soybean crop was mostly influenced by comparative humidity followed by temperature and rainfall. The decision tree analysis shows that the productivity of paddy crop was mostly inclined by Rainfall followed by comparative Evaporation and humidity. For Wheat crop, the analysis shows that the productivity is mostly influenced by Temperature followed by relative humidity and rainfall. The result of decision tree were confirmed from Bayesian classification. The rules formed from the decision tree are useful for identifying the conditions intended for high or low crop productivity.

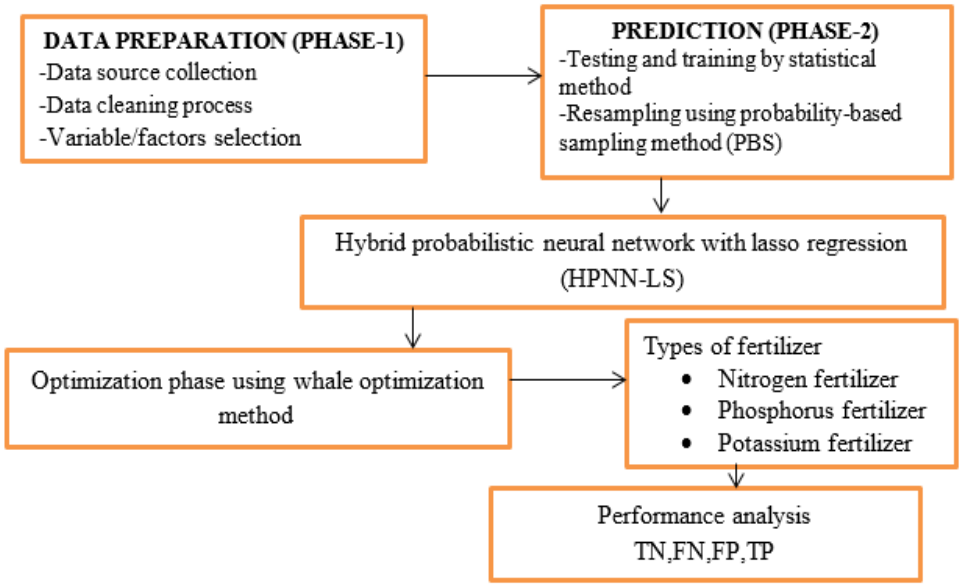
Shalvi D and De Claris N Bayesian network is a powerful tool and broadly used in agriculture datasets. The model developed for agriculture application based on the Bayesian network learning method. The results show that Bayesian Networks are feasible and efficient. Bayesian approach improves hydro geological site characterization even when using low-resolution resistivity surveys[27].

K. Verheyen et al., Data Mining is the process of discovering meaningful patterns and trends by shifting through huge amount of data, using pattern detection technologies as well as statistical and mathematical techniques. Data Mining

techniques are often used to studied soil characteristics. As an example, the K-Mean approach is used for classifying soils in combination with GPS based techniques[28].
AltannarChinchulunnet al.,The k-nearest neighbor classification algorithmic rule may be divided into 2 phases: coaching section and testing section. Bermejo associated Cabestany urged a reconciling learning algorithmic rule to permit fewer information points to be utilized in coaching information set. Several different techniques are projected to scale back procedure burden of k-nearest neighbor algorithms[29].

III. RESEARCH METHODOLOGY

Samples of soil are collected from individual farmers by the soil testing laboratory. The area wise analysis results and the gradation details about each input measure are publically available. The soil samples were analyzed for 15 parameters of immediate relevance to plant nutrition: soil reaction (pH), electrical conductivity (EC), OC, plant available primary nutrients (P, K), secondary nutrient (S) and micronutrients (Cu, Zn, Fe, Mn and B).



Data preparation:
The preprocessing of data is necessary for efficient representation of data and machine learning classifier which should be trained and tested in an effective manner. Preprocessing techniques such as removing of missing values, standard scalar, and MinMax Scalar have been applied to the dataset for effective use in the classifiers.The standard scalar ensures that every feature has the mean 0 and variance 1, bringing all features to the same coefficient. Similarly, in MinMax Scalar shifts the data such that all features are between 0 and 1. 0 missing values feature row is just deleted from the dataset.
Least absolute shrinkage and selection operator select features are based on updating the absolute value of features coefficient. Some coefficients values of features become zero, and these zero coefficients features are eliminated from features subset. Thee LASSO performs excellently with low coefficients feature values. The features having high values of coefficients will be included in selected feature subsets. In LASSO, some irrelevant features may be selected and include a subset of selected feature.

