Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning

Keerthan Kumar T G, Shubha C, Sushma S A

Abstract: ***In society the population is increasing at a high rate, people are not aware of the advancement of technologies. Machine learning can be used to increase the crop yield and quality of crops in the agriculture sector. In this project we propose a machine learning based solution for the analysis of the important soil properties and based on that we are dealing with the Grading of the Soil and Prediction of Crops suitable to the land. The various soil nutrient EC (Electrical Conductivity), pH (Power of Hydrogen), OC (Organic Carbon), etc. are the feature variables, whereas the grade of the particular soil based on its nutrient content is the target variable. Dataset is preprocessed and regression algorithm is applied and RMSE (Root Mean Square Error) is calculated for predicting rank of soil and we applied various Classification Algorithm for crop recommendation and found that Random Forest has the highest accuracy score.***

Keywords: ***Crop Recommendation, Fertility Grading, Machine Learning, Prediction, Random Forest, Linear Regression***

I. INTRODUCTION

India is a country which has huge no. of natural as well as human resources, and it's economy is growing at an rapid rate. A large part of Indian economy is dependent upon agriculture sector and to improve agricultural practices it is necessary to accurately predict responses of the crop yield which can be done with the help of Machine Learning. Agricultural soil quality depends on the soil macro as well as micro nutrient content like S, K, pH, C, Mg, P, Ca, B etc. [2]. Our main objective is the examination, adaptation, and formulation of soil properties and crops growth factors.

The main aim of this project is the examination of macro and micro soil properties such as organic content, essential plant nutrients, that affects the crop yield and find out the rank of a given soil based on the previously graded soil using Supervised Learning [1],[9]. Hai-Yang Jia et al. [1],[10] and E. Manjula et al. [2] investigated that based on soil type and soil nutrient content a suitable Regression Algorithm can be applied and using suitable Classification Algorithm the best suitable crop for the land is recommended.

To start with the project, we have first done the preprocessing of the data [8]. Some records have missing attribute values, that records were removed from the dataset. In the data conversion step, the preprocessed data was converted based on the nutrient's values. After data conversion, the macro and micro nutrients are analyzed.

Revised Manuscript Received on November 05, 2019.

Keerthan Kumar T G, Assistant Professor, Siddaganga Institute of Technology, Department of information science and Engineering, Tumakuru-572103, Karnataka State, India.

Shubha C, Assistant Professor, Siddaganga Institute of Technology, Department of information science and Engineering, Tumakuru-572103, Karnataka State, India.

Sushma S A, Assistant Professor, Siddaganga Institute of Technology, Department of information science and Engineering, Tumakuru-572103, Karnataka State, India.

II. BACKGROUND

This project is an idea to grade soil at a particular place and to recommend the best crop suitable for a land based on the previously fed data regarding the same. This project consists of two modules:

Module-1: Quantized rank [1] of the soil based on macro as well as micro nutrients is the main criteria to be recognized by models.

Module-2: Based on soil type and soil nutrient content a suitable Classification Algorithm will be applied for crop recommendation.

A. System Design

Module-1: Grading of Soil

This module consists of Machine learning model that is built to help farmer understand the quality of his soil by considering various Soil nutrients (both macro as well as micro nutrients) and based on these parameters the following model is made.

- Content of various soil nutrient (EC, K, pH, Mn, Zn, S, P, B, OC) are the feature variables, while the grade of soil nutrient criterion is the target variable.
- Quantized rank of the soil macro as well as micro nutrients is the main objective of the model.
- Preprocessing of the dataset is done.
- Regression algorithm (linear regression) is applied [10].
- Cost function is minimized by the algorithm, gradient descent is applied and appropriate learning rate is chosen.
- Root Mean Squared error between the predicted value and true value is calculated.

Module-2: Crop Recommendation

The Module is built keeping in mind the minimal soil nutrients requirements and Soil type that are necessary for growing a particular crop and helping farmers to maximize their yield [5] and good utilization of the agricultural field.

