iJRASET

International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** I **Month of publication:** January 2022

DOI: <https://doi.org/10.22214/ijraset.2022.40081>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Soil Health Prediction Using Supervised Machine Learning Technique

Pratiksha Patil

B.E. Student, Computer Engineering, Smt. Indira Gandhi College Of Engineering , Ghansoli , Navi Mumbai , Maharashtra

Abstract: Agriculture is one of the major fields in India that has been overlooked by technical touch. The application of artificial intelligence derivatives such as machine learning and deep learning to agricultural practises aids in crop production and soil health maintenance. The health of an agricultural field is primarily concerned with the preservation of soil nutrients, such as chemical and physical properties, by properly transmitting supplements. When soil health is managed scientifically, it gradually aids in high yield production and the long life of cultivation land. The soil data collected from soil testing centres is used to build the ontology. Ontology is constructed in such a way that it demonstrates the knowledge and relationship between soil and its chemical nutrients. The knowledge base is then used to connect the nutrient and soil type. Machine learning comes with useful and best-in-class algorithms for managing soil health and classifying it into healthy and unhealthy categories. In this study, obvious machine learning algorithms are used to efficiently classify the soil into two classes: healthy and unhealthy. To classify the data, algorithms such as logistic regression, Decision tree, Random tree classifier, Support Vector Machine, and XGBoost were used, and their algorithmic efficiency was increased through hyper parameter tuning using various techniques.

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

I. INTRODUCTION

It is clear that the soil nutrient is being harmed by the widespread use of chemical fertilisers. It is suggested that using fewer chemicals on the soil and replacing them with organic fertilisers will help the soil rejuvenate itself and produce a higher yield. It is critical to educate farmers on the benefits of switching from chemical fertilisers to organic fertilisers. When we think of soil data as a knowledge base that can be used to make decisions about maintaining soil health based on the information gathered. This type of data is highly unstructured. They are unstructured data because they are not coordinated, and it is extremely difficult to establish relationships between these unstructured data and make decisions based on them. By establishing a framework, ontology plays a critical role in knowledge management. It provides a clear and efficient understanding of stored knowledge to both humans and computers in order to process the knowledge into information. Ontology describes the knowledge that is stored in the form of classes, axioms, functions, relations, and instances. Ontology operates on the basis of three rules: acquisition, storage, and reuse. Using this method of knowledge storage for agricultural aspects such as soil nutrient management and fertiliser management is more advantageous and easily processed by machine learning and deep learning algorithms. This method illustrates how a machine learning model predicts whether the soil is healthy or unhealthy for crops.

II. LITRATURE SURVEY

Farmers can test their soil numerous times during the cultivation season to track soil fertility and maintain soil nutrient levels [1].Based on this theory, a prediction on the type of crop to be grown by accounting for soil fertility was made using a machine learning algorithm. They collected a data set that included all of the soil's chemical properties discussed above, as well as the texture and temperature of the soil, in order to predict the type of crop that the soil allows farmers to grow.by taking into account the target variable in the same data set on the labels present in the data set. It is stated that Supervised learning can be applied to classification and regression problems. The data set is divided into two types: training, which is used exclusively for training the prediction, and testing, which is not used for training but is used to test the prediction accuracy. The Tamil-Nadu data set was compile using this concept as well as the types of crops grown in that area. The model was efficiently built by analysing the training data, which was soil property, and taking the target variable, crops to be grown, into account, and predicted the type of crop to be grown within an hour. This model was also effective at predicting the type of fertiliser that would be used during the cultivation period. Nitrogen is regarded as the most important nutrient source for plant growth because it is directly involved in the photosynthesis reaction [2].Nitrogen is managed in the fields using Fuzzy algorithms and the k- mean algorithm by creating zones and managing the optimal levels in the field.

