Predicting Fertility of Soil Using Data Mining Techniques

Team N
School of Computing
National College of Ireland
Nikhil Deshmukh, Shailesh Rathi, Shital Thakare, Siddharth Chaudhary
{x16145003, x16138163, x16139739, x16137001}@student.ncirl.ie

Abstract— Abstract— **Agriculture industry in India is the greatest sector for employment but lack of research in this sector is the reason behind less productivity. It's important to implement computational research, Machine learning techniques in Agriculture industry to make India better quality and quantity producer in food sector. Machine Learning techniques are useful in abstracting patterns and establishing relationships between varied data sets and predicting reasonable outputs. It can be efficiently applied in Agriculture industry to improve efficiency in this sector. We have discussed application of Machine learning techniques in Agriculture sector to analyze fertility of soil. Agriculture industry has been always one of the interested area of research. This study venture to analyze soil data depending upon various factors, classify it and improve efficiency of each model using different combinations.**

Keywords— ANN; SVM; Decision Tree; kNN; Soil;

I. INTRODUCTION

Agriculture is one of the highest employer sector in India as well as worldwide. Around 51% population in India has been employed by Agriculture industry. Contradictory to that it is accountable for only 18% of annual GDP. The mismatch is due to lack of research and less use of technologies. Agriculture sector in India is less automated as compared to western countries. Western countries using various technology to do predictions in Agriculture. Indian agriculture sector is lacking behind in this area. Productivity and growth is very much less due to lack of technology. We have considered Indian soil data to predict the fertility of soil considering various elements like nitrogen, oxygen, Phosphorous, porosity and various environmental elements like air, temperature. Data has been taken from ISRIC -World soil information [1]. ISRIC is an institute which works on soil science. India has diverse weather conditions as well as soil type across different regions. Data set belongs to Haryana state in India which ground of moist soil.

fertility of soil the limiting factor in Agricultural industry India. Soil fertility defines growth of plants when other environmental factors like light, water, temperature are favorable. Soil fertility is influence by several factors like Climate, irrigation (Soil water), Soil, acidity, Soil alkalinity, Nutrition in Soil. Globalization, changing weather condition, urbanization, higher use of pesticides is the reason of decreasing quality of soil in India. Deficient soil type lead to less agricultural production and ultimately higher cost of food products. Different soil types are used to analyze fertility of soil. The ultimate goal of applying technology in Agricultural with minimal impact in fertility of soil and quality of food product [4].

Machine Learning techniques are useful in abstracting patterns and establishing relationships between varied data sets and predicting reasonable outputs. Agriculture industry in India is the greatest sector considered for employment and has been part of research. Machine learning techniques can be efficiently applied in Agriculture industry to improve research. In current scope of project, we have developed a model for fertility of soil based on different soil type. After receiving fertility depending upon various soil type, a comparative study of machine learning techniques such as ANN, linear regression, SVM and Decision tree is carried out.

II. RELATED WORK

Agriculture is one of the hot topic of research among academic as well as scientific researchers. A lot of previous research has been done in Agriculture industry and fertility of soil by using different data mining, classifications and statistical techniques.

Various factors affect fertility of soil. Among which Sulphur, Water and Zinc are the most influencing factors constituting fertility of soil. A agriculture case study carried out on 3622 soil samples from different districts in India. It is concluded in study that water or moisture level in soil is the most influencing factor. Soil fertility varies depending moisture level in soil[15].

Author[1]used segmentation algorithm to divide signals and features. Signals are extracted using boundary method and then classifiers divided into classes; they used SVM, Decision Trees, and ANN to classify surface soil data. Author[2]developed hierarchical neural network models to predict water retention and hydraulic conductivity. Author[3]developed regression and artificial neural network to check the water retention with the help of texture and bulk density. Ahmad [4] predicted soil moisture using SVM for data sensing remotely. Author[6] implemented linear regression technique for the forecasting of soil data along with

Naïve Bayes, J48 classification. Armstrong [5] implemented cluster analysis on soil data collected by the food department of Australia.

