Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July, 2021:1622 – 1630

Supervised Machine Learning Technique to Predict Soil Health

Kushala.V.M[a], Dr. M.C. Supriya[b],Suma.N.R[c], Dr. H. R. Divakar[d],Pranith Jain[e]

[a] Research Scholar, Sri Siddhartha Academy of Higher Education & Research, Tumkur,Karntaka,India. (email:vmkushala@gmail.com)
[b] Professor,Dept. MCA, Sri Siddhartha Academy of Higher Education, Tumkur,Karnataka,India, (email:supriya.mc9@gmail.com).
[c]Assistant Professor, Dept. of MCA, Bangalore Institute of Technology, Bangalore, Karnataka, India. (nrsuma2006@gmail.com)
[d]Assistant Professor, Dept. of MCA, PES College of Engineering, Mandya, Karnataka, India. (email:divakarhr@gmail.com)
[e]IDA Automations,Bangalore,Karnataka,India,(email:jpranit90@gmail.com)

Abstract

In India agriculture is one of the major field which is been ignored by technical touch. Applying Artificial intelligence derivatives like Machine learning and Deep Learning to agricultural practices helps in maximum production of crops and maintains field's soil health. Health of agricultural field mainly involves maintaining of soil nutrient like chemical and physical property of soil by properly channelizing the supplements. If soil health is managed scientifically, it progressively helps in high production of yield and long life of cultivation land.

Ontology is built on the soil data collected from soil testing centers. Ontology is built in a way which exhibits the knowledge and relationship between soil and its chemical nutrient. The knowledge base is then considered to relate the nutrient and the soil type.

To manage soil health and classify them to healthy and unhealthy class machine learning comes with handy and best in class algorithms. In this study evident algorithms of machine learning are used to classify the soil efficiently to two classes healthy and unhealthy.

Algorithms like logistic regression, Decision tree, Random tree classifier, Support Vector Machine and XGBoost was used to classify the data and their algorithmic efficiency was increased by hyper parameter tuning by using different techniques.

Keywords: Soil health, chemical fertility, Supervised Learning, SVM, Decision Tree, Logistic Regression, Ensemble technique

1. Introduction

It is evident that the soil nutrient is damaged by high usage chemical fertilizers. It is recommended that less usage of chemicals to soil and replacing organic fertilizers will help the soil rejuvenating itself and produce high yield. It is very important to educate farmers to replace chemical fertilizers to organic fertilizers.

When we consider soil data as knowledge base which can be later used to make decisions on maintaining soil health on the knowledge gathered. These kinds of data are highly unstructured. They are unstructured data as they are not organized and it's very difficult to establish the relation between these unstructured data and

make decision on these data. Ontology plays a very important role in knowledge management by creating a framework. It gives an evident and efficient understanding of knowledge stored to both humans and computer toprocess the knowledge to information. Ontology describes the knowledge stored using class, axioms, function, relations and instances. Ontology works on the bases of three rules they are acquisition, storage and reuse. Using this method of knowledge storage for maintaining agricultural aspects like managing nutrients of soil, fertilizers are more advantageous and easily processed by machine learning and deep learning algorithm. This method helps in depicting how machine learning model predicts whether soil is healthy or unhealthy for the crops.

2. Literature Survey

Farmers can test their soil many numbers of times during their cultivation time to track the fertility of the soil and maintain the nutrient of the soil [1]. By considering this theory, prediction on type of crop to be grown by accounting the fertility of soil was made based on usage of machine learning algorithm. The data set collected by them contained all the chemical property of soil discussed above, along with that they considered the soil texture and temperature of the soil to predict the type of the crop that the soil facilitates the farmers to grow. The algorithm used in this prediction was the usage of supervised machine learning, which learns the data based on the labels present in the data set by considering the target variable in the same data set. It is mentioned that supervised learning can be used for classification and regression problems.The data set is differentiated into two types one as training which is dedicatedly used for training the prediction and the other for testing which is not used for training but to test the prediction accuracy. Using this concept, the Tamil-Nadu data set was collected along with type crops cultivated in that area. By analysing the training data i.e. soil property and considering the target variable as crops to be grown, the model was built efficiently which predicted the type of crop to be grown within an hour. This model was also efficient in predicting the type of fertilizers to be used in cultivation period.

