# *Application of data mining classification techniques on soil data using R*

*Nikhita Awasthi*
*Amity University, Uttar Pradesh*
*Email: nikhitaawasthi@gmail.com*

*Dr. Abhay Bansal*
*HOD, Amity University, Uttar Pradesh*
*Email: abansal1@amity.edu*

*Abstract*— **As we are moving to a computerized and scientific world, data becomes an intrinsic part of our life. Every second petabytes of data are getting generated by social media, commercial websites, surveys, telecommunications, etc. and it needs to be analyzed and explored to discover some unknown facts, patterns, classifications and correlations. This task is accomplished through data mining and analysis. Data mining is a vast subject and involves use of databases, computer algorithms, computing performance, statistics and analysis. This research, compares two advance data mining techniques, Artificial neural network and Support vector machine on soil data. Apart from this, we have tried to enhance the performance of each model by applying different variations.**

***Keywords—****Data Mining, Artificial Neural Network, Support Vector Machine.*

## I. INTRODUCTION

In spite of the fact that Indian agriculture industry employs approximately 51 percent population of country, this sector accounts for just 18 percent of its annual GDP. The reason behind this mismatch is lack of use of technology and advance techniques in every step of production. When western countries are taking help of advance computer technologies in determining when, what, where and how of agriculture and animal husbandry, we have not even completely digitized our agriculture data. In this paper, we are developing a model of classification and prediction using some advance data mining techniques such as artificial neural network and support vector machine on the soil dataset. Another important purpose is to evaluate the relationship between different components of the soil. This data used in this paper has been from, ISRIC-World Soil Information [1] which is an independent institute founded by International Society of Soil Science and UNESCO.

Soil fertility is impacted by lot of factors like air, water, organic matter and nutrients. In present time soil fertility is on a verge of decreasing trend due to use of fertilizers, pesticides, insecticides, salinity, unscientific cultivation and urbanization. Most of the agricultural soil in India is deficient in primary nutrients which results in reduced production of food material and higher cost of food products [2]. It would be better to utilize saline, alkaline wastelands and fellow lands by analyzing different soil fertility parameters like fertility index and soil depletion index of diverse land use. To attain sustainability lot of research is being done with the help of bioinformatics, biotechnology and data analytics [3]. The eventual target of applying a technology in agriculture sector should be to augment the production while making less impact on fertility of soil and health of those who will be consuming the grown food products [4]. To attain this not only we need latest equipment and machines but also require to analyze the data of production and methodology which will help to identify the methods or processes that either are malpractices or advantageous.

## II. RESEARCH REVIEW

Various classification and statistical techniques have been used by researchers in agriculture and soil data. Bhattacharaya [5] et al applied Decision tree, ANN and support vector machines for classifying sub surface soil data measure from cone penetration testing. Schapp [6] et al applied neural network model to predict water retention parameter and saturated hydraulic conductivity from basic soil properties. Pachepsky [7] et al implemented artificial neural networks and regression for estimating soil water retention by using texture and bulk density. Ahmad [8] et al applied Support Vector Machine to predict soil moisture using remote sensing data. Baskar [9] et al applied Naive Bayes, J48, Jrip for classifying soil data. Armstrong [10] et al applied cluster analysis on soil profile data collected by department of agriculture and food of Australia. Paul [11] et al applied Naive Bayes and K nearest neighbor for predicting yield of crop with the help of soil dataset. Rossel [12] et al compared regression, partial regression, multivariate regression, Support vector machine, Random forest, boosted trees and artificial neural network to estimate soil organic carbon, clay content and pH measured in water on a dataset of !!00 soil profiles. Gholap [13] et al. did a comparative analysis of soil data using classification techniques Naïve Bayes, J48 and JRip and found out least median squares regression produce better result than classical linear regression. Chandrakar [14] et al. classified soil based on soil texture using various classification technique and found Bayesian

classification is the efficient technique. Ravindra [15] et al. used decision tree for evaluating the best suited pump for irrigation. Ruß [16] evaluated multiple regression technique on agriculture data and concluded support vector regression generated better model for yield prediction. Suman [17] et al. applied k-means clustering on soil data and later on used linear regression to classify the clusters. Behrens [18] et al. performed digital soil mapping using artificial neural network Foody [19] et al performed crop classification using support vector machine.

