Performance Analysis of Soil Health Classifiers Using Data Analytics Tools and Techniques for Best Model and Tool Selection

https://doi.org/10.3991/ijoe.v18i10.30149

Sushma Vispute(✉), Madan Lal Saini
Poornima University, Jaipur, India
sushma.vispute@pccoepune.org

Abstract—One of the most crucial stages in the building of a Machine Learning (ML) model is the evaluation and analysis of classifier model performance. The agricultural sector is the economic backbone of India and needs extensions to provide solutions to the problems faced by the farmers. This paper presents agriculture soil health analysis using Machine Learning approaches for best model and tool selection and also bibliometric analysis to identify different sources and author's keywords for finding the area of focus for proposed work. Models are built on SK-Learn, KNIME, WEKA and Rapid Miner tools using different ML algorithms. Nave Bayes, Random Forest (RF), Decision Tree (DT), Ensemble learning (EL), and k-Nearest Neighbor (KNN) are used to analyze soil data on these tools. Results show that Decision Tree model outperforms other algorithms, followed by RF algorithm which is a set of multiple Decision tree algorithms and SK-Learn tool gives better accuracy followed by WEKA tool then KNIME tool. Maximum accuracy obtained by Decision Tree algorithm is 98.40% using SK-Learn followed by KNIME tool with 73.07%, Maximum accuracy obtained by Naïve Bayes algorithm is 69.50% using SK-Learn followed by KNIME tool with 68.14%, maximum accuracy obtained by Random Forest algorithm is 85.00% using SK-Learn followed by 73.06% using WEKA tool, maximum accuracy obtained by Ensemble algorithm is 89.00% using SK-Learn followed by 73.06% using WEKA tool and for KNN it is 95.50% using SK-Learn followed by 71.85% using WEKA tool.

Keywords—classifier, model performance, analytical tools, machine learning, soil data analysis, bibliometric analysis

1 Introduction

1.1 Machine learning

One of the functions of machine learning algorithms is discovering previously unrevealed interesting-patterns [1] and techniques for classifying samples. Data Mining and ML Techniques are applied to derive an unusual data pattern from the dataset. These techniques play an important function in the agriculture sector also for data analysis. In the agriculture sector, data mining can provide guidance to farmers to gain profit and country development [2]. Different Data mining algorithms are identified that are used in the agriculture sector to provide solutions to the farmer's problems. For accurate predictions, the Prediction process performs inference on the current data [1].

1.2 Soil attributes

Nitrogen (N): Plants absorb nitrogen in the form of ammonium or nitrate ions. Indian soils are almost universally deficient in N. It should be present in the right proportion in the soil for the growth of the plants. The optimum N concentration is 2-10 ppm. **Phosphorous (P):** Phosphorus has been called the Master key to agriculture. It is essential in plant growth, fruit growth, cell division and early ripening. The optimum P concentration is 30-50 ppm. **Potassium (K):** Potassium is an essential micronutrient and is associated with water movement, nutrients and carbohydrates present in the plant tissue. The optimum K concentration is 20K ppm. **Soil pH:** It is an indicator of the alkalinity and acidity in the soils. **The range of pH values: 0–14** (Neutral value: 7, Acidic: <7, Alkaline: >7, Optimal: 6.5 to 7.5).

2 Literature survey

2.1 Related research work

The research efforts carried out in the related systems are discussed in this portion of the article. Most of the documents are referred from IEEE transactions. The IEEE document count analysis referred to in this survey is shown in Table 1.

Table 1. IEEE document count analysis

Publication Year	IEEE Conference	IEEE Journal	Other journals
2021	4	2	2
2020	5	3	6
2019	1	0	1
2018	3	0	1
2017	5	1	0
2016	6	0	1
2015	3	0	1
2000 to 2014	3	0	3
Total Count	30	6	15

In a research carried out by Gholap, Ingole, et al. [21], an automated system for soil classification based on its fertility was proposed. Under this, various classification algorithms like NBTree, SimpleCart, J48 have been studied, with the conclusion that the J48 decision tree algorithm works best with the soil dataset, showing an accuracy of 91.90%.

Hot and Popović-Bugarin analyzes the agriculture problem for clustering of soil contents, and also for visualizing the analyzed output using visualization techniques [22].

P. Vinciya, et al. [10] used model of multiple regression for analyzing Agriculture problems for data mining enabled - High Tech farming for next generation.

Abhishek B. et al. [11] used classification data mining techniques for forecasting of Rainfall status accurately and required Water for Crops using these Techniques.

Authors of papers [18, 19, 23] represented surveys of different analytical tools and techniques for soil health analysis and for student performance analysis.

Table 2 shows summary of recent work done by different authors from IEEE transaction sources.

Table 2. Summary of recent work from IEEE transaction source

Ref No.	Research area focus	Dataset	Objective	Results Description/ Accuracy	Publication Year
[27]	Classification of the plant images into Crop and weed, Deep Learning	Weed and Crop image dataset	To minimize the usage of the herbicide	A maximum efficiency of 96.3%	2020
[28]	Multi-Label Classification, Remote sensing - Maximum Likelihood, Minimum Distance, k-NN, Support vector machines.	Land cover dataset	Classifier analysis of land cover	Better results with the Multi-Label method classifier	2017
[29]	Local Transylvanian areas. Soil classification.	Soil dataset	To improve satellite image training dataset quality	Viable dataset with eliminated noise	2020
[30]	Extraction of Pattern	Soil characteristics	To determine the soil's susceptibility to the presence of Panama disease	Biological suppression of plant pathogens	2018
[31]	Naïve bays. Decision tree. SVM	N, P, K Soil dataset	To suggest the optimal crop based on the soil's NPK concentration.	Decision tree gives higher accuracy. To provide solutions to the farmer's questions in order to boost profit margins.	2021
[32]	IOT, Soil sensors, Image classification, Local binary thresholding	Soil sensors, water quality sensors, temperature sensors, Image dataset	To utilize a robotic arm to harvest the crop autonomously, to maintain crop health and quality	Image recognition will be used to identify the crop, and the batch will be placed in the proper basket for the farmer to consider for examination.	2021
[33]	Support Vector Machine. Gabor Wavelet. Soil type classification.	Soil dataset	To work with soil images in order to create a high-level soil classification scheme	Framework achieved a 97.12% accuracy rate with a low error rate.	2021