Author[8] used different classification techniques on soil texture and found Bayesian classification is more accurate and performance is also good. [9]Gholap did the comparative analysis between classification techniques; such as Naïve Bayes, J48 and classical linear regression and found least median regression produced better result. Author [10]implemented artificial neural network and digital terrain-analysis for high quality soil maps. Author[12] Foody implemented single Support Vector Machine against series of the classifier for crop classification.

[13]implemented decision tree to the fields to help the farmers in making the decision to select pump for the irrigation and it's depend on irrigation types, total area coverage of the field, capacity of the motor, and the height. Several techniques applied different classifications machine learning methods, such as J48, Naïve Bayes, and Random forest algorithm to classify the fertility of the soil. J48 gives better result than other algorithms. Rub implemented multiple regression techniques on soli data and concluded SVM generated better model for the predication.

III. METHODOLOGY

Cross Industry Standard for Data Mining Model (CRISP-DM) has been used among various methodologies for our data mining project as it best suits our business due to its extensive step by step development ability. This is the six-phase process which will be covered in the remaining sections. **[CHAPMAN]**

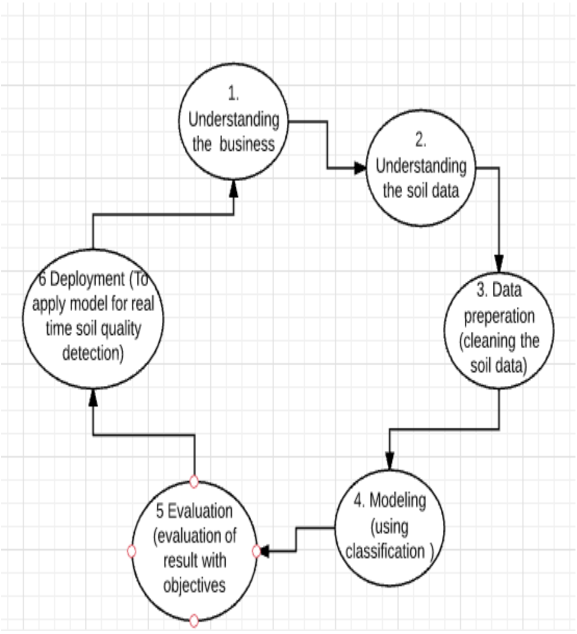

1. Understanding the business
2. Understanding the soil data
3. Data preperation (cleaning the soil data)
4. Modeling (using classification)
5 Evaluation (evaluation of result with objectives
6 Deployment (To apply model for real time soil quality detection)

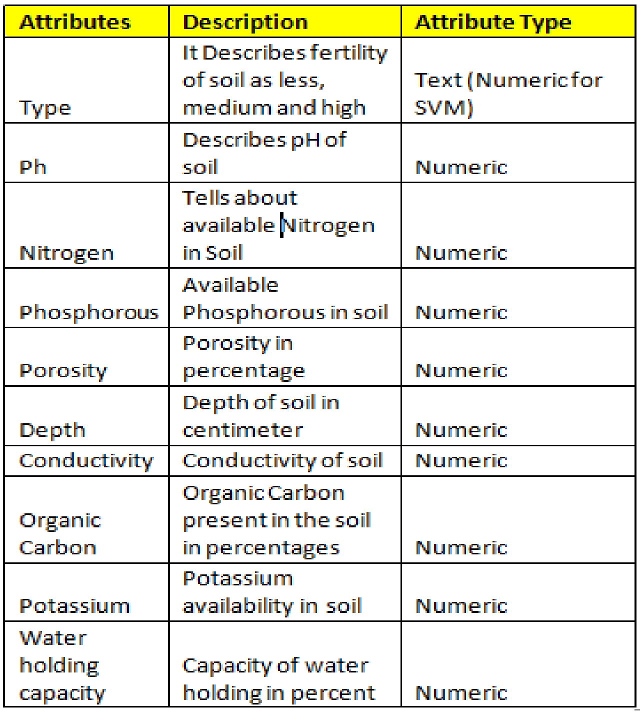
Fig 1: Understanding the business

A. *Understanding the Bisiness :* This is the first step in the process of data mining, before proceeding to the next step all the informations has gathered related to the Agriculture and soil business.