Nitrogen is considered as the major source of nutrient for the growth of the plant as it's directly involved in the reaction of photosynthesis [2]. Using Fuzzy algorithms and k- mean algorithm nitrogen is managed in the fields by creating zones and the optimal levels are managed in the field. Machine learning technique was used in hyper spectral image data to review physical and structural characteristics in plants and understand its physical effects by the external environment.

ML technique was successfully used in early identification of weed, plant diseases, insects using ANN (Artificial neural network), and Random forest algorithm. Also, it's proven that cost saving and auto decision making can be done.

Using ML techniques i.e. SVM, Random forest, extremely randomized trees and Deep learning corn yield production was successfully estimated.

Soil Knowledge based built using ontology assist the search of soil which is stored in various sources [3]. Ontology helps in providing knowledge of specific domain by building relationship among the objects in the form of classes and subclasses. The soil knowledge was built on the basis of feature extraction and storing of the knowledge base, where the unstructured data is processed and cleaned by considering the important features and storing them as knowledge. Automatic term weight processing is used to weightage the soil terms automatically which helps in retrieval of the soil. Lastly using the XML and Xpath algorithm the retrieval of the knowledge is made.

Deep learning (DL) which is concluded to be more efficient in prediction of the complex structural data, is based on the human brain structure. Where the model built using DL has multiple layer which process the information in every layer to give the output. Precision agriculture is the most advanced way of cultivation which involves adoption of many technologies. Amy, John used DL to predict the yield and protein of wheat based on fertilization [4]. They used special kind of ANN called as Stacked Encoder. Here there is an involvement of phases. In the first auto encoder is trained with input considering input as also target variable. In second phase separate hidden layer is stacked to the auto encoder of the first phase, which makes the hidden layer values map to the values of auto encoder which makes the hidden layer construct the new values. It proves the usage of fertilizers creates a saturation in the production of yield and doesn't improve production. This inversely causes in the emission of high nitrogen in the environment. It's important to manage the fertilizers. Using DL algorithms like linear regression, non-linear regression, ANN and SAE were used to predict the protein and yield production estimation.

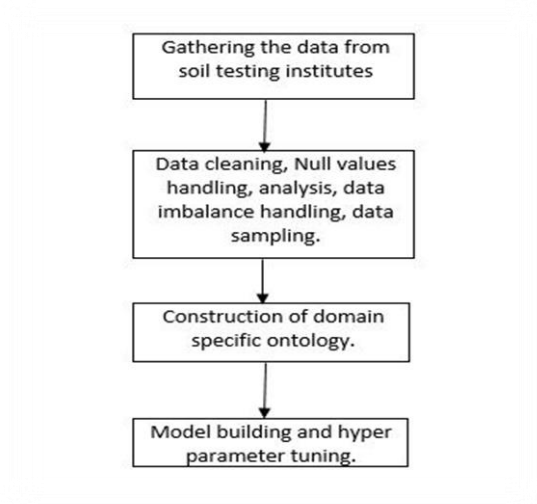
As many ontologies emerged with huge knowledge base for agriculture, it became very difficult to dig the huge ontology which was a combination of n-dimensions of ontology. They came up with more supervised ontology model called as AgroPortal, vocabulary for agronomy [6]. This model reused the Nation centre biomedical ontology to build AgroPortal. They successfully build which performed on hosting, searching,

commenting, recommendation, interoperation for the semantic web. They successfully implemented ontology for all the specific requirement of agronomy. Ontology applications were implemented in many fields of agriculture few are Semantic Web portal in sustainable agriculture this ontology is dedicated for the betterment of agriculture in France. This not only involves farmers but also the state community in improving the agriculture. It has two phases where it has a query processing phase to search for the input and it matches the input taken from framers to detect the type of problem they are facing [9]. Reduce in use of pesticides were achieved. The system involved semantic search results.