[34]	Decision tree J48 algorithm, sensors	Soil dataset	To make recommendations on the crop, fertility of soil, Level of toxicity, and water supply.	Calculates the soil's toxicity level	2018
[35]	C4.5 algorithm	Climatic parameters, crops dataset of Madhya Pradesh	To develop `Crop Advisor'	Determine which climate factor has the greatest impact on crop yields	2014
[36]	PID control. Type-2 fuzzy logic.	External Camera shake	To investigate the active control and stabilization of camera	Active control has been established, and vibration has been reduced.	2020
[37]	Wavelet Technique in Image Fusion	Image dataset	In image fusion, the Wavelet Technique is used.	An early detection system to stop plant pests from spreading further in the Philippines' agriculture sector	2018
[38]	Electrical sensing, optical imaging, Classification	Synchronized optical images, Electrical signal	To categorize pollen grains moving through a device of micro fluidic at 150 grains per second rate using a combination of electrical sensing and optical imaging.	Electrical classifier accuracy: 82.8, Optical classifier accuracy: 84.1%, Multimodal classifier accuracy: 88.3 %	2021
[39]	Deep learning Survey	Agriculture Dataset	To look into the benefits of employing deep learning in agricultural applications.	Bibliography analysis in the different categories.	2020
[40]	Cloud based and sensor based irrigation and an automated agricultural monitoring system	Soil parameters temperature, moisture, fertility	To make the most efficient use of labor and land, maximize output of crop, and reduce energy waste	various characteristics remotely sensed and monitored	2016
[41]	Multitemporal deep learning model	Ppixel-based, time series dataset with 16 crops	To generate the dataset	The dataset's construction is discussed, as well as Deep learning methods for crop type mapping are compared.	2021
[42]	Neural Network, KMeans, SVM, PCA, image processing	Agriculture image dataset	Study of many domains related to agricultural image processing	Plant disease classification and recognition	2015
[43]	Naive Byes, SVM, K-NN, LDA and QDA	Activity dataset	The goal was to create a smart-shirt for farmers.	Provide with an uncertain evidence of reported activities, a priori information related with the crop protocol to recognize the principal activity	2015

[44]	Unmanned aerial vehicle (UAV), FCN-AlexNet,	Image dataset	Yield prediction Assessment of crop growth, fertilizer management	SegNet outperformed FCN-AlexNet. The semantic picture segmentation model has an average inference speed of 0.7s and an 89 percent segmentation identification accuracy.	2020
[45]	Artificial neural networks. Image processing	Soil image dataset	To determine the pH and soil nutrients	Soil nutrients and pH level were determined to be accurate.	2017
[46]	Deep learning (DL) network	Loamy types of soil. silt clay dataset	For spectroradiometer data, determine the quantity of urea fertilizer mixed soils.	R^2 for urea and silt clay soil mixed samples = 0.945 and For urea-mixed loamy soil, $R^2 = 0.954$.	2020
[47]	The Improved Mahalanobis Taguchi System. Multiclass model	26 crop cultivation input factors	Classify 3 crops: paddy, sugarcane, and groundnut.	The classifier is perfect in terms of accuracy (100%), recall, precision, and error rate (0%) .	2020
[48]	Data mining, Machine learning	Soil data	To analyze and classify soil data and to increase the effectiveness of each model by combining different models.	Analyze fertility of soil, improve efficiency	2020
[49]	Knowledge-based classification non-parametric classifiers such as decision tree classifiers or neural networks	Agriculture	Survey of existing work	Appropriate use of the large number of features in remotely sensed data and selection of the best classifier	2020
[50]	Classification algorithm of K nearest neighbor	Soil and crop dataset	Soil quality analysis of to suggest crops	It maps soil and crop data that are suited for the soil, as well as information on nutrients that are insufficient in the soil for the specific crop.	2020
[51]	Hadoop, Map Reduce, neural network, the grey wolf optimization (GWO)	Harmonized World Soil Database	Apply method for classifying soils that is effective.	A NN-GWO accuracy=90.46%. CNN accuracy= 75.3846% and KNN accuracy= 75.38%	2020
[52]	Optical spectroscopy sensors, Least-Square ANN, Random Forest, , Naïve Bayes, SVM, Decision Tree	Soils nutrients dataset of Slovenia	To improve the precision with which soil properties are predicted	The impact of the nutritional characterization, category chosen was explored, and it was discovered that using a multi-component technique resulted in superior prediction.	2021
[53]	Machine Learning, Deep Learning	Soil dataset	Survey of ML and DL application in Agriculture	Identified limitations in existing work	2021

Survey shows that descriptive and predictive analytical methods of ML are the backbone of any decision support system in different areas such as Medical, Agriculture, and Transport and so on. These techniques plays important roles to solve problems easily and so present work mainly focuses on soil data analysis using these techniques on different analytical tools for best model and tool selection for providing solution to the problem.

2.2 Scopus bibliography analysis

This work discusses the bibliometric analysis of Soil Health Analysis research activities from the Scopus database for analyzing the research in this area. Year 2013 to September 2021 are considered for this bibliography analysis work. It is found that for a given query total 602 documents are retrieved and agriculture research activities for soil data analysis using Machine learning are gradually increased from year 2013 to 2021 and maximum work is done in the year 2021. Computers and Electronics in Agriculture journal is leading among all sources. United States followed by China then India are top 3 countries leading in these research activities. Agricultural and Biological Sciences is leading in subject area analysis [24].

Data collection. Following search query is executed to retrieved Scopus documents for analysis. This search query includes following primary keywords: Soil, Data, Analysis, Machine and Learning.

```
soil AND data AND analysis AND machine AND learning
 AND PUBYEAR > 2013 AND PUBYEAR<2021 AND ( LIMIT-
TO ( SUBJAREA , "AGRI" ) OR LIMIT-
TO ( SUBJAREA , "ENGI" ) OR LIMIT-
TO ( SUBJAREA , "COMP" ) )
```

It is found in the Figure 1 that, for a given query total 602 documents are retrieved and agriculture research activities for soil data analysis using Machine learning are gradually increased from year 2013 to 2021 and maximum work is done in the year 2021.