Data Preprocessing:

Soil attributes like pH, EC, OC, S, K, Zn, Mn, B and Soil Type are taken as feature Variables and the NULL values as well as redundant values from the dataset are removed. Data collection is done by the ICRISAT Development Center and Government of Andhra Pradesh [8]. This helps to transform the Agriculture sector in all 13 districts of Andhra Pradesh, by enhancing crop productivity and improving lifestyle of farmers. This project helps to restore the soils of Andhra Pradesh using soil health plotting as an entry point in order to unleash the potential of agriculture sector which relies on rainfed. Data is crucial for any project on machine learning.

Retrieval Number: L36091081219/2019©BEIESP
DOI: 10.35940/ijitee.L3609.119119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Researching the Model best fit for the Data:

Application of various classification algorithms like Support Vector Machine, Random Forest Classification and Decision Tree is done and based on the RMSE (Root Mean Squared Error) the suitable model is selected. After having collected all the information we need to select the relevant column of data required and discard the rest. This is what was done in this step. This was done directly in spreadsheet. MLaaS is another tool which can prove beneficial [12]. This is generally used in places with really huge datasets.

Training and Evaluating the Model on Data:

After verifying the suitable model for our problem, the model will be subjected to evaluation where the correctness of the model is checked by passing a real time data to validate the model. A visual representation of how the data flows from one step to another is shown in Fig.1.

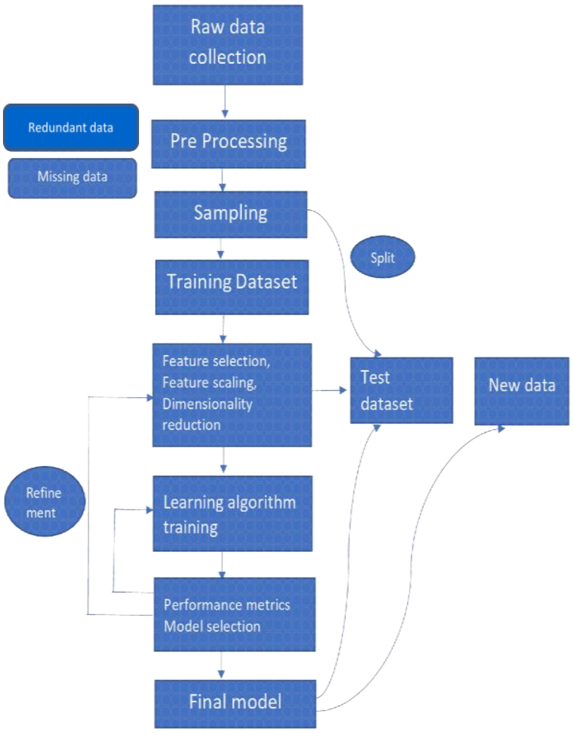

Fig.1 Flow Chart of the System

B. Algorithms

MVLR Cost Function:

provide the best fit line (minimize Mean squared error).

$$\text{Minimize } \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \tag{1}$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \tag{2}$$

Gradient Descent:

An algorithm that reduce the cost function (MSE) and gets best fit line. For different m and b, bowl shape plot [1] is obtained. In this algorithm we start with assigning some random values to m and b and then changing them recursively until a convergence is obtained to reduce the cost. To find the gradients, we took partial derivative with respect to m, b. Update m and b simultaneously.

Alpha is the learning rate and must be explicitly given. With a larger learning rate, the convergence is achieved quickly but there is a chance that you could miss the minima whereas with a smaller learning rate the time taken is more to reach the minima but it will take you closer to minima as shown in Fig.2.

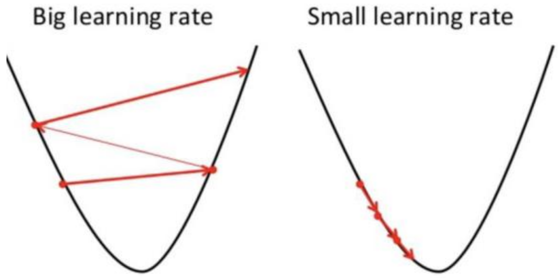

Fig.2. Gradient Descent Curves

$$J = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{pred_i} - \mathrm{y_i})^2 \tag{3}$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{a_o} + \mathrm{a_1}.\mathrm{x_i} - \mathrm{y_i})^2 \tag{4}$$