Domain ontology is an ontology which has a dedicated knowledge to a particular field with the same dedicated field terminology. Task ontology in combination with domain ontology explains how the tasks (procedure) which are performed or involved in the domain to make the model complete. Building a domain specific in combination with task ontology helps in more understanding and interpretation of any field. On considering the same advantage a domain specific ontology was built for maintaining crop cultivation standards [7]. An ontology was built maintain a crop cultivation process which involved complete life cycle of plant growth to production. The domain ontology contained the type of crop, fertilizer needed, soil type, climatic condition, growth time which is the basic idea of the crop growth. Task ontology was combined with domain with V-shape structure explaining the tasks that need to be followed. The task included like, how to plant, water and fertilize etc.

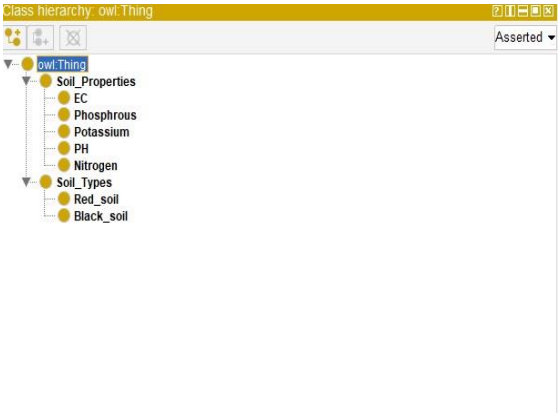
The logical analysis and decision making from the stored data is a phenomenon process in model using machine learning and deep learning. Knowledge representation is better done using ontology as it provides relationship between concepts, describes the concepts and classes. Logic based knowledge representation and reasoning using machine learning and deep learning is still a channel and has not given evident results [8]. The knowledge representation and reasoning which is the main source of data used by artificial intelligence was implemented using ontology efficiently, then the reasoning was made using recursive reasoning network (RRN). The RRN was trained against the ontology built which is capable of encoding all the information available about the domain.

3. Methodology



In this paper implementation was aimed to build domain ontology which contains chemical nutrients of soil which is collected in and around of Mysore District and which are tested at soil testing centres. With these data the soil can be categorised as red soil, black soil. The soil was tested for PH, EC, Potassium, Nitrogen and Phosphorous. The intension of the ontology is to facilitate the structured knowledge of soil nutrient of Mysore district.

The ontology has two entity soil type and soil properties. The below figure depicts the hierarchy of the built ontology. Property class contains and exhibits all the property that was tested for at soil centre and the type of soil is categorised on the data collected.



The object property depicts the relationship between the individuals in the ontology. The relation that exists here are soil types class entities Red soil and Black soil hasProperties EC, pH, Phosphorous, Potassium and Nitrogen. This relationship exhibits inverse-of as special property. The data property determines the type of data literals linking the entities. The data property for soil name is defined as strings and property class defined for pH,EC and inkgs pH defined as float representing pH value, EC defined as float representing EC value, inkgs defined as float representing Nitrogen, Phosphorus, Potassium present in the soil.

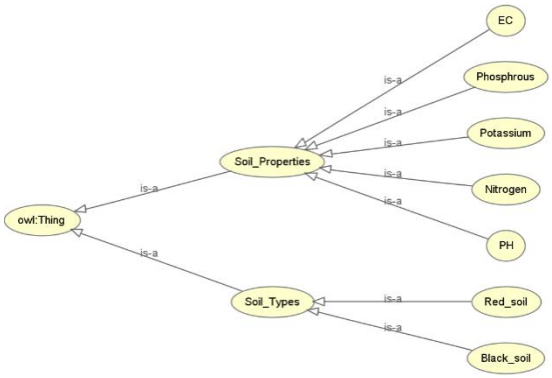


Fig 2. Asserted class hierarchy and the inferred class hierarchy produced by protégé.