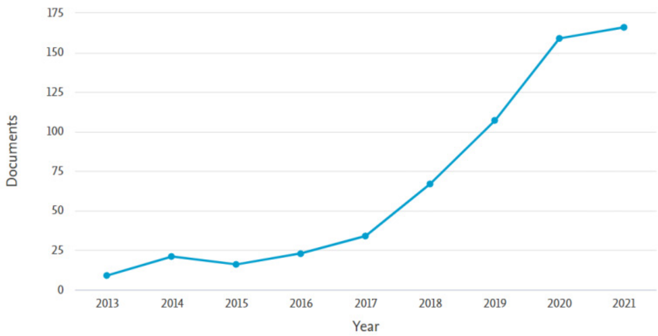

Fig. 1. Soil health Analysis using ML -Documents by year. Source: http://www.scopus.com (September 2021)

Analysis based on document type. As shown in the Figure 2 most of the work on Soil data analysis research has been published in Article papers followed by conference papers then in review papers, book chapters, etc. 69.8% work is published as articles followed by 19.9 % in conference papers.

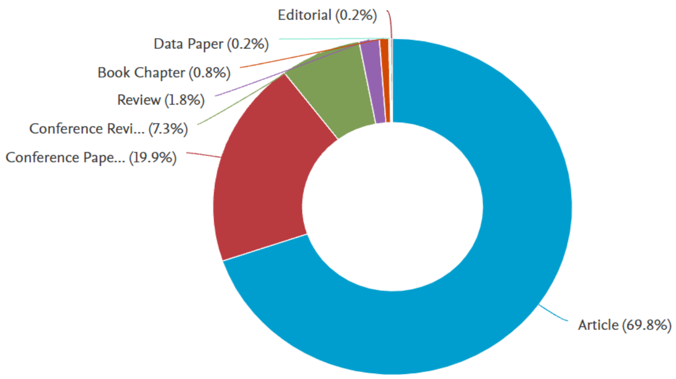

Fig. 2. Documents by Paper type. Source: http://www.scopus.com (September 2021)

Subject-based analysis. Scopus database survey in the figure 3 shows, most of the research activities are carried out in Agricultural and Biological Sciences (23%), Engineering (18%) and Computer Science (17%).

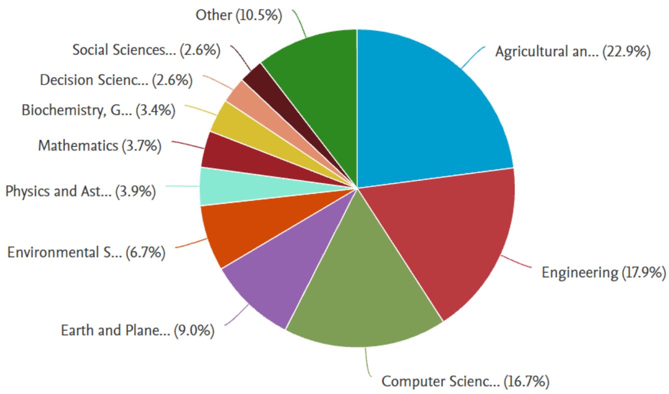

Fig. 3. Documents by Subject area

Sources-based analysis. Figure 4 depicts the document analysis by source. "Computers and Electronics in Agriculture" reported the majority of the research findings. Computers and Electronics in Agriculture journal is leading among all sources.

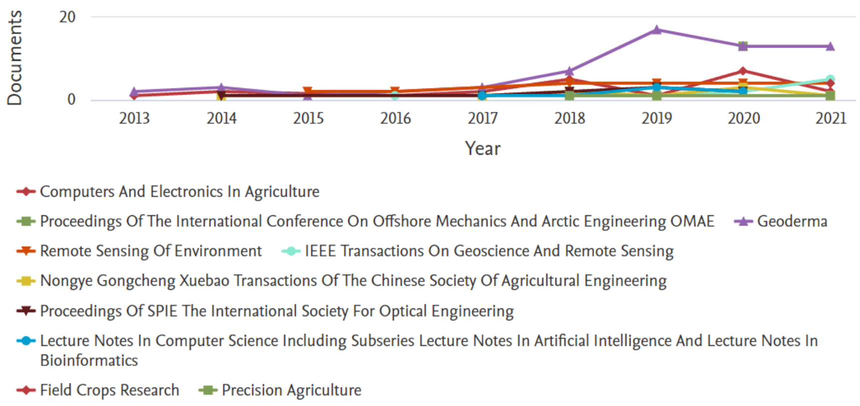

Fig. 4. Documents by year and top 10 sources, Source: http://www.scopus.com (September 2021)

Analysis based on Authors work. In the Figure 5, author's survey shows Minasny B et al. leading among all authors works.

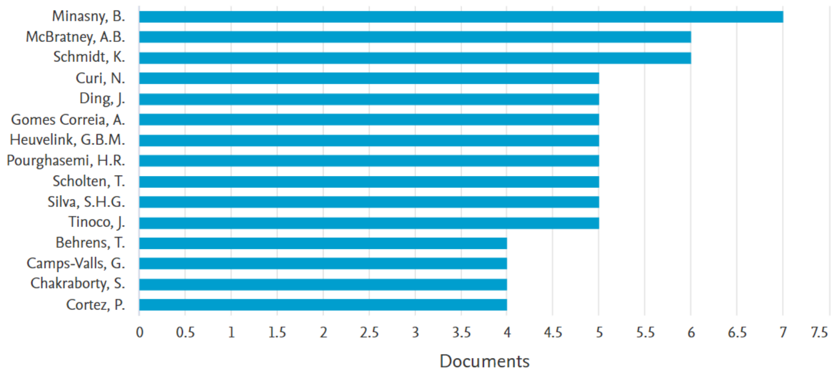

Fig. 5. Documents by Top 15 Authors, Source: http://www.scopus.com (September 2021)

Analysis by affiliations. Figure 6 shows Chinese Academy of Sciences is leading among all sources.

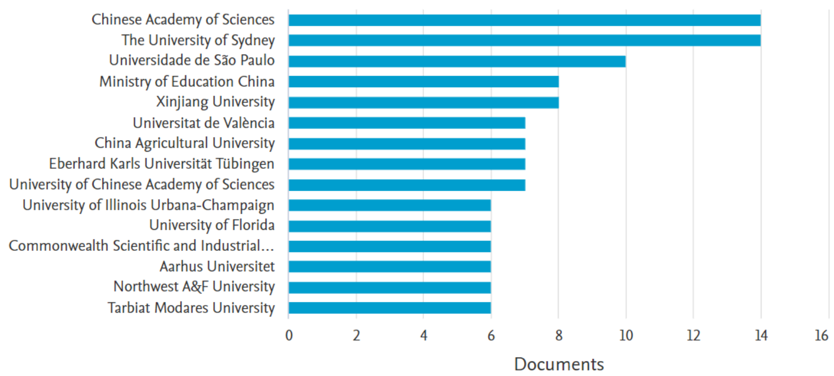

Fig. 6. Documents by top 15 affiliations, Source: http://www.scopus.com (September 2021)

Geographical region analysis. As shown in Figure 7, United States followed by China then India are top 3 countries leading in these research activities.