## III. DATA MINING

Data mining derives its basics from statistics, artificial intelligence and machine learning. Data mining is a dynamic process that enables a more intellectual use of a data warehouse than data analysis. It helps in building models that can be used to make predictions and analyzing massive amounts of data present in any dimension [20]. It helps to determine relationships among various key factors. There are various data mining software available in market as free ware and commercial product. One of them is R programming.

### *A. Why R*

R is one of the most preferred tool for data mining and analysis. It makes statistical computing easy and the programming effort is reduced. The graphs are easy to plot and depict. With the help of R various statistical and graphical techniques can be implemented. Advance statistical and data mining packages are provided by R. R programming software also provides us with various packages and in built functions which makes statistical analysis very easy. R provides well designed plots, effective data handling and storage facility. R is used in data pre-processing, data visualization, predictive analytics, statistical modeling and deployment [21].

### *B. Classification in Data Mining*

Data analysis can be done using classification and prediction, it helps in dividing the data into classes by which we can easily predict the future trends. Classification helps in putting labels on the data and in prediction helps in forecasting the value of data. There are various techniques of classification like decision tree classifier, rule based classifiers, Bayesian classifiers, support vector machine, k-n-n classifier.

### *C. Classification Algorithms*

It is a twostep process. In first step a model is built on collected and classified data also known as training data. This is the learning step in which classification algorithm is made by learning through training data. Since the class of each row of data is provided, this step is also called supervised learning. In second step the model is used for classification of unclassified data. The later one is prepared by imputing a part of training data which is only used for testing. Then the model prepared in first step is tested in second step. The results are compared with the actual classes of test data [22].

### *D. Artificial Neural Networks*

The working of Artificial neural networks is based on the functionality of neuron in life science. A neuron is a cell which transmits electrical signal after getting stimulated. This signal can be excited(increased) or inhibit(decreased) as required by the nervous system. Moreover, system has three types of cell, one for receiving information, second for transferring it and third for sending the information outside Similarly artificial neural network builds a network between input and output by connecting them with artificial neurons. These neurons carry some weight which can be decreased or increased with a motive to achieve the output [23].
All these neurons are interconnected to each other and model a parallel processor. There are various aspects of ANN needs to be defined for modelling.

- Activation function: It helps in transforming neuron's multiple input signal into single output signal.
- Network topology: It explains the model in terms of number of neurons and layers as well as the modus operandi in which they are connected.
- The training algorithm: It explains how much weights are assigned to each neuron to inhibit or excite them in order to achieve the input signal.

The potential of a neural network is embedded in its topology that is the manner neurons are connected to each other. There are infinite forms of ANN due to following factors

- Layers in the network
- Backward or forward movement of data.
- The count of nodes in each layer

During the first step the ANN is trained on soil data along with the class information. In second stage, the model uses network to predict the class of test dataset.

### *E. Support Vector Machine*

Support vector machine is also a black box technique used for classification and prediction problems. SVM combines the concept of regression as well as clustering, hence become one of the most powerful technique. It can be thought as surface which creates a two-dimensional boundary between two different classes of data point. This plane should be equidistant from both the datasets. The distance of this plane from boundary points are the support vectors. There are some datasets which cannot be classified with a simple SVM. In that case, various types of kernels are used as polynomial kernel, sigmoid kernel and Gaussian RBF kernel. There is no thumb rule for defining which kernel will work for which data. The fit depends a lot on data and relationship among the attributes [24]

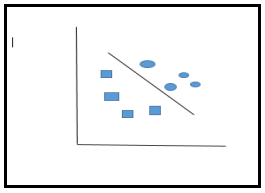

Fig 1. SVM with Linear Kernel

## IV. Data Modelling

The data for this research has been collected from ISRIC (International Soil Reference and Information Centre). [1] is an independent organization of soil data which records and stores the soil profiles from different part of world. They have the soil profiles for different parts of world. They have more than 50 attributes, out of which we selected 10 attributes [25] as follows: Depth, pH, Organic Carbon, Available Nitrogen, Available Phosphorus, Available Potassium, Porosity and Water Holding Capacity. The data was in the form of multiple excels. We downloaded the data into system and striped out extra attributes and rows where some data was missing. After data cleaning, we validated the data to check that all values are in correct units and range. All this work was done in excel. After data cleaning this data was loaded in R. R is a statistical tool which has all the data mining algorithms implemented in itself.