The above figure demonstrates the class pecking guidelines in an OWL cosmology to be seen and incrementally extended of the asserted class sequence of control and gather class evolution.

A. Data Overview

The soil data which was collected was examined for the major type of soil from the region it was collected and it projected major part of land containing red soil with 57% and black soil covering 43% of cultivation land.

The collected data dominated with highest number of unhealthy soil by belonging to class 0. Total 87 percent of red soil was unhealthy with 21% being healthy. Whereas 84 percent of black soil was found unhealthy with 15 percent being healthy. The above analysis on the data highly indicates there is an imbalanced classification ratio between healthy and unhealthy soil when model trained on this data the model is said to produce high accuracy low recall model. This was evident in a model we built which resulted in high accuracy with 0 precision and recall value. The data was hence handled manually to balance the healthy and unhealthy class equally.

B. Model evaluation methods

All the algorithms that are built are measured for accuracy, precession, recall and ROC curve and they are used as a benchmark to compare the performance.

Accuracy is defined as the number of perfectly classified class divided by the total number of predictions on the classes made.

Confusion matrix of a model looks like this. True negative signifies the number of predicted values that are actually negative, false negative signifies the number of values which are predicted as negative in spite of class being positive [4] [11]. False positive signifies the number of classes predicted as false in spite class being positive, True positive signifies the number of classes predicted as positive and actual value also being positive.

C. Algorithms

- Logistic regression is the simplest form of algorithm used in classification purpose. It does the classification on the basis of probability. The loss function used by logistic is the sigmoid function. Sigmoid function is used to map the predicted classes to probabilities [15].

$$f(x) = \frac{1}{1 + e^{-x}}$$

A threshold value is fixed, if the probability of the value is above the threshold the value is classified to class 1 if less the value is classified to class 0. Cost function is considered as optimization objective which will effectively reduce the errors in the model.

$$\text{cost}(h_\theta(p),q) = \begin{cases} -\log(h_\theta(p)) & \text{if } q = 1 \\ -\log(1 - h_\theta(p)) & \text{if } q = 0 \end{cases}$$

The cost value is reduced using gradient descent. To reduce the cost function using gradient descent every parameter is involved. Gradient descent for any parameter can be done using the below equation.

$$\frac{\partial K(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)x_k^i$$

With the above model implementation, the model evaluation score was obtained. With tuning the hyper parameters our accuracy improved by 6 percent with better ROC curve value. The loss function combined with gradient descent & L2 (Ridge) regularization is used in tuning the model.

- Support Vector machine (SVM) being the most used and simplified algorithm in classification of the data in to classes. SVM uses a hyper plane to separate the data points to classes. The hyper plane which is drawn in the space of data points acts as a decision boundary and it is considered or drawn in such a way that the distance between the points and hyper plane is maximum [17]. When the data points are centric, not separable the data is transformed to a higher dimension space which makes it possible to understand and draw hyper plane that maximum separates the data points. The main aim is to maximize the margin between the data points and to do this we use hinge function which acts as loss function and helps to maximize the hyper plane.

The cost of this function becomes zero when the predicted value and the actual value are same. If they are not, we calculate the loss function. Along with cost function q we add regularization parameter to handle loss function along with maximization of the hyper plane.

$$p(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \text{ else} \\ 1 - y * f(x), & \end{cases}$$

Hinge loss function

$$\min_m \lambda \, ||m^2|| + \sum_{i=1}^{n}(1 - y_i \, (x_{i,m}))$$

Loss function of SVM

The weights are updated by applying partial derivation to the weights which helps to find the gradients. Using gradients, we can update the new weights.

- Decision Trees are predictive modelling approach that splits the data into different conditions in a tree format. They are non-parametric way to classify the data into classes. When target variable for decision tree is discrete then it is classification tree. Splitting of the data is layer basis homogenous data spit to one side and non-homogenous to other side. Splitting of the data can be both binary and multi-way split depending on the advantage. There are different types of DT like CART, ID3 and C4.5 by using different metrics to split the tree [16].