Business Objectives and Business Questions are as follows:

1. Which properties are important in defining the fertility of soil?
2. Which properties of soil are highly correlated to each other?
3. Which properties do not vary with respect to fertility of soil?
4. Which Classifier performs best for predicion of soil data?

B. *Understanding the soil data :* Raw data set collected from the ISRC *(*International Soil Reference and Information Centre) an organization which works in soil research area as well as different types of soils across world. Data set selected had 60 attributes, relevant attribute as per research has been selected. Final dataset contains 10 different attributes and around 1300 records. Attributes arewhich are as follows:



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

Table1: Data set details

C. *Data Preperation :*

. Data cleaning and formatting needed before using it as final input. Removed unnecessary columns, null values, extra blank

spaces in dataset using R programming language and basic excel functionality. Some packages like readr, tidyr have been used.
Following step has been performed in data cleaning process.

1. Null Values handling: There were very few null records in dataset. Null records replaced with NA values to remove inconsistency in the data set.
2. Removing irrelevant columns: Data set contain some columns which were irrelevant for our research, removed those unnecessary columns, only relevant data has been taken as input for processing.
3. Inconsistent Data Types: Numeric values in data set were not in consistent format ex. Precision and scale defined for 'Porosity' column was not consistent, unique precision modified for this column.

4.Unessary Spaces handling:

Used 'TRIM' function in excel to remove unnecessary spaces in column values.
Processed clean data is used as input for further analysis.

D. Data Modeling and Evaluation

Sequence for data modeling and evaluation is selecting the technique, generate, test design, building a model, model assessment, evaluation of result, Review and determine the next step **[ASTESJ]**

1) Decision Tree

QUESTION 1: Which properties are important in defining the fertility of soil?

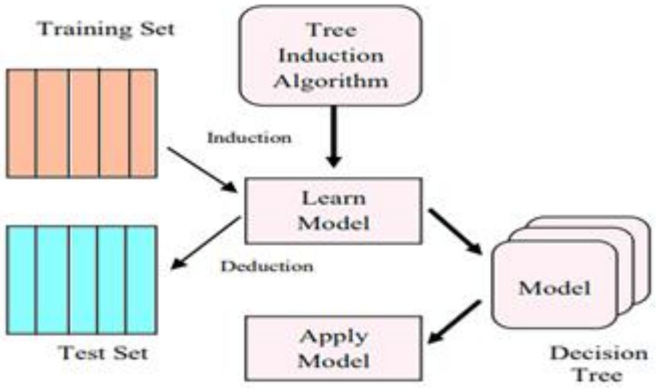


Fig : 2 Decision tree model

In our project, we used decision tree as classification model. There are three labels in the class in dataset depending upon fertility of soil which are High Medium and low. Various dependent factors considered are ~ ph,depth, conductivity ,carbon ,Nitrogen, Phosphorus, Potassium, WHC, Porosity. Once data is loaded, shuffling is done on dataframe. Shuffled dataframe is then divided into train and test, 70:30 percent ratio is maintained for train and test. Decision tree is applied on factors mentioned dependent factors mentioned above. R part function is used to create decision tree and classes are predicted depending variables. A prediction model is prepared using decision tree as input on test data set. Graph is plotted using prediction model.