### *A. Depth*

It is the depth of the soil at which sample is collected. It is measured in cm. There are four depth, 0-15, 15-30, 30-45, 45-60.

### *B. Mineral Particle*

Mineral particles measure the presence of mineral like phosphorus, potassium in the soil. It is measured in ppm (particle per million) or gm/Kg.

### *C. Organic Carbon*

It is Organic Carbon is the carbon present in soil organic matter. It is measured in the form of percentage. It is directly proportional to the fertility of soil.

### *D. pH Value*

pH is a measure of acidity or alkalinity in soil. The ideal pH value for agriculture is 6.5 which is slightly acidic. In given data pH ranges from 3 to 10 approximately.

### *E. Nitrogen*

There are two types of measurement in soil data related to Nitrogen. First is available Nitrogen. which can be used by plants rest of the Nitrogen is in a form which cannot be consumed by plants.

| Data | Description |
|---|---|
| **Type** | **It describes the type of soil as less fertile, medium fertile and high fertile** |
| **Depth** | **The depth of soil in cm** |
| **Ph** | **pH of soil** |
| **Conductivity** | **Electrical conductivity of soil** |
| **Organic Carbon** | **Percentage of organic carbon in soil** |
| **Nitrogen** | **Available nitrogen in soil** |
| **Phosphorus** | **Available phosphorus in soil** |
| **Potassium** | **Available potassium in soil** |
| **Water holding capacity** | **Water holding capacity in percentage** |
| **Porosity** | **Porosity in percentage** |

Table. 1. Soil Attributes of Input Data

## V. Results and Discussion

The first step was analyzing the statistical results like minimum, maximum, mean and median, first quartile and third quartile. All the results have been generated with the help of R. The results are as shown in the below table.

| | Depth | pH | Electrical Conductivity | Organic Carbon | Available Nitrogen | Available Phosphorus | Available Potassium | Water Holding Capacity | Porosity |
|---|---|---|---|---|---|---|---|---|---|
| Min. | 1 | 5.6 | 40 | 0.015 | 3.98 | 4 | 60 | 27.6 | 29.3 |
| 1st Quartile | 1 | 8.2 | 265 | 0.12 | 17 | 14.43 | 200 | 42 | 40 |
| Median | 2 | 8.81 | 410 | 0.235 | 30.52 | 19.84 | 290 | 46.33 | 44.5 |
| Mean | 2.427 | 8.804 | 499.1 | 0.3949 | 37.08 | 24.29 | 379 | 47.37 | 44.96 |
| 3rd Quartile | 3 | 9.7 | 660 | 0.5887 | 50.4 | 32.3 | 400 | 52.54 | 49 |
| Max. | 4 | 11.5 | 1720 | 2.35 | 185 | 82.42 | 3000 | 76.8 | 65.72 |

Table. 2. Statistical summary of data

Below is the graphical view of data along with the classes of each variation.