We have used ID3 a typical classification algorithm which uses Information Gain as a metric. Information gain refers to the amount of information a feature can give to a class. Information gain is a statistical property which can be calculated using entropy which measures the errors and randomness in the data. Measure of decrease in entropy is nothing but a highest information gain. The attribute with highest information gain is choose to be the decision criteria on splitting the node. The tree is made sure it doesn't face the over fitting by controlling the depth of the tree, it's said that DT gets more complicated and try to outperform when allowed to grow fully by splitting all the nodes. Which increases in bias of the output.

By building the classification algorithm using DT in the initial stage the accuracy score of the model was achieved with 77 percent. The model was made better by tuning it with Grid search CV method. Grid search CV

takes many hyper parameter values as an input and chooses the best one among them. The best hyper-parameters obtained by the Grid search CV is max_depth: 10, min_samples_leaf: 2, min_samples_split : 2. With these as a parameter the accuracy score was increased.

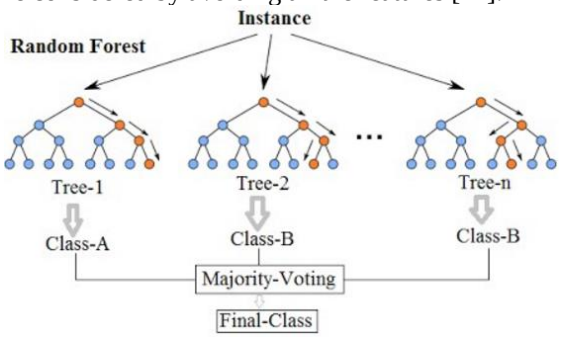
D. Advancing the algorithm using Ensemble Methodology

Ensemble is the idea about merging many models which are solving the same problem which will eventually merged together to get the best result.

Bagging: This method considers all the model which are solving the same problem and it learns from each other in parallel and combines them on some deterministic averaging process resulting in a better efficiency.

Boosting: This method considers all the model which are solving the same problem and it learns from each other sequentially and merge them on some deterministic strategy resulting in a better efficiency [10].

a) Random Forest Classifier is same as the decision tree classifier but the cleverest idea involved here is the adoption of ensemble's bagging method. Large number of trees working as a team predicting the classes randomly and not relating to each other is believed to outperform than an individual constituent tree built on the data. In DT the root node is chosen specifically to split the data but in case of RF the nodes are chosen randomly and only subset of the features are considered by avoiding all the features [14].



b) XGBoost Classifier is again a DT classification method which uses gradient boosting to improve the efficiency of the prediction. This is one of the best of algorithms which is super powerful by combining both software and hardware efficiency to reduce the computational speed and increase the efficiency of the model. XGBoost uses max depth as a specific parameter and a criteria to split the data tree and starts to prune the trees backwards. This has high efficacy by adopting cross-validation which avoids to explicitly mention. It also uses L1, L2 regression if needed to optimize the loss function [15]

4. Results And Discussion

Table I: Represents the results of all algorithms

Model	Accuracy	Precision	Recall	F1 Score	ROC
Logistic	0.6	0.285714	0.857143	0.428571	0.701299
SVM	0.725	0.25	0.2857	0.266667	0.551948
Decision Tree	0.775	0.37500	0.428571	0.4000	0.638528

Random tree Classifier	0.85	0.666667	0.285714	0.4	0.627706
XGBoost	0.8	0.444444	0.571429	0.5	0.709957

Below curve is the ROC curve of all the algorithms which is used as a metric for binary classification problems. The curve plots the true positive rate against false positive rate at various threshold values. This also separates the signal from noise. ROC curve was plotted for XGBoost, Random tree classifier as they showed better accuracy.

