Prediction Of Soil Quality Using Machine Learning Techniques

T. Venkat Narayana Rao, Gaddam Rishitha Reddy

Abstract: Agriculture is a non technical sector where in technology can be incorporated for the betterment. Agricultural technology needs to be quick in implementation and easy in adoption. Farmers usually follow a method called crop mutation after every consequent crop yield. The crop mutation allows the soil to regain the minerals that were used by the crop previously and use the left over minerals for cultivating the new crop. To know if the soil has reached the point where it is unfit to yield the particular crop, farmer has to experience a loss in yield. One financial year for a farmer is very crucial to accept the loss. This paper implements a that would help in maintaining the soil fertility consistently. This method is traditionally implemented in many countries where the change in crop is done after a loss in yield for cultivating the same crop continuously. There are three soil parameters that come into consideration when we have to predict the soil quality. This method suggests the solution for the above stated problem using Machine Learning Techniques. This paper suggests a software enabled solution considering crucial soil parameters and soil factors to predict the soil quality.

Keywords: Machine learning, soil quality, mutation, Artificial Intelligence

1. INTRODUCTION

Agriculture is a non technical sector where in technology can be incorporated for the betterment. Agricultural technology needs to be quick in implementation and easy in adoption. Farmers usually follow a method called crop mutation after every consequent crop yield. This method is traditionally implemented in many countries where the change in crop is done after a loss in yield for cultivating the same crop continuously. The crop mutation allows the soil to regain the minerals that were used by the crop previously and use the left over minerals for cultivating the new crop. This process will help in maintaining the soil fertility consistently. To know if the soil has reached the point where it is unfit to yield the particular crop, farmer has to experience a loss in yield. One financial year for a farmer is very crucial to accept the loss. This method suggests the solution for the above stated problem using Machine Learning Techniques. There are three soil parameters that come into consideration when we have to predict the soil quality as shown in table 1.0.

- Chemical Parameters
- Physical Parameters
- Biological Parameters

Table 1.0 Soil Factors

Chemical Parameters	Soil Texture	It defines the soil particle size using the weight proportion by weight percentage of different soil textures available.
	Water retention character	This property of the soil is used determine the soil water holding capacity after irrigation.
Physical Parameters	Extractable N, K, P	Indicators of plant nutrients, productivity and quality of the surroundings.
	pH	Defining thresholds for biological and chemical activity; vital for modelling processes.
Biological Parameters	Micro-organisms biomass of C and N	It is the bacteria and pillows that break down crop residues and organic matter in the soil that would assist release nutrients such as nitrogen (N) into the soil for plant growth.
	Natural manure availability	Biological manure like animal excretions and the vegetable wastes that saturates the organic matter in soil for plant growth.

1.1 The Existing System and Scope for Development

The crop mutation allows the soil to regain the minerals that were used by the crop previously and use the left over minerals for cultivating the new crop. This process will help in maintaining the soil fertility consistently. To know if the soil has reached the point where it is unfit to yield the particular crop, farmer has to experience a loss in yield. One financial year for a farmer is very crucial to accept the loss.

The following are the drawbacks of the existing manual System:

- Scope for redundancy
- Time Delay
- Less accuracy
- Needs more human effort
- Requires more laboratory

Machine Learning, the basic concepts could make the complex job flexible and automated using a little bit of programming. Machine Learning Algorithms in further

- *T. Venkat Narayana Rao, Professor, Department of C.S.E, Sreenidhi Institute of Science and Technology Yamnampet, Hyderabad, TS, India.*
- *Gaddam Rishitha Reddy Student, Department of C.S.E, Sreenidhi Institute of Science and Technology Yamnampet, Hyderabad, TS, India.*

implementations would make the hell bound jobs easier and automated. The prediction can be extended to any sort of purpose if studied like in prediction of diabetes, prediction of gold purity etc. This project helpful to predict the quality of soil by using some machine learning algorithms like Decision Tree Algorithm and Random Forest Algorithm. The classification of soil according to the quality also can be made easy by using this project without the interference of any human. It takes less time to predict the quality of soil than the humans as the machines work faster and more efficient than the humans. We can also note the quality and the factors of the soil very easily and efficiently and within less time. This project can be used among public to get them know about their land and climate for crop growth.