Dataset description:

fertilizer	Physical form
Diammonium phosphate	Solid, granule
Monoammonium phosphate	Solid, granule
Triple superphosphate	Solid, granule
Ordinary superphosphate	Solid, granule
Ammonium poly phosphate	liquid
Potassium chloride	Solid, granule
Potassium sulfate	Solid, granule

Potassium nitrate	Solid, granule, crystal
Calcitic Limestone	solid
Manganese Oxy-sulfate	Solid, granule
Zinc Oxy-Sulfate	powder
Amide nitrogen	powder
Ammonium nitrogen	Solid, granule, crystal
Sodium nitrate	Solid, granule
Calcium nitrate	Solid, granule, crystal

Hybrid probabilistic neural network with lasso regression

The PNN provides a general solution to pattern classification problems by following an approach developed in statistics, called Bayesian classifiers. The network paradigm also uses Parzen Estimators which were developed to construct the probability density estimation required by Bayes theory . The PNN structure was used in this study has a multilayer structures consisting of an input layer, a single hidden layer (radial basis layer), and an output layer (competitive layer) . In this system, real valued input vector is feature's vector and two outputs are index of two classes (healthy and mesothelioma). All hidden units simultaneously receive the 15-dimensional real valued input vector. The pattern layer consists of a set of radial basis functions. In the training procedure, the training data is given to the network and passes from the input layer through the pattern layer to the output layer. PNN uses training input data to set up the weights (Wi) between input and pattern layer as follows.

C(ij)=f(j) (1)

f(j)= (f1,f2,f3……f15) = W(i) = W(Ij) whereas, i=1,2,3,4…..15 and j= 1,2,3….30

where wij is the j^{th} weight value of weight (Wi). Then the network will define the weight (Wo) among the pattern layer and output layer as follows:

$$w(ij) = (1, if\ the\ jth\ sample\ and\ neuron\ is\ belong\ to\ same\ class\ and\ 0, others) \quad (2)$$

If the j^{th} sample is associated with class c, c = 1,2 the weight value is defined as 1 and others are 0. In the recall procedure, the testing data is given to the network and the weights between the input and pattern layer calculates the Euclidean norm (Edj) of training and testing data. The Euclidean norm function can be described as:

$$ED = \sqrt{\sum W(ij) - f(i)\text{^}2} \quad (3)$$

where w is the weight vector and f is the input vector. The pattern units generate the probabilistic vectors and after calculating the probabilities of each condition, the weight Wo transports these probabilities to the summation layer. The summation layer neurons sum the inputs from the pattern layer and transmit them to output layer. Output layer uses 'winner takes all' attitude to compare the probability density of each condition in the output layer.

The hidden layer neurons (41 neurons for each hidden layer) and the output layer neurons use nonlinear sigmoid activation functions. In this system, 34 inputs are features, and two outputs are index of two classes.

Outputs of the first hidden layer neurons are,

$$X1(n) = \frac{1}{(1+\exp(w1(n)*f(n)+b1(n)} \quad (4)$$

Outputs of the second hidden layer neurons are,

$$X2(n) = \frac{1}{(1+\exp(w2(n)*f(n)+b2(n)} \quad (5)$$

Outputs of the network are,

$$Y(n) = \frac{1}{(1+\exp(wh(n)*f(n)+bh(n)} \quad (6)$$

where W1 (n) are the weights from the input to the first hidden layer and b1(n) are the biases of the first hidden layer, W2 (n) are the weights from the first hidden layer to the second hidden layer and b2(n) are the biases of the second hidden layer, Wh(n) are the weights from the second hidden layer to the output layer and bh(n) are the biases of the output layer, f(n) values are the features, y(n) values are the outputs for the class index, and n is training pattern index.

The regression line represents the estimated disease chance for a given combination of the input factors. Scatter plot is defined by a linear equation of,

y=β0 + β1 x1+ β2x2+.....+ βixn for i = 1…n (7)

The deviation between the regression line and the single data point is variation that our model cannot explain. This unexplained variation is also called the residual. The method of least squares is used to minimize the residual.

The lasso regression's variance α^2 is estimated as $S^2 = \frac{\sum e\ (i)\text{^}2}{n-p-1}$ (8)

Where,

p is the number of independent variables and

n the sample size.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

After getting lasso regression equation, the validity and usefulness of the equation is evaluated. The key measure to the validity of the estimated linear line is R^2. R^2 = total variance / explained variance.

Optimization using whale method:

Phase 1: Initialization The algorithm recognized by randomly making weight value that interconnects to the outcome in the examine space. Here the weight represents the random wi (i=1, 2, 3,.. n) where n represents the number of weight value. And initialize the factor vectors of value such as a, A, and C.

Phase 2: Fitness Calculation Evaluate the fitness on basis of the equation (9) to become the finest weight value the fitness point of the solution is calculated. It"s exposed in under

Fiti=MinMSE (9)

Phase 3: Update the situation of current weight point Encircling prey: The situation of prey is documented by humpback whale and it is encircled the prey. Towards the finest finding operator, the other finding operators will accordingly effort to inform their locations when the finest finding agent is considered. The update methods, is projected by the underneath equations:

U = /C.w (t)best- w(t) / (10)

W(t+1) = w(t)best –A.U (11)

Where,

t designates a current repetition,

A and C designates a Coefficient vector,

w (best) designates a position vector for best solution,

w signifies a current location Vector,

// represents an absolute point.