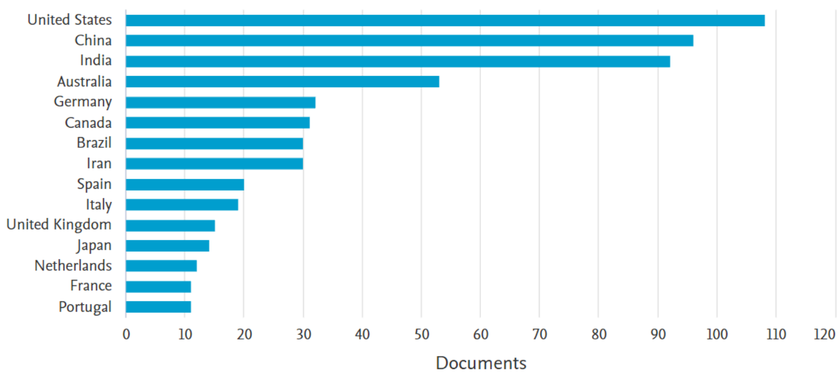

Fig. 7. Documents by top 15 countries, Source: http://www.scopus.com (September 2021)

As shown in the Figure 8 for the network analysis for cluster of co-occurrence of author keywords, most of the research work used "Machine Learning" keyword in their research activities. Second highest word is "Random Forest" followed by "Digital Soil Mapping and Deep Learning". Proposed work keywords can be identified where less work has been done.

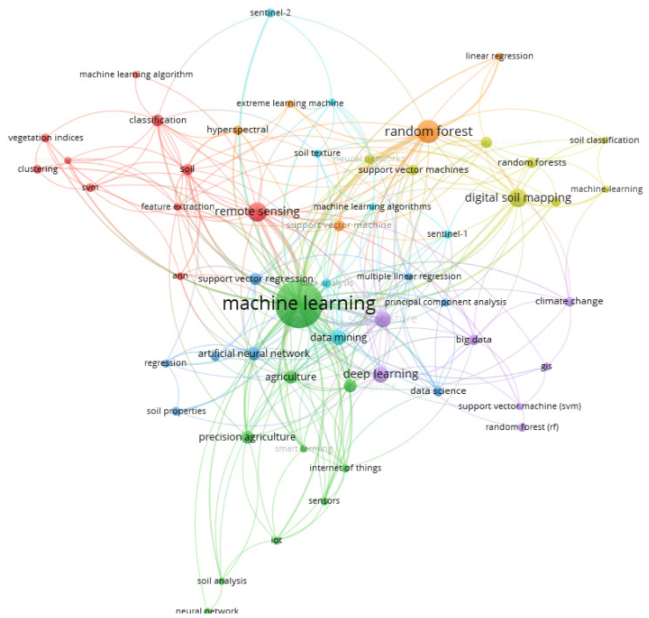

Fig. 8. Network map of Author's Keywords based on bibliographic data

3 Classifier model for soil data analysis

3.1 Classifier model

The Classification Model includes the following components for classification of new samples.

a) Input training and testing data in suitable format
b) Classifier learner to train model
c) Classifier predictor to predict class of new sample
d) Output Visualization
e) Performance scorer for model evaluation

Figure 9 shows the classifier model for Naïve Bayes classifier and designed using KNIME tool. In Naïve Bayes classifier model's design, two file readers are used; represented by Node1 and Node2. One file reader used for providing training dataset to the Naïve Bayes learner and second for providing testing dataset to the Naïve Bayes Predictor. The NB learner is used to train the model, and the predictor is used to predict class labels in the testing dataset. There are two inputs for predictor one is output of NB learner and second input is from file reader (Node2) for testing dataset. With the help of NB learner, predictor predicts the class labels of testing dataset. Interactive table is used to visualize the output of the predictor. In KNIME tool scoring nodes are available to measure the accuracy of different models [26]. There are 3 types of scorer available. Scorer for classifier with categorical outputs: Confusion Matrix, Accuracy, F-Score etc.; Numeric scorer for numerical outputs: R2, MSE etc.; Entropy scorer for clustering output.

Following classification models are implemented in the work using SK-Learn, KNIME, WEKA and Rapid miner tools.

a) Decision Tree Soil health classifier
b) Naïve Bayes Soil health classifier
c) Random Forest Soil health classifier
d) Ensemble Soil health classifier
e) KNN Soil health classifier

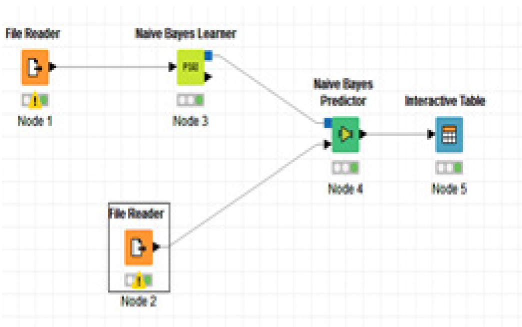

Fig. 9. Naive Bayes classifier model using KNIME tool

3.2 Mathematical model for Naive Bayes classifier

Naive Bayes falls under Supervised Classifier, where we provide a training dataset with the correct answers. Training samples are used to develop a model for predicting correct answers of new queries [1]. The Naive Bayes Classifier classifies the most likely class label by given attribute values a1, a2, ..., an. Naive Bayes is a conditional probability model. This results in the equation (1) given below:

$$p\ (C_k \mid a_1, \ldots, a_n) \tag{1}$$

System representation

S2= {Is, Es, I, O, Fu}

Where,

Is = Initial State: Input samples for classification

Es = End State: Classified samples with decision

I = Input to the Learner in formats such as ARFF, CSV, XLS.

O = Output from predictor: Classified samples.

Fu = NaïveBayesLearner_function(), NaïveBayesPredictor_function( ), NBScorer().