2.PROPOSED SYSTEM WITH REQUIREMENT SPECIFICATION

The proposed system using machine learning overcomes the drawbacks of the existing system. Machine learning is a computer science subset of Artificial Intelligence(AI) that commonly uses statistical techniques to provide pcs with the ability to "learn" (i.e. gradually increase effectiveness on a specific task) with data without explicit programming.Machine Learning is heavily related to (and often overlaps with) computational statistics, which also uses computers to make projections. It has strong mathematical optimization links, offering techniques, theory and application domains for the field. Sometimes machine learning is associated with data mining, where the latter subfield is more focused on exploratory data analysis and is known as unsupervised learning. By using Machine Learning algorithms, we will calculate the quality of the soil. We use the accurate language for writing the code for the implementation. We calculate the accuracy to know how much accurate is our machine working. With this we can reduce the human effort in the industrial sector where they predict the quality of the soil.

MERITS

Using machine learning ideas, the handling of multi-dimensional and multi-variety information in dynamic settings can be performed.

- Fast processing and predictions in real time.
- Large and complicated process environments provide constant quality. Automation of tasks is easily implemented.
- The advertisements that Face book and Google present are an important application of machine learning.
- Feature learning.

This system analysis is strongly linked to the assessment of demands. It is also "specific investigation to help somebody identify a better course of action and make a better decision than he might have done otherwise." This step involves breaking down the system into separate components to assess the situation, analyzing the project's goals, breaking down what needs to be established, and attempting to engage clients to identify particular requirements[1].

Machine Learning:

Artificial Intelligence is providing a machine with human intellectuals like think and process based on human senses to listen, speak, see or reply through some action. Machine Learning is one such application of AI. In Machine Learning, the system is provided the capacity to learn and improve by itself based on training and experience.

There are three types to train a machine:

- Supervised Learning: In this type of learning, machine is trained using the data where in we provide questions and answers for them to analyse and train. When the similar question is raised then the machine would output the stored answer.
  - Example of one such type is Chabot, [1] where the questions and answers are given while training and if the similar queries are asked the given answers would be the output in the form of text or audio.
- Unsupervised Learning: In this type of learning, machines are left to themselves with the input data but the corresponding output is not known. This is like learning without guidance. Here these algorithms will come out with some interesting conclusions
  - Example of one such type is visual recognition and text recognition where the text from the given data is read and the conclusions or the related search takes place.
- ReinforcementLearning:

We have no input or output to learn from this type of machin e

Learning technique. The question we raise is the input and the feedback/credit system (+,-) is the output to learn.

- Example of one such type is Tesla car, where the cars learn driving on their own (automated driving) and the rewards or credits or feedbacks are their training models.

Requirement Specification

The system was recognized with the following modules after thorough evaluation.

a. Dataset Module:

This module consists of the information regarding the datasets which are given as Input to the next modules. The datasets from excel sheet are imported and converted as a raw data for the input in the further procedure.

b. Training Module:

This module is the essential part of the whole system as the whole system is trained

with machine learning algorithms and train the machine to perform the required task on its

own without human intervention.

c. Testing Module:

This module is used to test the project and verify if the training has to be modified to get the appropriate precision in the output. Several graphs and matrix forms are used to make the user understand [2].

3. Performance Requirements and Design Issues

Performance is measured in terms of the application's outp ut. Specification of requirement plays

significant role in the system analysis. A system can only be intended to fit into the required environment when properly defined requirement specifications. It depends mainly on the current system

customers to give the requirement specifications because t hey are the individuals who end up using the system.

This is because it is necessary to know the demandsduring the original phases so that the system can be designed acc ording to these demands. Once designed, changing the system is very hard and, on the other hand, designing a system that does not fulfil the demands of the user is of no use. The requirements for any scheme can be specified as follows:

• The system should be able to interface with the existing system
• The system should be accurate
• The system should be better than the existing system.

The existing system depends entirely on the user to perform all the tasks.
Systems design is the method by which the architecture, elements, modules, interfaces and information of a system are defined to meet specific requirements. It could be viewed as applying the system theory to product development. Object-oriented evaluation and design methods become the most widely used computer system design techniques as shown in figure 1.0 [10].