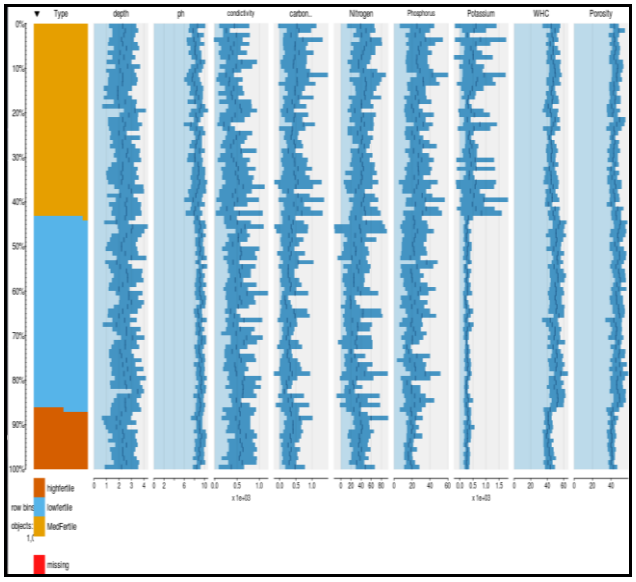


Fig. 2. Soil variation graph

Below is the correlation matrix among each variable of dataset. The upper triangle of matrix reports the correlation coefficient between each pair and lower triangle shows the that coefficient graphically. A few pair show very strong correlation as (ph, conductivity), (carbon, nitrogen) and (Porosity, Water holding capacity).

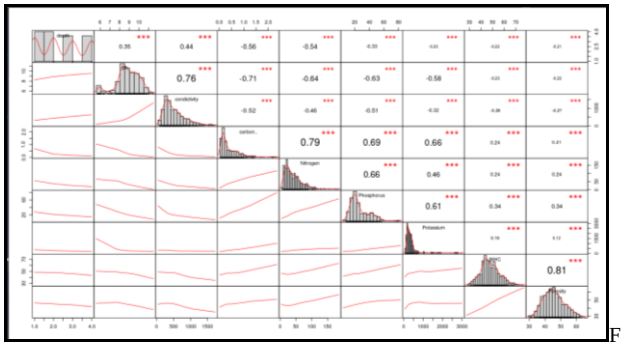

Fig.3 Correlation matrix among the variables of soil data

*A. ANN results*

To apply ANN data was loaded in R along with the information of classes. In this data classes are in character like low fertile, midfertile and high fertile. We converted these classes into numeric like 1,2,3. Since the range of each column is different data has been normalized by using following formula.

pH-min(pH)/max(pH)-min(pH) eq. 1

Max(pH) and min(pH) are max and min value of pH in data column After normalization data was divided into two parts to generate training data and test data. Next step was to apply ANN with one, five and seven hidden nodes. The performance of ANN with these three models is as follows.

| | Prediction percentage | Training Steps | RMS |
|---|---|---|---|
| ANN with single node | 48.00% | 3378 | 31.34 |
| ANN with 5 hidden node | 52.00% | 61268 | 19.45 |
| ANN with 7 hidden node | 55.00% | 73073 | 15.8 |

Table.3 Comparison of various ANN models

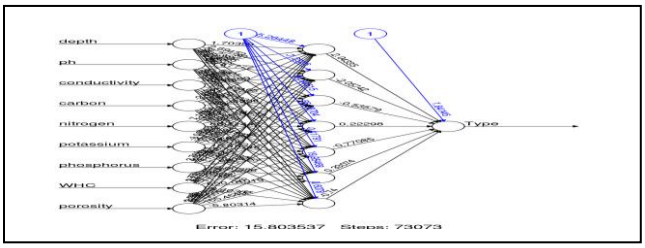


Fig 4: ANN Model with seven hidden nodes

*B. SVM Results*

For SVM we do not need to convert final class into numeric data. Hence after loading and normalizing the data SVM is applied with three different kernel, polynomial, Radial Basis and Hyperbolic tangent. The results of three variations is reported in below table

| kernel type | Prediction percentage | Support Vectors | Training error |
|---|---|---|---|
| Polynomial | 68.00% | 597 | 0.35 |
| Radial Basis | 74.00% | 543 | 0.2 |
| Hyperbolic tangents | 43.50% | 519 | 0.62 |

Table 4. Comparison of various SVM kernels

*C. Comparison of ANN And SVM*

We have tried to improve the two models with their respective variations. In case of ANN we achieved highest performance of 55 percent with 7 hidden nodes. The model took 73073 steps to train the classifier and root mean square error is of 15. In case of SVM we have achieved much better results with 74% using Radial basis kernel.

## VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a comparison of two data mining techniques of Artificial Neural Network and Support Vector Machine. This study is done with the help of data analytics tool R. In future, we would like to add more techniques for the classification and prediction, and will try to improve the performance of these classifiers with other available methods. We will also develop a fertility recommendation system. It will need more data to be collected from different parts of India and will prepare a digital map of soil. This will become input of new system and will help in identifying different profiles of soil.

## REFERENCES


[1] Batjes, N.H., 2002. Soil parameter estimates for the soil types of the world for use in global and regional modelling (Version 2.1; July 2002). ISRIC Report 2002/02c

[2] Sahrawat, K. L., Wani, S. P., Rego, T. J., Pardhasaradhi, G., & Murthy, K. V. S. (2007). Widespread deficiencies of Sulphur, boron and zinc in dryland soils of the Indian semi-arid tropics. Current Science, 93(10), 1428-1432.

[3] Sanghpriya, R., & Vohra, R. Application of machine learning in agriculture.

[4] Hooda, P. S., Henry, C. J. K., Seyoum, T. A., Armstrong, L. D. M., & Fowler, M. B. (2004). The potential impact of soil ingestion on human mineral nutrition. Science of the Total Environment, 333(1), 75-87.

[5] Bhattacharya, B., & Solomatine, D. P. (2006). Machine learning in soil classification. Neural Networks, 19(2), 186-195.

[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. [565]

[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. [320]

[8] Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. Advances in Water Resources, 33(1), 69-80.

[9] Baskar, S. S., Arockiam, L., & Charles, S. (2013). Applying data mining techniques on soil fertility prediction. International Journal of Computer Applications Technology and Research, 2(6), 660-meta.

[10] Armstrong, L. J., Diepeveen, D., & Maddern, R. (2007, December). The application of data mining techniques to characterize agricultural soil profiles. In Proceedings of the sixth Australasian conference on Data mining and Analytics-Volume 70 (pp. 85100). Australian Computer Society, Inc.

[11] 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.
[12] Rossel, R. V., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*(1), 46-54.
[13] Gholap, J., Ingole, A., Gohil, J., Gargade, S., & Attar, V. (2012). Soil data analysis using classification techniques and soil attribute prediction. *arXiv preprint arXiv:1206.1557*.
[14] Chandrakar, P. K., Kumar, S., & Mukherjee, D. (2011). Applying classification techniques in Data Mining in agricultural land soil. *International Journal of Computer Engineering*, *2*, 89-95.
[15] Ravindra, M., Lokesha, V., Kumara, P., & Ranjan, A. Study and Analysis of Decision Tree Based Irrigation Methods in Agriculture System.
[16] Ruß, G. (2009, July). Data mining of agricultural yield data: A comparison of regression models. In *Industrial Conference on Data Mining* (pp. 24-37). Springer Berlin Heidelberg.
[17] Suman, B. B. N. (2013). Soil classification and fertilizer recommendation using WEKA. *IJCSMS Int. J. Comput. Sci. Manag. Stud*, *13*(5).
[18]" Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E. D., & Goldschmitt, M. (2005). Digital soil mapping using artificial neural networks. *Journal of plant nutrition and soil science*, *168*(1),21-33.
[19] Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on geoscience and remote sensing*, *42*(6), 1335-1343.
[20] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.
[21] Team, R. C. (2013). R: A language and environment for statistical computing.
[22] Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.
[23] Sawaitul, S. D., Wagh, K. P., & Chatur, P. N. (2012). Classification and prediction of future weather by using back propagation algorithm-an approach. *International Journal of Emerging Technology and Advanced Engineering*, *2*(1), 110-113.
[24] Gupta, M., & Aggarwal, N. (2010, March). Classification techniques analysis. In *Proceedings of National Conference on Computational Instrumentation*(pp.120-8).
[25] Rahi, T. S., Singh, K., & Singh, B. (2013). Screening of sodicity tolerance in Aloe Vera: An industrial crop for utilization of sodic lands. *Industrial Crops and Products*, *44*, 528-533.