The vectors A and C are calculated as follows:

A = 2 a . r –a (12)

C=2.r (13)

Where, a is the sequence of repetitions linearly abridged from 2 to 0 (in both exploration and exploitation phases), r ∈(0,1). Bubble-net attacking method: To model the bubble net performance of hump back whales scientifically two methods are enhanced as trails.

Shrinking encircling method: The point of a in the equation (9) is reduced to attain this performance. Note that a ρ is used to reduction the difference range of A ρ . In other words, in the interval [−a, a] A is an accidental point where a is reduced from 2 to 0

Search for prey: To find for prey the difference of the A ρ vector can be castoff by same method. In fact, rendering to the location of each extra hump back whales find randomly. So, to force find agent to transfer so far from a situation whale we use A ρ with the accidental points better than 1 or less than −1. The location of the find agent is efficient in manipulation phase. To perform a global search, this method and A >1 ρ highlight examination allows the WOA algorithm. The precise model is given below:

U = /C.w_{rand} / (14)

w (t+1) = wrand – A. U (15)

Where, wrand is a current population chance location vector. Find agents update their locations in every repetition with admiration to either the finest solution got so far or an arbitrarily chosen find agent. In order to deliver examination and manipulation the limit a is reduced from 2 to 0, correspondingly.

Phase 4: Termination criteria: The WOA algorithm is terminated when finest weight point is obtained for the satisfaction of a closure principle.

Performance analysis

The accuracy score which computes the accuracy by the count of correct predictions is the measure used for classification performance. The fraction of correct predictions over ksamples is defined as:

Accuracy $= \frac{1}{k(samples)} \sum_{i=0}^{nk\ samples-1} 1(y(i) = Y(i)$

Where,

y(i) is the predicted value of the i^{th} sample,

Y(i) is the corresponding true value and

1(x) is the indicator function

The other four measures used for calculating the classification and prediction performance are Kappa, Precision, Recall and F Score

Kappa (K) =$\frac{p(o)-p(e)}{1-p(e)}$

Where,

po is the probability of the actual observed value of the classification and

pe is the probability of a right classification by chance.

The confusion matrix obtained after the methodology analysis helps to compute the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) values.

Precision (%) = $\frac{TP}{TP+FP}$ *100

Recall (%) = $\frac{TP}{TP+FN}$ *100

f-Scorel (%) = $\frac{2(precision*recall)}{precision+recall}$ *100

Table- 1 Comparison of proposed and existing method

parameter	Hybrid probabilistic neural network with lasso regression	PNN with linear regression method
Accuracy	89%	81.4%
kappa	67%	56%
F-score	78.3%	57.4%
precision	78.4%	48.2%
recall	67.5%	54.2%