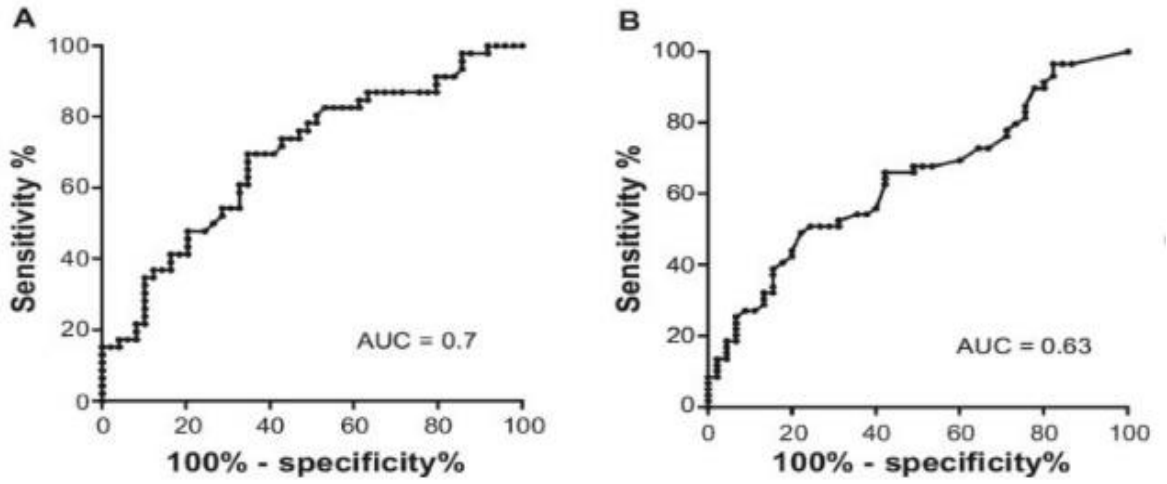


Fig 3. ROC Curve for the top two accuracy algorithms.

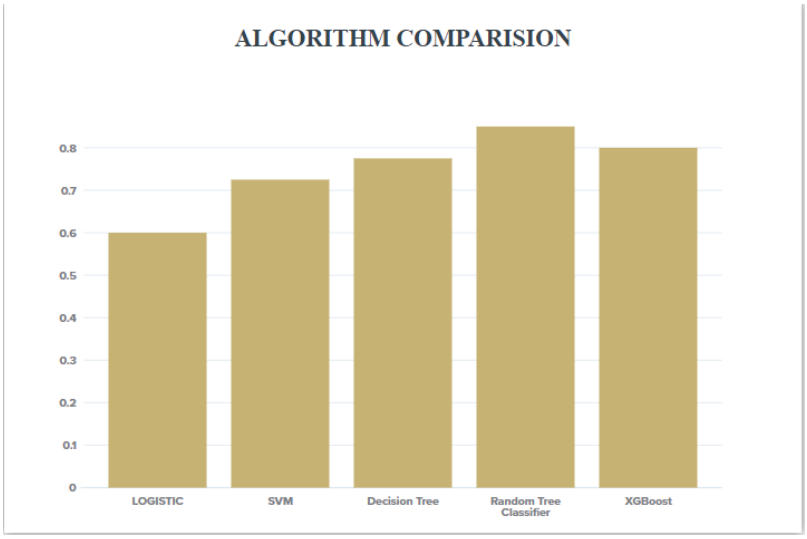


Fig 4. Bar-graph representing the performance

5. Conclusion And Future Scope

Agricultural data is completely haphazard and more of unhealthy soil is leading to the crop depreciation and yield loss. An effort is made to classify the data to healthy and unhealthy classes using machine learning algorithms. It is evident from the above results that accuracy of prediction model increased with advancement of the algorithms. Adopting to ensembles method is evident to produce better accurate results to haphazard, misclassified, unclear data's. To increase the accuracy on these types of data we can adopt new enhancements made on ensemble algorithms like LightGBM.

References

[1] Soil Analysis and Prediction of Suitable Crop for Agriculture using Machine LearningS. Panchamurthi. M. E1, M. D. Perarulalan2, A. Syed Hameeduddin3, P. Yuvaraj. International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 7 Issue III, Mar 2019.

