British Journal of Multidisciplinary and Advanced Studies:
Earth Sciences 4(6),1-11, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by the European Centre for Research Training and Development UK

Soil Analysis Using Machine Learning

Prakash Kanade
Researcher in Robotics, Artificial Intelligence, IoT, USA
Email: Prakashsrs@gmail.com

doi: https://doi.org/10.37745/bjmas.2022.0350 Published November 25, 202

Citation: Kanade P. (2023) Soil Analysis Using Machine Learning, *British Journal of Multidisciplinary and Advanced Studies*: Earth Sciences 4(6),1-17

ABSTRACT*: India's agriculture sector employs the most people. Here it is: Agriculture employs around 60% of the Indian population and accounts for about 18% of India's GDP; yet, low productivity is due to a lack of research in this industry. Water logging, soil erosion, nitrogen shortage, and other issues plague Indian agricultural land. These are the primary causes of agriculture's low productivity. A farmer must spend a significant amount of time and money on farming, which is more extensive and time intensive than using a tractor. As a result, the cost of agriculture has increased. It is critical to use machine learning techniques and computational research to the agriculture industry in order for India to become a better quantity and quality food producer. ML approaches are particularly beneficial for building relationships and abstracting patterns between disparate data sets, as well as forecasting a realistic outcome as an output. It can be successfully implemented in the Indian agriculture sector to increase efficiency. We've talked about how machine learning techniques can be used in the Indian agriculture industry to assess soil fertility. Agriculture has long been one of the most fascinating study and analytical topics. This research aims to assess soil data based on a variety of characteristics, classify it, and increase the efficiency of each model using multiple terms and classifications. The major goal of this study and analysis is to classify soil fertility (behavior) indices by area using village-level soil fertility data.*

KEYWORDS*:* Artificial Neural Network (ANN), LeenaBOT, Support Vector Machine (SVM), Decision Tree, k-nearest neighbors (KNN), Soil.

INTRODUCTION

Agriculture, as we all know, is a non-technical industry where technology can help with growth and improvement. Agricultural technology needs to be implemented quickly and in a growing number of places. This industry has employed almost 60% of the population of our country. On the other hand, it accounts for only 18 percent of India's annual GDP. The disparity arises from a lack of research and innovative technology. In comparison to other countries, our agriculture sector is less automated. Except for Asian countries, all countries use a variety of approaches to make agricultural projections. In this technical area, our agriculture sector lags

behind. Due to a lack of technological understanding, growth and production are significantly reduced. The fertility of Indian soil is a limiting element in the country's agricultural business. When other environmental elements such as light, temperature, and water are suitable, soil fertility determines plant development. Soil fertility is influenced by a variety of factors such as soil nutrition, climate, irrigation (soil water), and soil alkalinity, and urbanization, globalization, changing soil acidity, meteorological conditions, and increased pesticide use in India. A lack of certain soil types results in lower agricultural production and, as a result, higher food costs. The fertility of the soil is assessed using various soil types. The major purpose of implementing technology in this sector is to have the least amount of impact on food and soil fertility and quality. Soil organic carbon is an important determinant of plant productivity, soil fertility, and soil quality. It also plays an important function in delivering nutrients to the soil for better soil fertilization. Diverse soil researchers have reported on the variability of SOC in various ecological settings around the world in recent years.

These investigations support various hypotheses, including the fact that crop production variance within a specific field reflects SOC fluctuation. These studies also stated that in order to accomplish significant soil nutrients management for utilized crop output, it is necessary to understand where soil nutrients, as well as low SOC, dwell inside a given field, as well as the amount of carbon or soil nutrient available. It is fundamentally the significance of soil quantity mapping. The application of site specific factor management for structural match able variable situations using the most up to date, precise, and correct information gathered in this classification procedure. Because of the non-uniformity of o/p over different sections of the same field, various nutrients are a key constraint for imperishable (effective) Indian agricultural production. Advanced soil mapping is one technique to lessen soil discordance that results in varied crop yields, but it is unpleasant due to differences within the site. That challenge became the focus of Agricultural India's specific agricultural system.

All crop yield approaches can find venues for specific management. On the basis of geo-information technology and utilizing soil characteristics (pH, phosphorous, etc.), micro climatic data, (DEM), remote sensing data, and geology, Indian agriculture has progressed to numerous types of nutrients and varied types of soil properties within a particular field soil. Individuals can manage within-field variability with site-specific agriculture to optimize the suggested crop structure's cost cut ratio.