Equation (2) shows a formula for the Naïve Bayes conditional probability model.

$$p(Ck\ |a) = \frac{p(Ck)\ p(a\ |Ck)}{p(a)} \tag{2}$$

3.3 Mathematical model for Decision Tree classifier

The Decision_Tree_learner in KNIME produces a decision tree for making decisions [1]. Decision Tree classifier is based on below 3 main equations:

- Amount of information - I(p,n)
- Entropy- ET
- Information Gain- IG

Consider,

- Dataset contains S set of examples,
- Assume C and D is the two classes.
- c denotes C class elements and d denotes D class elements

As a result, the amount of information I(c, d) is given by equation (3).

$$I(c,d) = -\frac{c}{c+d}log2\ \frac{c}{c+d} - \frac{d}{c+d}log2\frac{d}{c+d} \quad (3)$$

Equation (4) represents formula for calculating Entropy ET for attribute A and for set of partitions w.

$$ET(A) = \sum_{i=1}^{w}\frac{pi+ni}{p+n}I(pi+ni) \quad (4)$$

Formula for calculating Information Gain (IG) is given in the equation (5).

$$IG(A) = I(p,n) - ET(A) \quad (5)$$

System representation
S2= {Is, Es, I, O, Fu}
Where,
Is = Initial State: Input samples for generating Decision Tree
Es = End State: Classified samples with decision
I = Input to the Learner in formats such as ARFF, CSV, XLS.
O = Output from predictor: Decision Tree for taking decisions, classified samples.
Fu = DecisionTreeLearner_function(), DecisionTreePredictor_function(), DTScorer().

3.4 Ensemble learning and random forest classifiers

Ensemble learning is a generic machine learning approach that aims to improve prediction performance by combining predictions from a group of models. Figure 10 shows basic ensemble model architecture, Where M= Models, P= Predictions. Architecture includes cluster of n models M1, M2…. Mn and Predictions from each model P1, P2…Pn. Voting algorithm is applied to generate final prediction. Random forest is an ensemble learning-based supervised machine learning technique, which consists of cluster of decision trees to generate final prediction.

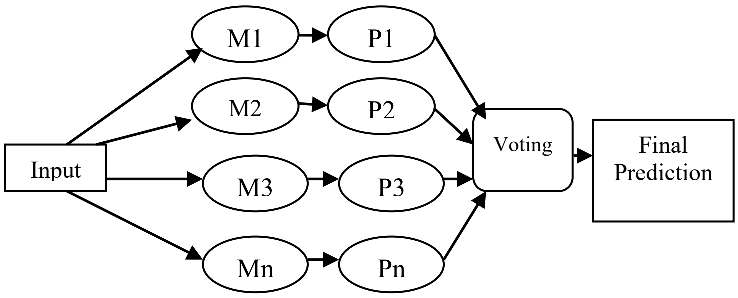

Fig. 10. Basic ensemble model architecture

3.5 Mathematics for KNN classifier

KNN is a straightforward method that maintains all available examples and categorizes new ones using a similarity metric (e.g., distance functions). Here k indicates number of neighbors. An object is classed by a majority of its neighbors, with the object being allocated to the class with the most members among its k closest neighbors [12]. KNN has the following basic steps:

1. Calculate distance using one of the Distance measure (Euclidean or Manhattan)
2. Locate the K nearest neighbours
3. Labels are up for a vote.

Equation (6) shows formula for Euclidean Distance measure for calculating distance between objects P and Q.

$$ED(P,Q) = \sqrt{\sum_{i=1}^{k}(Pi - Qi)^2} \quad (6)$$

Equation (7) shows formula for Manhattan Distance measure.

$$ED(P,Q) = \sum_{i=1}^{k} | Pi - Qi | \quad (7)$$

4 Dataset and result discussion

The dataset has the following soil parameters with Class label as a Soil quality as shown in Table 6. Dataset is collected from following sources:

- Agriculture office Pune
- Agro-assistant (Khed sub-district)
- www.soilhealth.dac.gov.in

Total 2718 training data samples are used for developing models and testing results. Preprocessing is done for feature selection and converting the dataset into suitable format. Table 3 shows a sample training dataset.

Table 3. Sample training dataset

Sr. No	N	P	K	label
0	919.8	13.6	332.69	1
1	693	13.6	509.07	4
2	617.4	13.16	829.29	0
3	667.8	13.6	734.37	0
4	7.56	13.38	318.96	2
5	756	13.38	318.96	1
6	894.6	13.38	268.26	1
…	…	..	…	…

Table 4 shows training data accuracy obtained by SK-Learn, KNIME, WEKA and Rapid miner tools for different ML algorithms.

Table 4. Algorithmic analysis using analytical tools

Classifier	Tool	DT	NB	RF	EL	KNN
Correctly Classified Instances	KNIME	1986	1852	1840	1795	1952
Incorrect Classified Instances	KNIME	732	866	878	923	766
Accuracy (%)	KNIME	73.07	68.14	67.70	66.04	71.81
	WEKA	72.87	68.05	73.05	73.06	71.84
	Rapid Miner	70.05	67.07	71.34	68.05	69.86
	SK-Learn	98.40	69.50	85.00	89.00	95.50

Figure 11 shows accuracy obtained using KNIME tools for different classifiers such as KNN, Decision Tree, Ensemble Learning Random Forest and Naïve Bayes. Here Decision tree out performs followed by KNN, Naïve Bayes and so on.

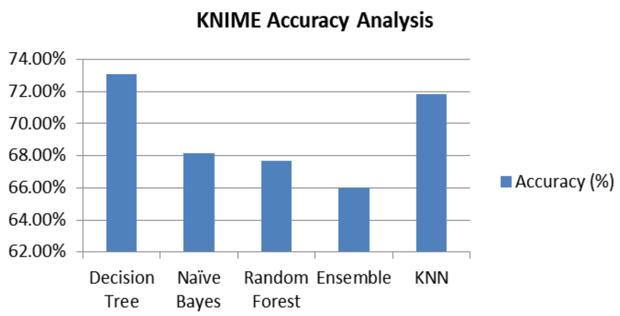

Fig. 11. Accuracy analysis using KNIME tool

Figure 12 shows analysis of classified instances by the different classifiers using KNIME tool.