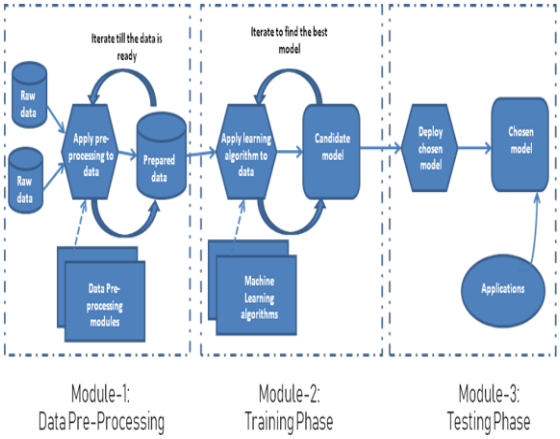


Figure 1.0 *Design of projects using Machine Learning Algorithms*

The above data flow diagrams explains the whole procedure starting from raw data collection to the end application. Raw data is collected and apply data processing modules and we have to pre-process the data and we should check whether there are any null values or not. If there are any null values we should replace them by NaaN. Then we get the prepared data [3] mentioned in figure 1.1 and 1.2.

```
RangeIndex: 99 entries, 0 to 98
Data columns (total 10 columns):
Sample No    99 non-null int64
PH           99 non-null float64
EC           99 non-null float64
N            99 non-null float64
P            99 non-null float64
K            99 non-null int64
Avg Rain     99 non-null float64
Max Temp     99 non-null int64
Min Temp     99 non-null int64
quality      99 non-null int64
dtypes: float64(5), int64(5)
memory usage: 7.8 KB
```

Fig 1.1 *Input Dataset Columns*

	Sample No	PH	EC	N	P	K	Avg Rain	Max Temp	Min Temp
0	1005	6.50	0.16	189.0	55.76	150	1009.5	36	23
1	1006	7.77	0.25	214.0	27.33	174	1009.5	36	23
2	1007	8.08	0.33	91.0	46.25	124	1009.5	36	23
3	1008	7.51	0.51	147.0	37.60	145	1009.5	36	23
4	1009	7.70	0.10	133.0	34.00	123	1009.5	36	23

Figure 1.2 *Input Dataset after Data Pre-Processing*

We have to apply the pre-processed data to learning algorithm i.e. (Decision Tree, Random Forest) after that we will get the candidate model of the algorithm. We have iterated the same process to get the bestmodel of the algorithms. This is called training phase. After that we have to deploy the chosen model and the chosen model is used in the applications. This is the testing phase. The output quality is given as a percentage based on the inputs.

4. SYSTEM IMPLEMENTATION

The execution stage of any project is a true display of the defining moments that create a success or failure of a project. The execution stage is defined as the scheme or system modifications in a manufacturing environment being installed and operated._The stage starts after the user has tested [4] the scheme and approved it. This stage continues until the system operates according to the specified user requirements in manufacturing.

4.1 Algorithm

A) Decision Tree Algorithm: It is a supervised learning algorithm that is mostly used for classification problems. It works for either discrete or ongoing variables. A decision tree is plotted with its root at the top and branches at the bottom. The picture's courageous text reflects an internal condition /

node on the basis of which the tree divides into branches / edges [5].

Example 1:
Consider an instance of using titanic information set to predict whether or not a passenger is going to survive as given in figure 1.3. The following model utilizes three information set features / attributes / columns, i.e. sex, age and sibsp (no spouse / children). In this case, it is represented as red and green text respectively whether the passenger died or survived.