RELATED WORK

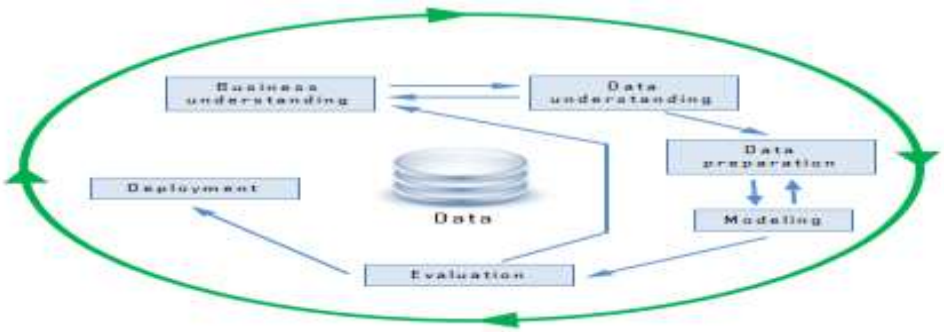
Agriculture is one of the most researched topics in both scientific and academic circles. Using various statistical tools, data mining, and classifications, past research has been reported on the agriculture industry and soil fertility. The fertility of the soil is influenced by a number of

things. Zinc, sulphur, and water are the most important components in determining the soil's richness. A case study of structure was conducted on 3526 soil samples from various Indian states. Moisture or water level in the soil is the most important component, according to the study. Soil fertility is determined by the amount of moisture in the soil. It divided features and signals using the dissolving algorithm. It will be classified and separated into classes using the boundary technique; they will employ Decision Tree, ANN, and SVM to classify surface soil data. To forecast water retention and compressor conductivity, hierarchical neural network models were created. With the help of bulk density and texture, regression and ANN were built to verify water retention. It used SVM to forecast soil moisture from data collected remotely. It uses the linear regression technique, as well as "Nave Bayes" and "J48-classification," to anticipate soil data. Cluster analysis is used on soil data obtained by the food department (Australia). On soil texture, multiple classification techniques were used, and it was discovered that Baysian classification is more accurate and performs well. For high-quality soil maps, artificial neural networks and digital terrain analysis are used.

It is installed in the fields as a decision tree to assist farmers in selecting an irrigation pump based on irrigation kinds, motor capacity, total area coverage of the field, and height. Several machine learning algorithms were used to classify soil fertility, including Nave Bayes, J48, and the random forest algorithm. J48 outperforms other algorithms in terms of accuracy. Rub used various regression algorithms on soil data and found that SVM produced a better model for prediction. Because of the ensuing non-uniformity of o/p over different areas of the field, differences in soil nutrients are a substantial barrier for livable crop output. Digital soil mapping (DSM) is one technique to reduce soil diversity that leads to varying crop yields, although it is typically hampered by within-site variability. Rub used a variety of regression algorithms on soil data and determined that the Support Vector Machine produced a better model. Food used in a single Support Vector Machine versus a succession of classifiers for crop categorization can produce optimal results, resulting in improved crop performance.

METHODOLOGY

Various techniques, such as, are used in Data Mining Technology.



British Journal of Multidisciplinary and Advanced Studies:
Earth Sciences 4(6),1-11, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by the European Centre for Research Training and Development UK

Figure 1. Data Mining Process

1. **Understanding the Business:** It is the first step in the data mining process, and it collects all of the information linked to agriculture and soil industry before moving on to the next step.

2. **Understanding the soil data:** There were 60 attributes in the data set that was chosen. The dataset's final output has 1300 records and 10 distinct attributes.

3. **Data Preparation:** Using simple excel features and the R programming language, remove extraneous columns, additional blank spaces, and null values from the dataset. Some packages, such as tidy and reader, were utilized. In the data cleansing process, the next step has been completed.

Figure 2. Data Preparation Process

4. **Null Values handling:** The collection had a small number of null records. To eliminate irregularities in the data set, null records were replaced with NA values. Removing unnecessary columns: The data set contained several columns that were unrelated to our research; these columns were eliminated, and only relevant data was used as input for further processing, such as numeric values and inconsistent data types in the data set. The 'Porosity' column's scale and accuracy were inconsistent, therefore the column's unique accuracy was changed.

5. **Handling Unnecessary Spaces:** To remove extraneous spaces in column values, I used Excel's 'TRIM' function.

6. For data modelling and evaluation in selecting the approach, test design, generate, model assessment, and developing a model evaluation of the result, use the evaluation sequence and data modelling.