$$\frac{\partial J}{\partial a_o} = \frac{2}{n}\sum_{i=1}^{n}(\mathrm{a_o} + \mathrm{a_1}.\mathrm{x_i} - \mathrm{y_i})$$

$$\rightarrow \frac{\partial J}{\partial a_o} = \frac{2}{n}\sum_{i=1}^{n}(\mathrm{pred_i} - \mathrm{y_i}) \tag{5}$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(\mathrm{a_o} + \mathrm{a_1}.\mathrm{x_i} - \mathrm{y_i}).x_i$$

$$\rightarrow \frac{\partial J}{\partial a_o} = \frac{2}{n}\sum_{i=1}^{n}(\mathrm{pred_i} - \mathrm{y_i}).x_i \tag{6}$$

$$a_o = a_o - \alpha.\frac{2}{n}\sum_{i=1}^{n}(\mathrm{pred_i} - \mathrm{y_i}) \tag{7}$$

$$a_1 = a_1 - \alpha.\frac{2}{n}\sum_{i=1}^{n}(\mathrm{pred_i} - \mathrm{y_i}).x_1 \tag{8}$$

Module-1

- Regression: Statistical technique to make prediction when target quantity is continuous (using Linear Regression here) [13].
- One independent variable: Simple Linear Regression
- Multiple independent variables: Multi Variate Linear Regression
- We are using Multi-Variate Linear Regression Algorithm so as to predict the Soil fertility on a scale of 5.

Step-1

Selection of Hypothesis and Cost function: A hypothesis h(x) is a predicted value of the response variable. Cost function defines the cost associated with the wrong prediction of hypothesis. It should be minimum. We have chosen hypothesis function as linear combination of features X.

$$h(x^i) = \theta_0 + \theta_1 . x_1^i + \cdots \ldots \ldots + \theta_n x_n^i \qquad (9)$$

In equatuion (9,10), where

$\theta = [\theta_0 + \theta_1 + \theta_2 + \cdots \ldots . + \theta_n]^T$ is the parameter vector, and x_i^j= value of i^th feature in j^th training example, and the cost function as sum of squared error over all training examples.

$$J(\theta) = \frac{1}{2m*\sum(h_\theta(x^i)-y^i)^2} \qquad (10)$$

Step-2

Minimization of the Cost function: Now the Gradient Descent Algorithm is executed over the datasets which adjusts the parameters of the hypothesis and minimizes the cost function. Once the cost function is minimized for the training dataset, it should also be minimized for a random dataset to check if the relation is universal. Gradient descent algorithm proves to be best for minimizing the cost function in case of multivariate regression.

Step-3

Hypothesis Testing: The hypothesis function testing is done over the test set to check for its correctness and efficiency.

Module-2

Random Forest Classifier:

Random Forest is a supervised learning algorithm. As in Fig.3 random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. By using this algorithm, we can add randomness to our model. Random forest looks for the most important parameter among all while doing splitting of any node, then from the subset of random features it searches for the best among them. This eventually generates a model which has higher accuracy in wide diversity [4],[11].

In this algorithm only selective features are taken into account for the splitting of a node [14],[16]. The trees can be made more random, by using random thresholds for the feature set rather than searching for the best thresholds possible. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set where, X = x_1, … x_n with responses Y = y_1, ..., y_n, continuously bagging b times by selecting a random sample with replacement of the training set and fitting trees to these samples as shown in Fig.3:

For, b = 1,…..,B:

1. Sampling, with replacement these n training sets from Y_b, X_b.
2. And training a regression tree f_b on Y_b and X_b.

after this process is complete, unknown samples x' predictions are applied by taking average of these predictions from all individual regression trees on x':