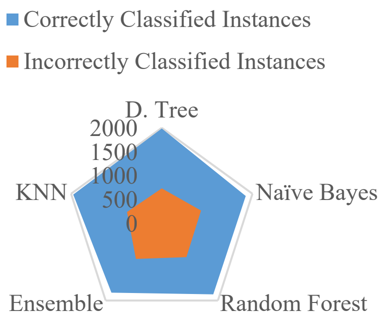

Fig. 12. Classified instance analysis using KNIME tool

Figure 13 shows accuracy obtained using WEKA tool for different classifiers. Here Random Forest out performs followed by Decision Tree, then KNN and so on. Figure 14 shows accuracy obtained using Rapid Miner tool for different classifiers. Here also Random Forest out performs followed by Decision tree and so on. Figure 15 shows analysis of classified instances by the different classifiers using Sci-Kit Learn.

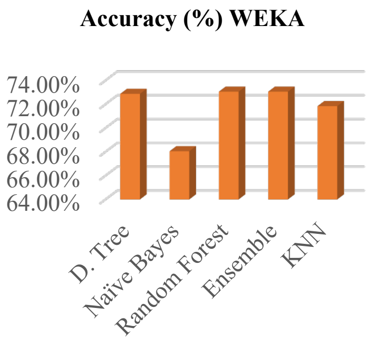

Fig. 13. Accuracy analysis using WEKA

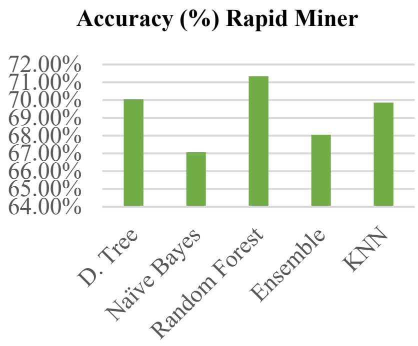

Fig. 14. Accuracy analysis using Rapid Miner

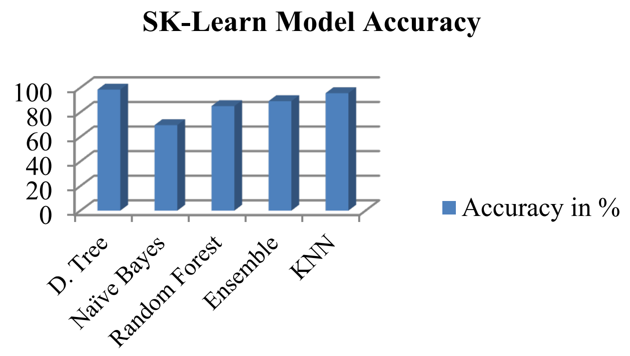

Fig. 15. Accuracy analysis using SK-Learn library

Figure 16 shows comparative analysis of accuracy obtained using SK-Learn, WEKA, Rapid Miner and KNIME tools for different classifiers such as Decision Tree, KNN, Ensemble Learning, Random Forest and Naïve Bayes, together.

Decision Tree algorithm's accuracy in forecasting soil quality is highest as compared to all classifiers followed by Random Forest. Results shows, overall accuracy of algorithms is better in SK-Learn followed by WEKA tool as compared to KNIME and Rapid Miner so SK-Learn and WEKA tool can be selected for proposed work on soil data. Maximum accuracy obtained by Decision Tree algorithm is 98.40% using SK-Learn followed by KNIME tool with 73.07% accuracy, Maximum accuracy obtained by Naïve Bayes algorithm is 69.50% using SK-Learn followed by KNIME tool with 68.14% accuracy, maximum accuracy obtained by Random Forest algorithm is 85.00% using SK-Learn followed by 73.06% using WEKA tool, maximum accuracy obtained by Ensemble algorithm is 89.00% using SK-Learn followed by 73.06% using WEKA tool and for KNN it is 95.50% using SK-Learn followed by 71.85% using WEKA tool.

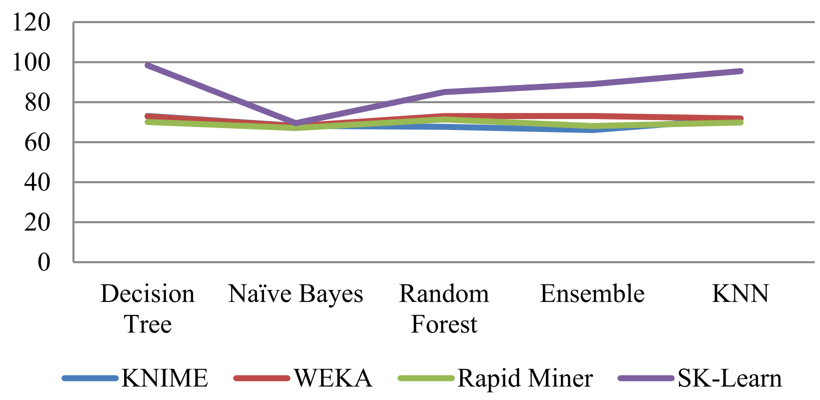

Fig. 16. Tool’s accuracy comparative analysis

5 Conclusion

In this paper Agriculture Soil data Analysis for Soil health prediction has been done using KNN, Naïve Bayes, Decision Tree, Random Forest and Ensemble Learning algorithms on SK-Learn, WEKA, Rapid Miner and KNIME tools. Also, paper represents bibliometric analysis for research data retrieved from Scopus database. Results show that Decision Tree model outperforms other algorithms, followed by Random Forest algorithm and SK-Learn gives better accuracy followed by WEKA than Rapid Miner and KNIME tools. The work is limited to only four analytical tools and limited 5 machine learning algorithms. Analysis of soil dataset can be further tested on different tools such as R language, Orange etc. and also model can be built and can be tested for different machine learning algorithms such as associative classifier, deep learning etc. to find the more accurate solution to the agriculture problem using Artificial Intelligence technology.

6 References

[1] J. Han and M. Kamber. (2011) Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. The Morgan Kaufmann Series in Data Management Systems.

[2] G. Nasrin Fathima and R. Geetha. (2014). Agriculture Crop Pattern Using Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering.

[3] S.Pudumalar , E.Ramanujam , R.Harine Rajashreen , C.Kavyan , T.Kiruthikan , J.Nishan. (2016). Crop Recommendation System for Precision Agriculture. Eighth International Conference on Advanced Computing (IcoAC), 978-1-5090-5888-4/16/$31.00@2016 IEEE.