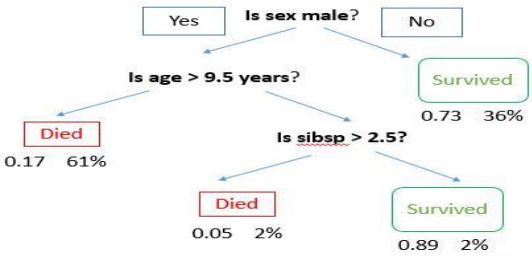
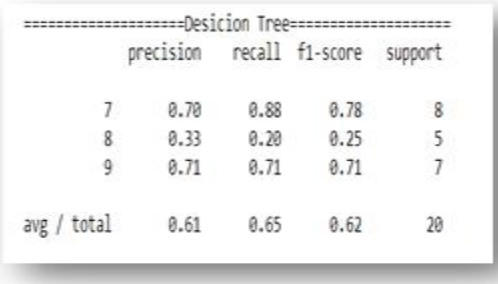


Figure 1.*3 Working of Decision Tree Algorithm*

We see in some examples that the population is categorized into distinct groups based on various characteristics in order to define ' if they do something or not. ' It utilizes multiple methods such as Gini, Information Gain, Chi-square, Entropy, etc. to divide the population into distinct heterogeneous groups. Playing Jezzball – a classic Microsoft game – is the easiest way to know how decision tree algorithm works. Essentially, you have a space with moving walls in this match and you need to build walls so that without the balls the total area is cleared. So, you try to generate 2 distinct populations inside the same room every time you divide the space with a wall. By dividing a population into as many groups as possible, decision trees function very likewise. This is performed on the basis of the most important characteristics to make the organizations as distinct as possible [6]. Decision trees are widely used in machine learning, covering classification as well as regression. A decision tree is used in decision analysis to portray choices and decision making visually and explicitly. It utilizes a tree-like decision-making mode[9].


==========Desicion Tree==========

	precision	recall	f1-score	support
7	0.70	0.88	0.78	8
8	0.33	0.20	0.25	5
9	0.71	0.71	0.71	7
avg / total	0.61	0.65	0.62	20

Figure 1.*3 Working of Decision Tree Algorithm*

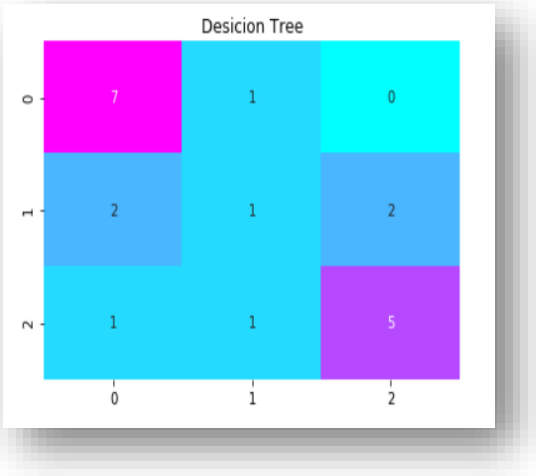


Figure 1.*4 Decision Tree output in confusion matrix and correlation matrix*

B) Random Forest Algorithm: For a supervised ensemble, Random Forest is a popular learning algorithm. Ensemble means that working together to form a strong predictor involves a lot of' weak learners.' In this scenario, the weak learners who are brought together to form the strong predictor— a random forest— are all randomly introduced choice trees.
Like decision trees, trees forests also apply to issues with multiple outputs (if Y is a size range [n samples, n outputs]).
Unlike the original publication [B2001], the scikit-learn implementation mixes classifiers by averaging their probabilistic prediction instead of enabling each classifier to vote for a single class.
Random Forest is a trade-mark term for a collection of decision trees. In Random Forest, we have a collection of decision trees called "Forest."[9] Each tree offers a scheme for the classification of a new item based on features and we say the tree "votes" for that class. The forest chooses the classification of the most votes (over all forest trees).
Each tree is planted & grown as follows:

- If the amount of instances in the training collection is N, the sample of N instances will be drawn randomly but replaced. We use this sample as a practice set for growing the tree.

- If there are M input variables, the number m < M shall be specified to divide the node by randomly selecting the M variables at each node and using the best split at each m. The value of m is held constant when the forest grows[7][8].

- To the maximum extent possible, each tree is cultivated. No pruning is available.

These two algorithms have been used to train our model of machine learning. In case of random forest, we received 70 percent precision and 67 percent precision in case of decision tree as evident from the above results.