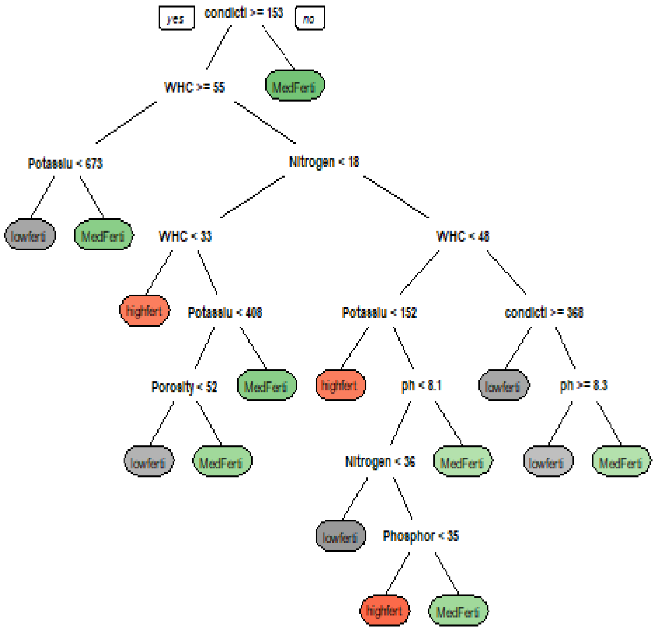
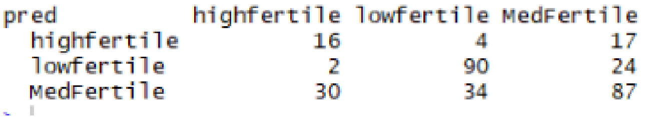


Fig :3 Decision tree for soil data



```
pred          highfertile lowfertile MedFertile
  highfertile          16          4         17
  lowfertile            2         90         24
  MedFertile           30         34         87
```

Table 2: Confusion matrix for DT

Accuracy= sum(diag(table))/sum(table)
accuracy = (16 + 90 + 87)/ (16 + 4 + 17 + 2 + 90 + 24 + 30 + 34 + 87)
accuracy = 193 / 304 = 0.63486
accuracy = 63.48 %

ANSWER 1: To answer this question we have applied decision tree classifier on the soil data, decision tree evaluates the entropy of each attribute to decide the root of the tree.
Decision tree classification is shown in the figure, the root of the tree is **Conductivity**, which we can assume is the most important criteria, next one will be the water holding capacity.

Below is the list of factors which are arrange in order of significance of role in classification

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

2) ANN

QUESTION 2: Which properties of soil are highly correlated to each other?

We have used ANN as it gives good result for classification [5]. For ANN, we have changed the class as numeric as ANN does not take the string as an input. We have used different hidden node to compare our result 1, 3, 5 and 7 respectively.
While implementing ANN first step was to load the complete soil data in R data frame. In this, classes were in the form of string High, mid and low fertile which was changed into 3,2,1 respectively. Due to different Range of ever column, data has normalized. Using the min max formula. Data has been split into train and test with 80:20 split. Data has been trained using neuralnet function by using neuralnet package for different hidden nodes. Using cor function percentage of corret classified data in ANN has been identified between class and type of the soil.
Accuracy and error can be seen in below table:

ANN nodes	Training Steps	RMS	Prediction percentage
ANN with 1 node	2085	31.59	48
ANN with 3 nodes	34672	19.22	50
ANN with 5 nodes	42683	15.42	52
ANN with 7 nodes	97794	13.92	55