Decision Tree

A branch, root node, and leaf node are all part of the structure. Each internal node represents an attribute test, each leaf node represents a class label, and each branch represents the test result. A test on an attribute is represented by each internal node. A decision tree is employed as a classification model in our research. There are three labels in the class in the dataset based on the soil fertility: low, medium, and high. Depth, carbon, ph, conductivity, nitrogen, phosphorus, potassium, WHC, and porosity are some of the dependent parameters evaluated. After the data has been loaded, the data frame is rearranged. The rearranged data frame is then separated into test and train, with test and train maintaining a 70:30 percent ratio. The decision tree is applied to the previously specified dependent factors. To build classes, the R part function is utilized, and a decision tree is predicted based on the inputs. On the test data set, a prediction model is created using a decision tree as i/p. The following is a list of factors, sorted in order of importance of function in classification:

1) Conductivity
2) WHC
3) Potassium
4) Nitrogen
5) pH
6) Phosphorous

Artificial Neural Network (ANN)

It's a component of a system created to mimic how the human brain evaluates and processes data. It is the cornerstone of AI, and it addresses problems that are difficult or impossible to solve using human or statistical standards. Artificial Neural Networks have self-learning characteristics, allowing them to produce reasonable outcomes when more data is available. We chose ANN because it produces superior categorization results. The class has been converted to numeric for this purpose because artificial neural networks do not accept strings as input/output. Using the minimum-maximum formula. By using the neural net package for different nodes, all of the data was trained using neural net function (hidden). The initial step in using ANN was to load all of the soil data into an R data frame. In this case, the classes were string low, medium, and high fertile, which were modified to 1, 2, and 3 accordingly. Data has been adjusted as a result of the varied Ranges of each column. The min-max formula is used.

British Journal of Multidisciplinary and Advanced Studies:
Earth Sciences 4(6),1-11, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by the European Centre for Research Training and Development UK

The data was separated into two groups: test and training, with a 20:80 ratio.

Attributes	Description	Attribute Type
Type	It Describes fertility of soil as less, medium and high	Text (Numeric for SVM)
Ph	Describes pH of soil	Numeric
Nitrogen	Tells about available Nitrogen in Soil	Numeric
Phosphorous	Available Phosphorous in soil	Numeric
Porosity	Porosity in percentage	Numeric
Depth	Depth of soil in centimeter	Numeric
Conductivity	Conductivity of soil	Numeric
Organic Carbon	Organic Carbon present in the soil in percentages	Numeric
Potassium	Potassium availability in soil	Numeric
Water holding capacity	Capacity of water holding in percent	Numeric

ANN is the properties of soil are highly correlated to each other. ANN because it produces good classification results. We modified the class to numeric for ANN because it does not accept strings as input. To compare our results, we selected different concealed nodes (1, 3, 5, and 7) accordingly.

The initial step in using ANN was to load all of the soil data into an R data frame. The classes were string low, medium, and high fertile, which were altered to 1, 2, and 3 accordingly. The relationship between soil class and type has been established using the function percent of correctly categorized data in ANN. The table below shows the accuracy and error:

Table 1. ANN Accuracy and Error

ANN nodes	Training Steps	RMS	Prediction percentage
ANN with 1 node	2085	31.59	48
ANN with 3 nodes	3467	19.22	50
ANN with 5 nodes	4268	15.42	52
ANN with 7 nodes	9779	13.92	55

British Journal of Multidisciplinary and Advanced Studies:
Earth Sciences 4(6),1-11, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by the European Centre for Research Training and Development UK

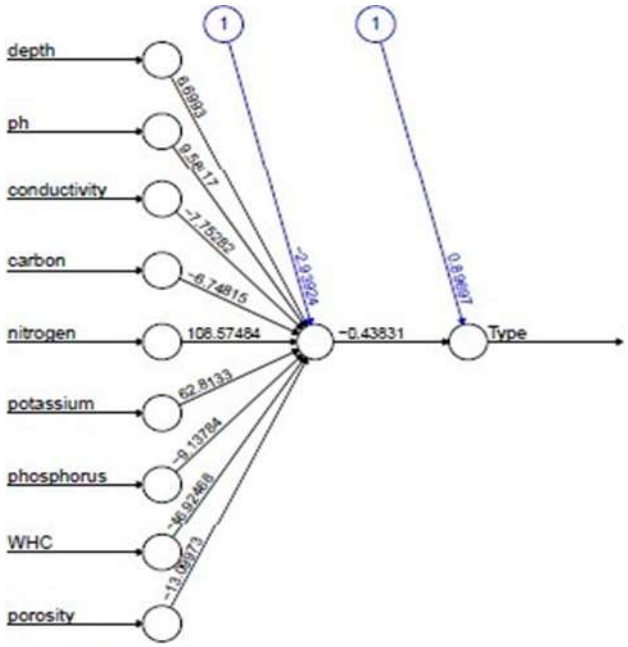


Figure 3. ANN with single hidden node

Support Vector Machine (SVM)