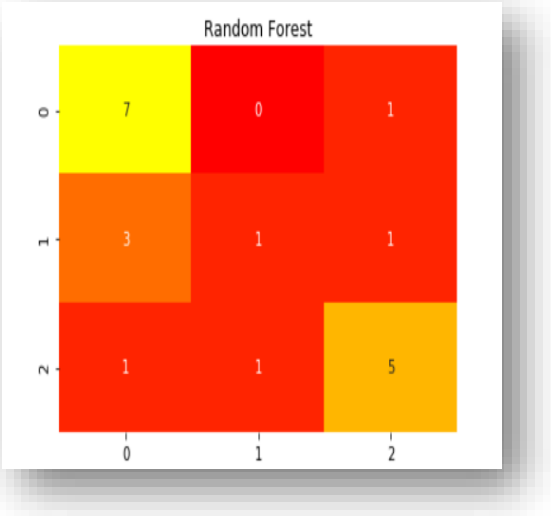


Figure 1.*5 Random Forest output in confusion matrix and correlation matrix*

5. CONCLUSION

Looking back at the motivation of this research, namely the attempt to mine a dataset composed of soils in order to attempt to develop a model for predicting the quality of a soil based on the chemical, physical and biological compositions of that soil and to examine the consistency of the soil testing field, A number of conclusions can be drawn from the outcomes of the different models built. Before considering any outcome that may be suggested, it is necessary to understand the constraints of this evaluation's scope. First, it is essential to note that the soil metrics set consists of a choice of distinct soils. Therefore, any findings drawn apply only to these specific soil varieties. In addition, since soil testing is a qualitative and, to a degree, subjective judgment by a selection of professional testers, the models developed in this research will be biased towards modeling the qualities of the particular testers who generated the data, which may differ from the ratings given by different testers. It is evident that either the information or the algorithms need to be tweaked in order to further improve the precision of our classifiers. We would suggest feature engineering, exploiting the prospective relationship between soil characteristics, or applying algorithms to the most precise techniques already available. It is evident that to further improve the accuracy of our classifiers, either the information or the algorithms need to be tweaked. We would suggest feature engineering, exploiting the potential connection between the features of the soil, or applying algorithms to the most accurate available methods. This means that fixed acidity and pH should be extremely correlated and could possibly be combined to decrease the amount of characteristics, simplifying the issue. Because we understand that Random Forests was the best classifier, with regression tailing right behind it, we estimate that Gradient Boosted Decision Trees has high ability to perform even better owing to the heavy dependence on Random Forests, using regression to increase efficiency. By taking and weighing our weak learners, one could possibly attain better outcomes.

REFERENCES:

[1] P. Vinciya, Dr. A. Valarmathi, “Agriculture Analysis for Next Generation High Tech Farming in Data Mining” IJARCSSE,vol. 6, Issue 5, 2016.
[2] Shivnath Ghosh,Santanu Koley, “Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks” IJRITCC, vol. 2, Issue 2,292-297,2014.
[3] Zhihao Hong,Z. Kalbarczyk,R. K. Iyer, “A Data-Driven Approach to Soil Moisture Collection and Prediction” IEEE Xplore,vol. 2, Issue 2,292-297,2016
[4] Sabri Arik, Tingwen Huang, Weng Kin Lai, Qingshan Liu , “Soil Property Prediction: An Extreme Learning Machine Approach” Springer, vol. 3, Issue 4,666-680,2015.
[5] Vaneesbeer Singh,Abid Sarwar, “Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach” IJARCSE,vol. 5, Issue 8,2017.
[6] Okalebo et al. (2002). Laboratory Methods of Soil and Plant Analysis: A working Manual. Second edition TSBFCIAT and SACRED Africa: Nairobi, Kenya.
[7] Wikipedia. (2009). Soil Testing. Last modified March 1st 2009.Retrieved on March 4th 2009 from http://www.wikipedia.com/soil_testing.
[8] Echochem Online. (2009). Soil Health and Crop yields. Last modified January 28th 2009.Retrieved on March 4th 2009 from http://ecochem.com/healthy_soil.html
[9] Food and Agricultural Organization. (2006). The state of Agricultural Commodity Markets. 37-39.
[10] http://agricoop.nic.in/sites/default/files/Annual_rpt_2 016 17_E.pdf