[4] Cruz, Geraldin B. Dela et al. (2014). Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining. International Journal of Modeling and Optimization 4: 375-382. https://doi.org/10.7763/IJMO.2014.V4.404

[5] K. M. A. Patel and P. Thakral (2016). The best clustering algorithms in data mining. International Conference on Communication and Signal Processing (ICCSP), pp. 2042-2046. https://doi.org/10.1109/ICCSP.2016.7754534

[6] M.C.S. Geetha. (2015). 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. https://doi.org/10.17148/IJARCCE.2015.44119

[7] R. Zhong and H. Wang. (2011). Research of Commonly Used Association Rules Mining Algorithm in Data Mining. International Conference on Internet Computing and Information Services, pp. 219-222. https://doi.org/10.1109/ICICIS.2011.63

[8] Karan Kansara, Vishal Zaveri, Shreyans Shah, Sandip Delwadkar, Kaushal Jani. (2015). Sensor based Automated Irrigation System with IOT. International Journal of Computer Science and Information Technologies, Vol. 6(6), 5331-5333.

[9] U. K. Dey, A. H. Masud and M. N. Uddin (2017). Rice yield prediction model using data mining. International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 321-326. https://doi.org/10.1109/ECACE.2017.7912925

[10] P. Vinciya ,Dr. A. Valarmathi. (2016). Agriculture Analysis for next Generation High Tech farming in Data Mining. International Journal of advanced research in computer science and software engineering.

[11] B. Abishek, R. Priyatharshini, M. A. Eswar and P. Deepika (2017). Prediction of effective rainfall and crop water needs using data mining techniques. IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 231-235. https://doi.org/10.1109/TIAR.2017.8273722

[12] J. M. Brown. (2017). Predicting math test scores using K-Nearest Neighbor. IEEE Integrated STEM Education Conference (ISEC), pp. 104-106. https://doi.org/10.1109/ISECon.2017.7910221

[13] S. Sivaranjani, S. Sivakumari and M. Aasha. (2016). Crime prediction and forecasting in Tamilnadu using clustering approaches. International Conference on Emerging Technological Trends (ICETT), pp. 1-6. https://doi.org/10.1109/ICETT.2016.7873764

[14] Chen, Min et al. (2017). Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE Access 5 (2017): 8869-8879. https://doi.org/10.1109/ACCESS.2017.2694446

[15] S. Nagini, T. V. R. Kanth and B. V. Kiranmayee. (2016). Agriculture yield prediction using predictive analytic techniques. 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 783-788. https://doi.org/10.1109/IC3I.2016.7918789

[16] W. Fan, C. Chong, G. Xiaoling, Y. Hua and W. Juyun (2015). Prediction of Crop Yield Using Big Data. 8th International Symposium on Computational Intelligence and Design (ISCID), pp. 255-260. https://doi.org/10.1109/ISCID.2015.191

[17] R. Sujatha and P. Isakki. (2016). A study on crop yield forecasting using classification techniques. International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), pp. 1-4. https://doi.org/10.1109/ICCTIDE.2016.7725357

[18] S. Deshmukh, D. Dhannawat, M. Dalvi, P. Gawali, S. R. Vispute, S. Kekane (2019). Application of Data Analytics in Agriculture Sector for Soil Health Analysis: Literature Review. 5th International Conference on Computing, Communication, Control and Automation, pp. 1-4. https://doi.org/10.1109/ICCUBEA47591.2019.9129104

[19] Pratik Gawali, Mohit Dalvi, Devesh Dhannawat, Samihan Deshmukh, SR Vispute. (2020). An Application of Data Analytics in Agriculture Sector for Multi-advice Generator in Native Language. Journal of Critical Reviews, 2020, 7(19):2389- 2394.

[20] S. H. Akundi, S. R, and M. PM. (2020). Big Data Analytics in Healthcare using Machine Learning Algorithms: A Comparative Study. Int. J. Onl. Eng., vol. 16, no. 13, pp. pp. 19–32, Nov. https://doi.org/10.3991/ijoe.v16i13.18609

[21] Jay Gholap. (2012). Performance Tuning Of J48 Algorithm For Prediction Of Soil fertility. Asian Journal of Computer Science and Information Technology,Vol 2,No. 8,2012.

[22] E. Hot and V. Popović-Bugarin. (2015). Soil data clustering by using K-means and fuzzy K-means algorithm. 23rd Telecommunications Forum Telfor (TELFOR), pp. 890-893. https://doi.org/10.1109/TELFOR.2015.7377608

[23] D Labhade, N Lakare, A Mohite, S Bhavsar, S Vispute, G Mahajan. (2019). An Overview of Machine Learning Techniques and Tools for Predictive Analytics. Asian Journal For Convergence In Technology (AJCT) 5 (3), 63-66.

[24] Scopus database: www.scopus.com (Data accessed till September 2021)

[25] VOSviewer download website: https://www.vosviewer.com/download

[26] http://www.knime.com/knime-analytics-platform

[27] M. Yashwanth, M. L. Chandra, K. Pallavi, D. Showkat and P. S. Kumar. (2020). Agriculture Automation using Deep Learning Methods Implemented using Keras. IEEE International Conference for Innovation in Technology (INOCON), pp. 1-6. https://doi.org/10.1109/INOCON50539.2020.9298415