There no need to convert the final string class into numeric data for SVM. There are three labels on our Class label: low fertile, medium fertile, and high fertile. One of the most powerful machine learning techniques is the support vector machine. SVM incorporates both clustering and regression into one algorithm. SVM is a black box technique that is commonly used for classification and prediction problems. SVM can be considered of as a method for forming two distinct classes by forming a two-dimensional boundary on a surface between different data points. The choice boundary should be the one that is closest to both of the class's labels. The distance of a hyper plane is authenticated by support vectors. SVM cannot be used to classify this dataset. We used a variety of kernels, including Polynomial, Hyperbolic Tangent, and Radial Basis.

British Journal of Multidisciplinary and Advanced Studies:
Earth Sciences 4(6),1-11, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by the European Centre for Research Training and Development UK

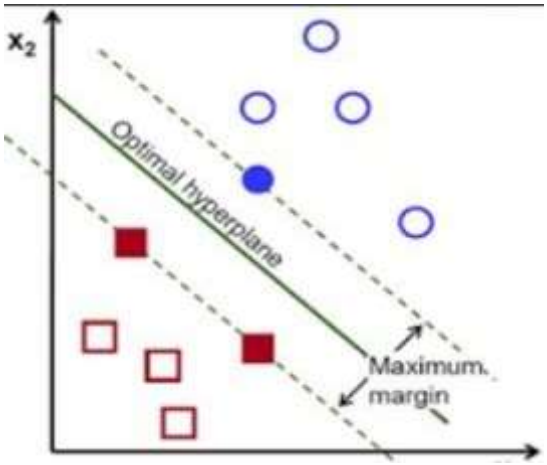


Figure 4. SVM hyper plane

The kernels used and their performance are described below. Table 2 shows the radial basis. Kernel provides the finest results.

Table 2. Description of SVM result for different Kernels

Kernel type	Prediction percentage	Support Vectors	Training error
Polynomial	69.00	767	0.32
Radial Basis	80.00	714	0.15
Hyperbolic tangents	44.00	751	0.62

Technique 2 for SVM:

For SVM, we've utilized a different nomenclature for classification called one vs all, as well as kernels like vanilla dot, RBF dot, poly dot, and spline dot for this dataset, all of which are included in the code. Perfection is calculated for each decent class label. The accuracy rate is given by spline-dot. For three different labels in that class, the accuracy, recall, precision, and F-measure of the SVM employing spline dot are shown below. The performance was calculated using the f measure.

Table 3. Accuracy, precision, recall & F- measure for SVM

Label Class	Accuracy	Precision	Recall	Fmeasure
High fertile	87%	66%	78%	71.5%
Med fertile	82%	76%	76%	76%
Low fertile	93%	83%	85%	83.9%

British Journal of Multidisciplinary and Advanced Studies:
Earth Sciences 4(6),1-11, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by the European Centre for Research Training and Development UK

K-nearest neighbors (KNN)

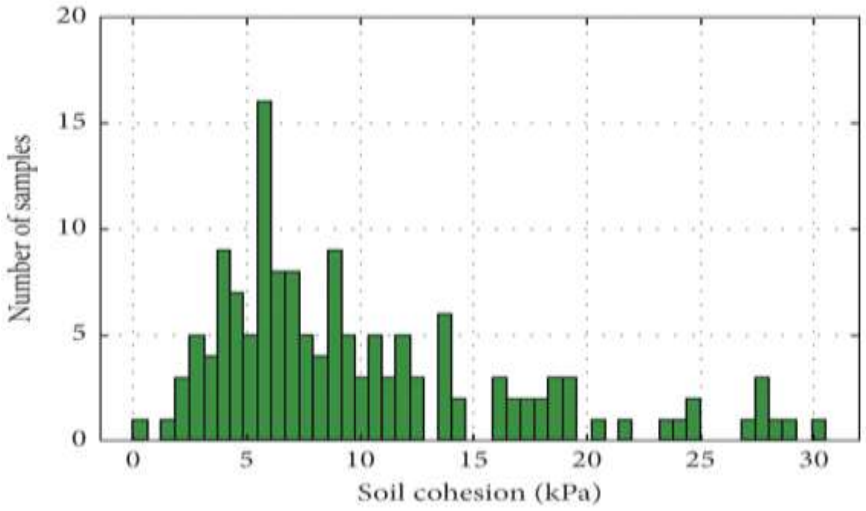


Figure 5. Fertile of Soil Rate

The marked class, which has three separate category values, was used. The values of this differ in the data, thus we utilized the normalization approach to convert the data to a common scale. Following that, the data was separated into two sets: training and testing, using a 70:30 ratio. Then I plotted the confusion matrix on the class label to check the precision Kappa statics using the K nearest neighbor function on the dataset with the value K = 17, which played a significant role in determining the model's competence, and the value calculated by the square root of observation by the test. Also, I plotted the histogram to see if the dataset was full.