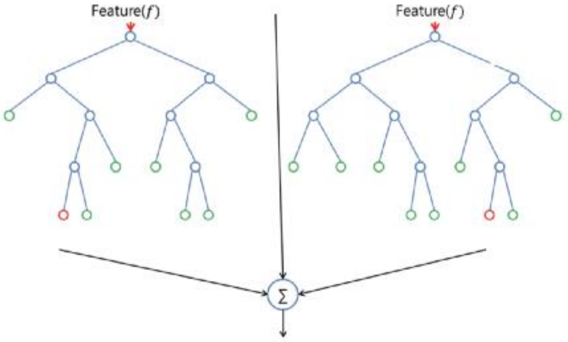

Fig.3 Random Forest Classifier

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad (11)$$

To decrease the variance of our model, without increasing the bias we have applied a bootstrapping procedure that eventually leads to a better model performance. We have seen that if the trees do not have any relation the average of these trees are not so sensitive towards noise but on the other hand predictions made for a single tree are highly sensitive to noise in the training set [11]. By training many trees on a single dataset we can generate strongly correlated trees, to de-correlate these trees we can use different training sets on them which is known as bootstrap sampling. In addition to this an estimate of the uncertainty of the prediction can be made by evaluating the standard deviation of the predictions from all the individual regression samples on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x')-\hat{f})^2}{B-1}} \qquad (12)$$

In equatuion (12), B is the no. of trees and is a free parameter. Generally, number of trees ranges from a few hundred to several thousand trees which mainly depends upon the size and nature of the training set.

III. RESULT ANALYSIS

While grading the soil based on its properties, Linear regression proves to be an efficient algorithm [3] with very less root mean squared error which is a metric for measuring accuracy in case of regression problem. In case of crop prediction, Random Forest proves to be a better classifier as compared to Gaussian Naïve Bayes [7] and Support Vector Machine [6],[14],[15]. This model helps at predicting soil fertility class using various algorithms like decision tree algorithm where large amount of data is present.

Module-1:

Algorithm: Linear Regression

Formula Used-

print(np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

Table I: Error Analysis of Module-1

Error	Values
Root Mean Squared Error	6.1683%
Intercept of the model	0.031866
List of coefficients	0.1148,0.1246,0.1346,0.1262, 0.1314,0.1433,0.1195,0.09474

In Table I, the error analysis of the model using Linear Regression is done and the following Root Mean Square Error of 6.1683% is obtained and latter are the values of the intercepts and coefficients in Linear Regression Model.

Module-2:

Formula Used- print (accuracy score (y_test, predicted))

Table II: Comparison of Accuracy Score of Various Classification Algorithms

Algorithm	Corresponding Accuracy score
Random Forest Classifier	72.74%
Support Vector Machine (Linear Kernel)	63.33%
Gaussian NB	50.78%

Table II shows accuracy Score of various classification Algorithms are compared and the best algorithm based on the accuracy score is chosen for module-2.

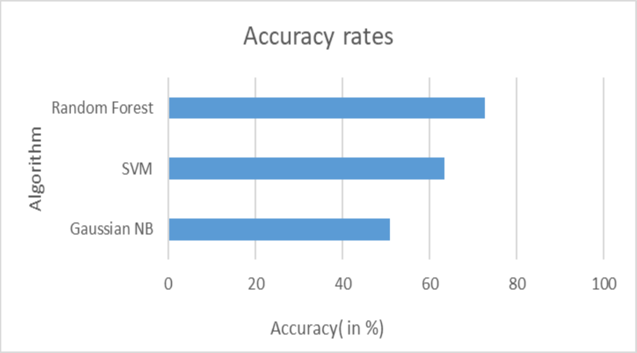

Fig.4 Bar Graph for Result Analysis

Fig.4 shows Graphical Representation of the Accuracy Score of various Classification Algorithms.

IV. CONCLUSION

The system uses supervised Machine learning algorithms like Linear Regression Multi-Variate, Support Vector Machine, Random Forest Classifier and gives best result based on error analysis. The results of these algorithms will be compared and the best among them i.e., Random Forest Classifier which gives the best and accurate output is chosen. Therefore, this system will help reduce the struggle faced by the farmers. Analysis of the important soil properties and based on that we are dealing with the Grading of the Soil and Prediction of Crops suitable to the land. This will act as simple solution to equip the farmers with necessary information required to obtain great yield and therefore maximizing their surplus and therefore will reduce his difficulties.