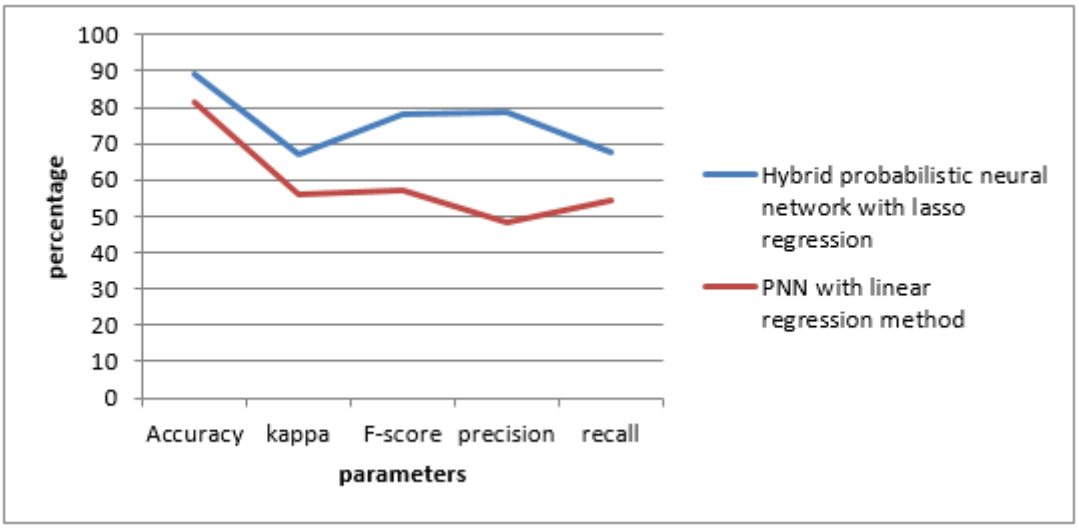


Figure-2 Analysis of various parameters between proposed and existing method

IV. CONCLUSION

The main reason for a heavy loss in soil quality is due to the incorrect soil and crop management strategies. Also, excess use of chemical fertilizers has created imbalances in the availability of soil nutrients. The main objective of this work is to classify area wise soil fertility indices based on the area level soil fertility information, which can be used in making a soil fertility index analysis report with preferred fertilizers. The very fast and simple learning algorithm called hybrid probabilistic neural network with lasso regression with its optimization concept are used to predict the fertility levels as low, medium and high using the soil parameters as input values; and to predict the pH levels of the corresponding soil. The classifiers lasso regression with 10-fold cross-validation showed best accuracy 89% when

selected by HPNN-LRmethod. Due to the good performance of lasso regression with Relief, it is a better predictive system in terms of accuracy.

REFERENCES

1. P. Mercy Nesa Rani, T. Rajesh and R. Saravanan Expert Systems in Agriculture, Journal of Computer Science and Applications. Volume 3, Number 1 (2011), pp. 59-71
2. B. Milovic and V. Radojevic Bulgarian Journal of Agricultural Science, 21 (No 1) 2015, 26-34 Agricultural Academy, Application of Data Mining in Agriculture
3. Namita Field Mirjankar, Smitha Hiremath International Journal of Computer Engineering &Applications,iCCSTAR 2016,SpecialIssue,May2016 Application of Data Mining In Agriculture.
4. Sheela PJ, Sivaranjani K, Phil M. A brief survey of classification techniques applied to soil fertility prediction. Int Conf Eng Trends Sci Hum 2015:80–3.
5. de Paul Obade V, Lal R. Towards a standard technique for soil quality assessment. Geoderma2016;265:96–102.
6. Mucherino A, Papajorgji P, Pardalos PM. A survey of data mining techniques applied to agriculture. Oper Res Int J 2009;9(2):121–40.
7. Romero JR, Roncallo PF, Akkiraju PC, Ponzoni I, Echenique VC, Carballido JA. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput Electron Agric 2013;96:173–9.
8. Pantazi XE, Moshou D, Alexandridis T, Whetton RL, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric 2016;121:57–65.
9. Hill MG, Connolly PG, Reutemann P, Fletcher D. The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. Comput Electron Agric 2014;108:250–7.
10. Ritz C, Putku E, Astover A. A practical two-step approach for mixed model-based kriging, with an application to the prediction of soil organic carbon concentration. Eur J Soil Sci 2015;66(3):548–54.
11. Schillaci C, Acutis M, Lombardo L, Lipani A, Fantappie M, Ma¨ rker M, et al. Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: the role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. Sci Total Environ 2017;601:821–32.
12. Wang B, Waters C, Orgill S, Cowie A, Clark A, Li Liu D, et al. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. Ecol Ind 2018;88:425–38.
13. Terhoeven-Urselmans T, Vagen TG, Spaargaren O, Shepherd KD. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Sci Soc Am J 2010;74(5):1792–9.
14. Jia HY, Chen J, Yu HL, Liu DY. Soil fertility grading with Bayesian Network transfer learning. In: International conference on machine learning and cybernetics. IEEE; 2010. p. 1159–63.
15. Veronesi F, Korfiati A, Buffat R, Raubal M. In: Assessing accuracy and geographical transferability of machine learning algorithms for wind speed modelling. Cham: Information Science. Springer; 2017. p. 297–310.
16. Elavarasan D, Vincent DR, Sharma V, Zomaya AY, Srinivasan K. Forecasting yield by integrating agrarian factors and machine learning models: A survey. Comput Electron Agric 2018;155:257–82.
17. Reashma SJ, Pillai AS. Edaphic factors and crop growth using machine learning—A review. In: International conference on intelligent sustainable systems (ICISS). IEEE; 2017. p. 270–4.
18. Sirsat MS, Cernadas E, Ferna´ndez-Delgado M, Khan R. Classification of agricultural soil parameters in India. Comput Electron Agric 2017;135:269–79.
19. Sirsat MS, Cernadas E, Ferna´ndez-Delgado M, Barro S. Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. Comput Electron Agric 2018;154:120–33.
20. P. Jasmine Sheela, M.Sc., K. Sivaranjani, M.Sc., M.Phil., “A Brief Survey of Classification Techniques Applied To Soil Fertility Prediction”, ICETSH-2015.
21. Jay Gholap, “Performance tuning of j48 Algorithm for prediction of soil Fertility”, AJCSIT, 2012.
22. S. S. Baskar, L. Arockiam, S.Charles “Applying Data Mining Techniques on Soil Fertility Prediction”, IJCAT, 2013. [
23. VrushaliBhuyar, “Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District”, IJETTCS, 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details