In hyper spectral image data, a machine learning technique was used to review physical and structural characteristics in plants and understand their physical effects by the external environment. Using ANN and Random forest algorithms, ML technique was successfully used in early identification of weeds, plant diseases, and insects. It has also been demonstrated that cost savings and automated decision making are possible. Corn yield production was successfully estimated using ML techniques such as SVM, Random forest, extremely randomised trees, and Deep Learning. Soil knowledge based on ontology aids in the search for soil stored in various sources [3]. Ontology aids in the provision of knowledge in a specific domain by establishing relationships between objects in the form of classes and subclasses. The soil knowledge was created on the basis of feature extraction and knowledge base storage, in which unstructured data is processed and cleaned by taking into account the important features and storing them as knowledge. Deep learning (DL), which is thought to be more efficient in predicting complex structural data, is based on the structure of the human brain. Whereas the model created with DL has multiple layers that process the information in each layer to produce the output. Precision agriculture is the most advanced method of cultivation, requiring the use of numerous technologies. Amy and John used DL to forecast wheat yield and protein based on fertilisation [4]. They used a type of ANN known as a Stacked Encoder. There is a phase involvement here. The first auto encoder is trained with input, taking input into account as well as the target variable. As more ontologies with large knowledge bases for agriculture emerged, it became increasingly difficult to dig the massive ontology that was a combination of n-dimensions of ontology. They developed a more supervised ontology model called Agro Portal, which is a vocabulary for agronomy [6]. AgroPortal was built using the Nation Centre biomedical ontology, which was reused in this model. They successfully implemented ontology for all agronomy-specific requirements. Few ontology applications have been implemented in many fields of agriculture. This ontology is a semantic web portal in sustainable agriculture that is dedicated to the improvement of agriculture in France. This involves not only farmers, but also the state community, in the improvement of agriculture. It has two phases: a query processing phase in which it searches for input and a matching phase in which it matches input from farmers to determine the type of problem they are facing [9]. Pesticides were used less frequently. Semantic search results were used in the system. An ontology with a dedicated knowledge to a specific field and the same dedicated field terminology is known as a domain ontology. Task ontology, in conjunction with domain ontology, explains how tasks are performed. (procedure) that are performed or involved in the domain to complete the model. Creating a domain-specific ontology in conjunction with a task ontology aids in the understanding and interpretation of any field. Considering the same benefit, a domain-specific ontology was created to maintain crop cultivation standards [7]. An ontology was created to support a crop cultivation process that included the entire life cycle of plant growth to production. The domain ontology included the type of crop, fertiliser required, soil type, climatic condition, and growth time, which is the fundamental concept of crop growth. The task ontology was combined with a domain with a V-shape structure to explain the tasks that must be completed. The task included instructions on how to plant, water, and fertilise plants, among other things. The process of logical analysis and decision making based on stored data is a phenomenon in models that use machine learning and deep learning. Ontology is a better way to represent knowledge because it provides relationships between concepts, describes concepts, and classes. Logic-based knowledge representation and reasoning using machine learning and deep learning is still an open channel with no clear results [8]. The knowledge representation and reasoning, which are the primary sources of data used by artificial intelligence, were efficiently implemented using ontology, and the reasoning was made using a recursive reasoning network (RRN). The RRN was trained against an ontology that was created and is capable of encoding all of the domain's information.

III.METHODOLOGY

The goal of this paper's implementation was to create a domain ontology that contains chemical nutrients from soil collected in and around Mysore District and tested at soil testing centres. Using these data, the soil can be classified as red soil or black soil. The soil's PH, EC, potassium, nitrogen, and phosphorus levels were all measured. The ontology's goal is to facilitate structured knowledge of soil nutrients in the Mysore district. There are two entities in the ontology: soil type and soil properties.

The built ontology's hierarchy is depicted in the figure below. The property class contains and displays all of the properties that were tested for at the soil centre, and the type of soil is classified based on the data collected.

Fig 1. Soil Health Prediction Using Supervised Machine Learning

The object property depicts the ontology's relationship between individuals. The soil types class entities Red soil and Black soil have properties EC, pH, Phosphorous, Potassium, and Nitrogen. As a special property, this relationship has inverse of. The data property specifies the type of data literals used to connect the entities. The data property for soil name is defined as strings, and the property class for pH, EC, and in kgs is defined. pH is defined as a float that represents the pH value, EC is defined as a float that represents the EC value, and in kgs is defined as a float that represents the Nitrogen, Phosphorus, and Potassium content of the soil.