REFERENCES

1. Hai-Yang Jia, Juan Chen,"Soil fertility grading with Bayesian network transfer learning, presented at the Proceedings of the Ninth International Conference on Machine Learning and Cybernatics, Qingdao, 2010.
2. E. Manjula, S. Djodiltachoumy, “A Model for Prediction of Crop Yield” (International Journal of Computational Intelligence and Informatics, March 2017).
3. A. Kumar & N. Kannathasan, (2011), “A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining “, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3.
4. Gholap, Jay. “Performance Tuning of J48 Algorithm for Prediction of Soil Fertility.” ArXiv abs/1208.3943 (2012): n. pag.
5. “Prediction of Crop Yield using Machine Learning” Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R, IJARCSE,vol. 5, Issue 8,2017.
6. Shivnath Ghosh, Santanu Koley (2014) “Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks” International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 2, ISSN: 2321-8169, pp 292 – 297.
7. Gorthi, Swathi and Huifang Dou. “PREDICTION MODELS FOR ESTIMATION OF SOIL MOISTURE CONTENT.” (2011).
8. http://dataverse.icrisat.org/dataset.xhtml?persistentId=doi:10.21421/D2/K3BPKW
9. Oyen, D. & Lane, T. Knowl Inf Syst (2015) 43: 1. https://doi.org/10.1007/s10115-014-0775-6
10. Shah A., Dubey A., Hemnani V., Gala D., Kalbande D.R. (2018) Smart Farming System: Crop Yield Prediction Using Regression Techniques. In: Vasudevan H., Deshmukh A., Ray K. (eds) Proceedings of International Conference on Wireless Communication. Lecture Notes on Data Engineering and Communications Technologies, vol 19. Springer, Singapore
11. 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
12. Vaneesbeer Singh,Abid Sarwar, “Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach” IJARCSE,vol. 5, Issue 8,2017.
13. Miss.Snehal, S. D. (2014). Agricultural Crop Yield Prediction Using Artificial. International Journal of Innovative Research in Electrical,Electronic, 1 (1).
14. Kajol R, Akshay Kashyap K, Keerthan Kumar T G," Automated Agricultural Field Analysis and Monitoring System Using IOT", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.10, No.2, pp. 17-24, 2018. DOI: 10.5815/ijieeb.2018.02.03
15. Shriya Sahu, Meenu Chawla, Nilay Khare, "An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach", Computing Communication and Automation (ICCCA) 2017 International Conference on, pp. 53-57, 2017.
16. Rashmi Priya, Dharavath Ramesh, Ekaansh Khosla, "Crop Prediction on the Region Belts of India: A Naïve Bayes Map Reduce Precision Agricultural Model", Advances in Computing Communications and Informatics (ICACCI) 2018 International Conference on, pp. 99-104, 2018.

AUTHORS PROFILE

Keerthan Kumar T G received his Master Degree in computer science and engineering from VTU University, Karnataka, India, in 2012. he was worked in VMware software India private Ltd. and Dell R and D, Bangalore for one and two years respectively in various domains like script writing, virtualization, cloud computing, Management software and automation. He is currently working as Assistant Professor in the Department of Information Science and Engineering at Siddaganga Institute of Technology, Karnataka, India. His research interests include software testing analysis and design, cloud computing, Internet of things, big data, mobile cloud, cloud security, formal verification system, SDN.

Shubha C received the M.Tech degree from SJB college of Engineering and Technology, Bangalore in 2012. She was working as Android developer at BORQS software solutions Pvt Ltd. She is presently working as Assistant Professor in Siddaganga Institute of Technology, Tumkur. She has 6 years of teaching experience and 1 years of industry experience. Her current research interests include Android, affective computing, data analysis and artificial intelligence.

Sushma S.A received the M.Tech degree from B.V. Bhomraddi college of Engineering and Technology, Hubli in 2014. She was working as GTO Adobe Consultant at Cognizant Technology Solutions. She is presently working as Assistant Professor in Siddaganga Institute of Technology, Tumkur. She has 4 years of teaching experience and two years of industry experience. Her current research interests include big data, machine/deep learning and artificial intelligence.