Deployment

This is the last stage of the life cycle, where data precision is checked for all of the outcomes and a final report is written.

Classifier performs best for prediction of soil data

Decision tree, KNN, Artificial Neural Network, and Support Vector Machine are four important techniques that we used to classify the data. Data is separated into train and test for each algorithm. The performance of classifiers is compared based on how well they correctly classify data.

Table 4. Classifier Technique

Techniques	Accuracy
Decision Tree	63.48 %
ANN	55 %
SVM	80%
KNN	70%

CONCLUSION

The main motivation for this study was to try to mine and uniquely identify a dataset made up of various samples of soil from various parts of India in order to improve a model for guessing soil quality based on the chemical, biological, and physical compositions of that sampled soil, as well as to examine the consistency of the testing field. The resulting of the many created models can be used to reach a variety of conclusions. It's critical to grasp the limitations of this evaluation's scope before contemplating any possible outcomes. To begin, it's important to understand that the soil metrics collection includes a variety of soils. As a result, any conclusions that can be reached apply only to these specific soil kinds. Different categorization algorithms, such as Decision Tree, Artificial Neural Network, Support Vector Machine, and KNN, have been implemented and executed in this article. SVM trumps all other techniques in terms of results. The R tool was used to implement the CRISP-DM technique. In the future, we may be able to collect additional data from other regions of the states and country, and a soil suggestion system for commercial usage may be developed, which will aid in the growth of the Indian agriculture business.

REFERENCES

[1] Amalu, U.C., Isong, I.A. Status and spatial variability of soil properties in relation to fertilizer placement for intercrops in an oil palm plantation in Calabar, Nigeria. Niger. J. Crop Sci. 2018, 5, 58– 72.

[2] Prakash Kanade, Fortune David, Sunay Kanade, Convolutional Neural Networks (CNN) based Eye-Gaze Tracking System using Machine Learning Algorithm, European Journal of Electrical Engineering and Computer Science, Volume 5, Issue 2, pp 36-40, 2021.

[3] Akpan, J.F.; Aki, E.E.; Isong, I.A. Comparative assessment of wetland and coastal plain soils in Calabar, Cross River State. Glob. J. Agric. Sci. 2017, 16, 17–30.

[4] Bhattacharya, B., & Solomatine, D. P. Machine learning in soil classification. Neural Networks, Vol 19, pp.186-195, 2006.

[5] P Kanade, S Kanade, Raspberry Pi Project–Voice Controlled Robotic Assistant for Senior Citizens, International Research Journal of Engineering and Technology (IRJET), 7 (10), pp. 1044-1049, 2020.

[6] Schaap, M. G., Leij, F. J., & Van Genuchten, M. T. (1998). Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Science Society of America Journal, 62(4), 847-855.

[7] Pachepsky, Y. A., Timlin, D., & Varallyay, G. Y. (1996). Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal, 60(3), 727- 733.

[8] Ahmad, S., Kalra, & Stephen, H. (2010).A machine learning approach. Advances in Water Resources, 33(1), 69-80.

[9] Armstrong, L. J., Paul, M., Vishwakarma, S. K., & Verma, A. (2015, December).Analysis of Soil Behavior and Prediction of Crop Yield Using Data Mining Approach. In Computational Intelligence and Communication Networks (CICN), 2015 International Conference on (pp. 766-771) IEEE.

[10] Foody, G. M., & Mathur, A. (2004).A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on geo science and remote sensing, 42(6), 1335- 1343.

[11] Kanade, Prakash, and Jai Prakash Prasad, "Arduino based Machine Learning and IoT Smart Irrigation System", International Journal of Soft Computing and Engineering (IJSCE), Vol.: 10, Issue: 4, pp. 1-5, March 2021

[12] U. Mishra, V. Gupta, S. M. Ahzam and S. M. Tripathi, "Google Map Based Railway Track Fault Detection Over the Internet", International Journal of Applied Engineering Research, Vol. 14, pp. 20-23, Number 2, 2019.

[13] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

[14] P Kanade, P Alva, JP Prasad, S Kanade, Smart Garbage Monitoring System using Internet of Things (IoT), 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 330-335, IEEE, 2021.