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

The diagram above shows the class pecking guidelines in an OWL cosmology that can be seen and incrementally extended of the asserted class sequence of control and gather class evolution.

A. Data Overview

The soil data collected was examined for the major type of soil from the region it was collected in, and it projected a large portion of land containing red soil (57%) and black soil (43%) of cultivation land. The collected data revealed that class 0 had the highest number of unhealthy soils. In total, 87 percent of red soil was unhealthy, while only 21 percent was healthy. Whereas 84 percent of black soil was found to be unhealthy, 15 percent was found to be healthy. The data analysis presented above strongly suggests that there is an imbalanced classification ratio between healthy and unhealthy soil. When trained on this data, the model is said to produce a high accuracy low recall model. This was evident in a model we developed, which yielded high accuracy with 0 precision and recall value. As a result, the data was handled manually in order to balance the healthy and unhealthy classes equally.

B. Model Evaluation Method

All algorithms that are built are measured for accuracy, precision, recall, and ROC curve, and their performance is compared. The number of perfectly classified classes divided by the total number of predictions on the classes made equals accuracy. A model's confusion matrix looks like this. True negative denotes the number of predicted values that are actually negative, while false negative denotes the number of predicted values that are predicted as negative despite the fact that the class is positive [4] [11]. False positive denotes the number of classes predicted as false despite the fact that the class is positive. True positive denotes the number of classes predicted as positive despite the fact that the actual value is also positive.

C. Algorithms

1) Logistic regression is the most basic type of algorithm used for classification. It makes the classification based on probability. The sigmoid function is the loss function used by logistic. To map predicted classes to probabilities, the Sigmoid function is used [15].

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

A fixed threshold value is set; if the probability of the value is greater than the threshold, the value is classified as class 1, otherwise it is classified as class 0.

The cost function is regarded as an optimization objective that will effectively reduce model errors.

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

Gradient descent is used to reduce the cost value. Every parameter is involved in reducing the cost function using gradient descent. The following equation can be used to perform gradient descent on any parameter.

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

The model evaluation score was obtained using the model implementation described above. With better ROC curve value after tuning the hyper parameters, our accuracy increased by 6%. The loss function is used to tune the model, along with gradient descent and L2 (Ridge) regularisation.

2) The most widely used and simplest algorithm for classifying data is the Support Vector Machine (SVM).

SVM divides data points into classes using a hyperplane. The hyperplane drawn in the space of data points serves as a decision boundary, and it is considered or drawn in such a way that the distance between the points and the hyperplane is as short as possible [17]. When the data points are centric and cannot be separated, the data is transformed to a higher dimension space, allowing for the understanding and drawing of a hyperplane that best separates the data points. The main goal is to maximise the margin between the data points, and to do so, we use the hinge function, which acts as a loss function and aids in the optimization of the hyperplane. When the predicted and actual values are the same, the cost of this function is zero. If they are not, the loss function is computed. Along with the cost function q , we add a regularisation parameter to handle the loss function as well as the hyperplane maximisation.

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

The weights are updated by applying partial derivation to them, which aids in the discovery of gradients. We can update the new weights using gradients.

3) Decision Trees are a predictive modelling approach that divides data into different conditions in the form of a tree. They are a non-parametric method of categorising data. When the target variable for a decision tree is discrete, the tree is referred to as a classification tree. The data is split on a layer basis, with homogeneous data split to one side and non-homogeneous data split to the other. Depending on the benefit, data can be split in binary or multi-way splits. There are various types of DT, such as CART, ID3, and C4.5, that use different metrics to split the tree [16]. We used ID3, a standard classification algorithm that employs Information Gain as a metric.