[28] K. Kulkarni and P. A. Vijaya (2017). A comparitive study of land classification using remotely sensed data. International Conference on Computing Methodologies and Communication (ICCMC), pp. 36-41. https://doi.org/10.1109/ICCMC.2017.8282720
[29] R. C. Margin and D. Gorgan. (2020). Qualitative Classification of Local Satellite Data. IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 589-596. https://doi.org/10.1109/ICCP51029.2020.9266198
[30] A. I. C. David and M. L. C. Guico. (2018). Presence or Absence of Fusarium oxysporum f. sp. cubense Tropical Race 4 (TR4) Classification Using Machine Learning Methods on Soil Properties," TENCON 2018 - IEEE Region 10 Conference, 2018, pp. 0689-0694. https://doi.org/10.1109/TENCON.2018.8650116
[31] V. Vanarase, V. Mane, H. Bhute, A. Tate and S. Dhar. (2021). Crop Prediction Using Data Mining and Machine Learning Techniques. Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1764-1771. https://doi.org/10.1109/ICIRCA51532.2021.9544724
[32] S. Suhag, N. Singh, S. Jadaun, P. Johri, A. Shukla and N. Parashar. (2021). IoT based Soil Nutrition and Plant Disease Detection System for Smart Agriculture. 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). pp. 478-483. https://doi.org/10.1109/CSNT51715.2021.9509719
[33] S. Shivhare and K. Cecil. (2021). Automatic Soil Classification by using Gabor Wavelet & Support Vector Machine in Digital Image Processing. Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1738-1743. https://doi.org/10.1109/ICIRCA51532.2021.9544897
[34] M. Pawar and G. Chillarge. (2018). Soil Toxicity Prediction and Recommendation System Using Data Mining In Precision Agriculture. 3rd International Conference for Convergence in Technology (I2CT), pp. 1-5. https://doi.org/10.1109/I2CT.2018.8529754
[35] S. Veenadhari, B. Misra and C. Singh (2014). Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics, pp. 1-5. https://doi.org/10.1109/ICCCI.2014.6921718
[36] S. Paul, A. Arunachalam, D. Khodadad and O. Rubanenko (2020). Fuzzy Tuned PID Controller for Vibration Control of Agricultural Manipulator. International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1-5. https://doi.org/10.1109/HORA49412.2020.9152848
[37] J. V. T. Medalla. (2018). Application of Wavelet Technique in Image Fusion and its Introduction as an Early Detection Tool for Spreading of Plant Pests in Philippines' Agricultural Sector: Initial Stage. IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2018, pp. 1-8. https://doi.org/10.1109/HNICEM.2018.8666424
[38] M. D'Orazio et al. (2021). Electro-optical classification of pollen grains via microfluidics and machine learning. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2021.3109384
[39] Z. Ünal (2020). Smart Farming Becomes Even Smarter With Deep Learning—A Bibliographical Analysis. IEEE Access, vol. 8, pp. 105587-105609. https://doi.org/10.1109/ACCESS.2020.3000175
[40] S. Srisruthi, N. Swarna, G. M. S. Ros, E. Elizabeth (2016). Sustainable agriculture using eco-friendly and energy efficient sensor technology. IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1442-1446. https://doi.org/10.1109/RTEICT.2016.7808070
[41] G. Weikmann, C. Paris and L. Bruzzone (2021). TimeSen2Crop: A Million Labeled Samples Dataset of Sentinel 2 Image Time Series for Crop-Type Classification. IEEE Journal

of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 4699-4708. https://doi.org/10.1109/JSTARS.2021.3073965

[42] N. Chahal (2015). A study on agricultural image processing along with classification model. IEEE International Advance Computing Conference (IACC), 2015, pp. 942-947. https://doi.org/10.1109/IADCC.2015.7154843

[43] S. Sarangi, S. Sharma and B. Jagyasi (2015). Agricultural activity recognition with smart-shirt and crop protocol. IEEE Global Humanitarian Technology Conference (GHTC), 2015, pp. 298-305. https://doi.org/10.1109/GHTC.2015.7343988

[44] M. D. Yang, H. H. Tseng, Y. C. Hsu and W. C. Tseng (2020). Real-time Crop Classification Using Edge Computing and Deep Learning. IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), pp. 1-4. https://doi.org/10.1109/CCNC46108.2020.9045498

[45] J. C. Puno, E. Sybingco, E. Dadios, I. Valenzuela and J. Cuello (2017). Determination of soil nutrients and pH level using image processing and artificial neural network. IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1-6. https://doi.org/10.1109/HNICEM.2017.8269472

[46] A. K. Patel, J. K. Ghosh, S. Pande and S. U. Sayyad (2020). Deep-Learning-Based Approach for Estimation of Fractional Abundance of Nitrogen in Soil From Hyperspectral Data," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 6495-6511. https://doi.org/10.1109/JSTARS.2020.3039844

[47] N. Deepa, M. Z. Khan, B. Prabadevi, D. R. Vincent P.M., P. K. R. Maddikunta and T. R. Gadekallu. (2020). Multiclass Model for Agriculture Development Using Multivariate Statistical Method. IEEE Access, vol. 8, pp. 183749-183758. https://doi.org/10.1109/ACCESS.2020.3028595

[48] Rajat Chaudhari, Saurabh Chaudhari, Atik Shaikh, Ragini Chiloba, T.D.Khadtare. (2020). Soil Fertility Prediction Using Data Mining Techniques. International Journal of Future Generation Communication and Networking Vol. 13, No. 3s, pp. 213–220.

[49] A. V. Deorankar and A. A. Rohankar. (2020). An Analytical Approach for Soil and Land Classification System using Image Processing. 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 1416-1420. https://doi.org/10.1109/ICCES48766.2020.9137952

[50] A.K.Mariappan, Ms C. Madhumitha, Ms P. Nishitha, Ms S. Nivedhitha. (2020). Crop Recommendation System through Soil Analysis Using Classification in Machine Learning. International Journal of Advanced Science and Technology Vol. 29, No. 03, pp. 12738 – 12747.

[51] Manjula, Aakunuri and Narsimha, G. (2017). Using an Efficient Optimal Classifier for Soil Classification in Spatial Data Mining Over Big Data. Journal of Intelligent Systems, vol. 29, no. 1, 2020, pp. 172-188. https://doi.org/10.1515/jisys-2017-0209

[52] Trontelj ml. J, Chambers O. (2021). Machine Learning Strategy for Soil Nutrients Prediction Using Spectroscopic Method. Sensors. 21(12):4208. https://doi.org/10.3390/s21124208

[53] Rao I., Shirgire P., Sanganwar S., Vyawhare K., Vispute S.R. (2022). An Overview of Agriculture Data Analysis Using Machine Learning Techniques and Deep Learning. In: Chen J.IZ., Tavares J.M.R.S., Iliyasu A.M., Du KL. (eds) Second International Conference on Image Processing and Capsule Networks. ICIPCN 2021. Lecture Notes in Networks and Systems, vol 300. Springer, Cham. https://doi.org/10.1007/978-3-030-84760-9_30

[54] Geng, L., & Dong, T. (2017). An Agricultural Monitoring System Based on Wireless Sensor and Depth Learning Algorithm. *International Journal of Online and Biomedical Engineering (iJOE)*, *13*(12), pp. 127–137. https://doi.org/10.3991/ijoe.v13i12.7885

7 Authors

Sushma Vispute is a research scholar in Poornima University, Jaipur and Assistant professor in department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India. She is ACM member and has very good research profile.

Dr. M. L. Saini is an associate professor in Department of Computer engineering, Poornima University, Jaipur, India. He has very good research profile.

Article submitted 2022-02-11. Resubmitted 2022-03-24. Final acceptance 2022-03-24. Final version published as submitted by the authors.