[2] Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review Anna Chlingaryana, Salah Sukkarieha, Brett Whelan. 0168-1699/ Published by Elsevier B.V.

[3] Soil Knowledge-based Systems Using Ontology, TongpoolHeeptaisong and AnongnartShivihok. Proceeding of the international Multi conference of Engineering and computer scientist 2012 Vol I, IMECS 2012. ISBN : 978-988-19251-1-4.

[4] Using Deep Learning in Yield and Protein Prediction of Winter Wheat Based on Fertilization Prescriptions in Precision Agriculture Amy Peerlinck, John Sheppard1, Bruce Maxwell,Gianforte School of Computing, Montana State University, Bozeman, MT.Land Resources & Environmental Science, Montana State University, Bozeman, MT.A paper from the Proceedings of the 14th International Conference on Precision Agriculture Montreal, Quebec, Canada.

[5] Ontology- Based Knowledge Management System and Application JunsongZhanga , Wu Zhaoa, Gang Xieb, Published by Elsevier Ltd. Selection and/or peer-review under responsibility of [CEIS 2011.

[6] AgroPortal: A vocabulary and ontology repository for agronomy Clément Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther DzaléYeumo, Vincent Emonet, John Graybeal, Marie-AngéliqueLaporte, Mark A. Musen, Valeria Pesce, Pierre Larmande.Computers and Electronics in Agriculture 144 (2018) 126–143.

[7] An ontology-based knowledge representation and implement method for crop cultivation standard

[8] DaiyiLia, Li Kanga, XinrongChenga, DaoliangLia, LaiqingJia, KaiyiWangb, YingyiChena,Mathematical and Computer Modelling 58 (2013) 466–473.

[9] Ontology Reasoning with Deep Neural Networks, Patrick Hohenecker,Thomas Lukasiewicz,arXiv:1808.07980v3 [cs.AI] 10 Dec 2018.

[10] Ontologies in Agriculture, C. ROUSSEY, V. SOULIGNAC, J-C CHAMPOMIER, V. ABT, J-P CHANET.

[11] Crop Prediction based on Soil Classification using Machine Learning with Classifier Ensembling. Vrushali C. Waikar, Sheetal Y. Thorat, Ashlesha A. Ghute, Priya P. Rajput4, Mahesh S. ShindeStudent, M. E. S. College of Engineering Pune, Maharashtra, India Professor, Dept. of Computer Engineering, M. E. S. College of Engineering Pune, Maharashtra, India.

[12] Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning Keerthan Kumar T G, Shubha C, Sushma S A. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019.

[13] Gholap, Jay. “Performance Tuning of J48 Algorithm for Prediction of Soil Fertility.” ArXiv abs/1208.3943 (2012): n. pag.

[14] A. Arooj, M. Riaz and M. N. Akram, "Evaluation of predictive data mining algorithms in soil data classification for optimized crop recommendation," 2018 International Conference on Advancements in Computational Sciences ICACS), Lahore, 2018, pp. 1-6. doi: 10.1109/ICACS.2018.8333275.

[15] Random forest available https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[16] Xgboost available https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.

[17] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 1310-1315.

[18] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O., Akinjobi J. "Supervised Machine Learning Algorithms: Classification and Comparison". International Journal of Computer Trends and Technology (IJCTT) V48(3):128-138, June 2017. ISSN:2231-2803.

[19] An Ontology Driven System to Predict Diabetes With Machine Learning Techniques Divakar H R, D Ramesh, B R Prakash, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2, December 2019.

[20] Deep Learning Ontology: Dimensions in the Field of Agriculture, A Survey Kushala. V. M ,Dr.Supriya. M.C, International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS) Volume VII, Issue I, January 2018 | ISSN 2278-2540.

[21] D Ramesh, Divakar H R, B R Prakash "An Ontology-Based System to Predict Diabetes Using Deep Learning". DogoRangsang Research Journal UGC Care Group I Journal ISSN : 2347-7180 Vol-10 Issue-08 No. 13, August 2020.