- 4) The amount of information that a feature can provide to a class is referred to as information gain. Information gain is a statistical property that can be calculated using entropy, which measures data errors and randomness. A measure of entropy decrease is nothing more than the greatest information gain. The attribute with the greatest information gain is chosen as the split-node decision criterion. By controlling the depth of the tree, the tree is ensured that it does not face overfitting. It is said that DT becomes more complicated and tries to outperform when allowed to grow fully by splitting all the nodes. This increases the output's bias.

D. Using Ensemble Methodology To Advance The Algorithm

Ensemble is the concept of combining many models that are solving the same problem and will eventually be merged together to produce the best result.

- 1) *Bagging*: This method takes into account all of the models that are solving the same problem and learns from them in parallel before combining them on some deterministic averaging process, resulting in higher efficiency.
- 2) *Boosting*: This method takes into account all of the models that are solving the same problem and sequentially learns from each other before merging them on some deterministic strategy that results in higher efficiency [10].
 - a) The Random Forest Classifier is similar to the Decision Tree Classifier, but the cleverest idea here is the use of the ensemble's bagging method. A large number of trees working as a team, predicting classes at random and unrelated to one another, is thought to outperform an individual constituent tree built on the data. In DT, the root node is specifically chosen to split the data, whereas in RF, the nodes are chosen at random and only a subset of the features are considered while avoiding all features [14].
 - b) XGBoost Classifier is another DT classification method that uses gradient boosting to improve prediction efficiency. This is one of the best algorithms because it combines both software and hardware efficiency to reduce computational speed and increase model efficiency. XGBoost splits the data tree using max depth as a specific parameter and criterion, and then begins pruning the trees backwards. This has a high efficacy due to the use of cross-validation, which avoids explicitly mentioning. If necessary, it also employs L1, L2 regression to optimise the loss function.[15].

IV. RESULT AND DISSCUSSION

Model	Accuracy	Precision	Recall	F1 Score	ROC
Logistic	0.6	0.285714	0.857143	0.428571	0.7012
SVM	0.727	0.25	0.2857	0.268667	0.55194
Decision time	0.775	0.37500	0.428571	0.4000	0.638528
Random Tree Classifier	0.85	0.666667	0.285714	0.4	0.627706
XG Boost	0.8	0.4444	0.571429	0.5	0.709957

Table I: Represents the results of all algorithms

The ROC curve of all algorithms used as a metric for binary classification problems is shown below. At various threshold values, the curve plots the true positive rate versus the false positive rate. This also distinguishes the signal from the noise. The ROC curves for XGBoost and Random tree classifiers were plotted because they performed better.

Fig 3. ROC Curve for the top two accuracy algorithms

Fig 4. Bar-graph representing the performance

V. CONCLUSION AND FUTURE SCOPE

Agricultural data is completely haphazard, and the increased use of unhealthy soil is causing crop depreciation and yield loss. Using machine learning, an attempt is made to classify the data into healthy and unhealthy categories. algorithms. The above results show that the accuracy of the prediction model increased as the algorithms advanced. It is obvious that using the ensembles method produces more accurate results than haphazard. misclassified and ambiguous data To improve the accuracy of this type of data, we can use new enhancements to ensemble algorithms such as LightGBM.

REFERENCES

- [1] Soil Analysis and Prediction of Suitable Crop for Agriculture using Machine Learning S. 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 Sukkariha, Brett Whelan. 0168-1699/ Published by Elsevier B.V.
- [3] Soil Knowledge-based Systems Using Ontology, Tongpool Heepaisong and Anongnart Shivihok. 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 Junsong Zhanga , 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, MarieAngélique Laporte, 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] Daiyi Lia, Li Kanga, Xinrong Chenga, Daoliang Lia, Laiqing Jia, Kaiyi Wangb, Yingyi Chena, 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. Vrushi C. Waikar, Sheetal Y. Thorat, Ashlesha A. Ghute, Priya P. Rajput4, Mahesh S. Shinde Student, 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.

10.22214/IJRASET

45.98

IMPACT FACTOR:
7.129

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)