Table 3 Accuracy and error table for ANN

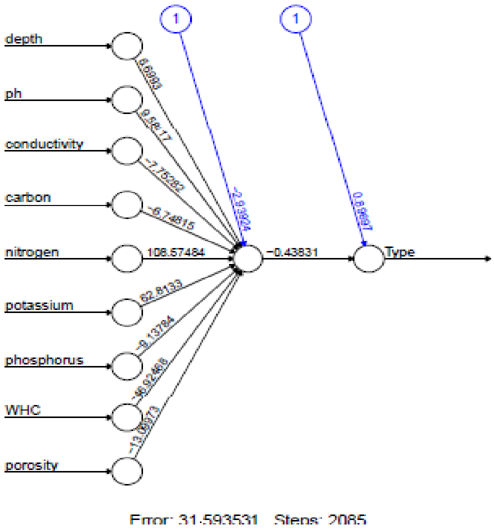


Fig4 : ANN with single hidden node

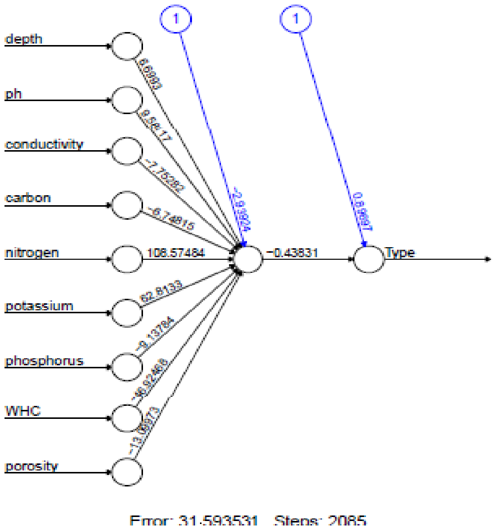


Fig 5: ANN with 7 hidden nodes

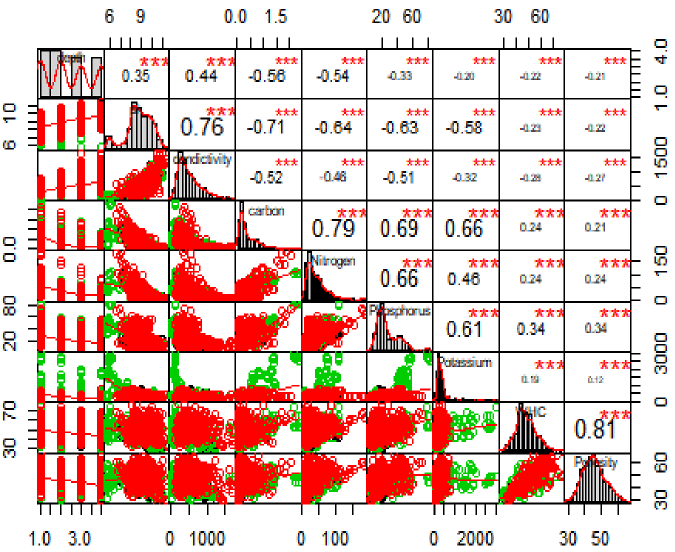


Fig : 6 Matrix correlation

ANSWER 2: To answer this question we have performed correlation analysis with the help of R Fig 6 Matrix correlation represents the relationship between the entire soil data as we can see from Fig 6, the highest correlation of 0.81 is between WHC and porosity. Next 2 properties which are highly correlate are Carbon and Nitrogen with coefficient value of 0.79, next one is conductivity and pH. The upper triangular matrix of this fig shows the correlation coefficient and lower triangular matrix is consisting of scatter plot between each pair of variable.

3) SVM

For support vector machine, we do not need to convert the final string class into numeric data. Our Class label contain 3 labels: low fertile, med fertile and high fertile.

Support vector machine is one of the most powerful technique in machine learning. SVM combines both the concepts: clustering as well as regression. SVM is a black box technique which is generally used for prediction and classification problems. SVM can be thought as which creates a two-dimension boundary on a surface between different data points to form two different class. The decision boundary should be equidistant from both the labels of the class. Support vectors validate the distant of a hyperplane.as shown in Fig 7. This dataset can't be classified using simple SVM. We have used different kernels like Polynomial, Radial Basis, Hyperbolic Tangent.

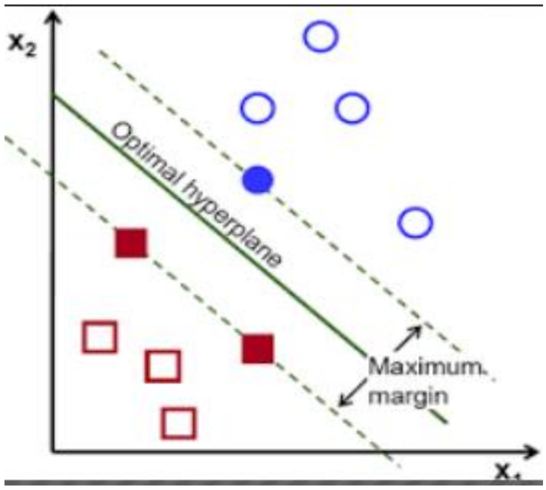


Fig 7: SVM hyper plane

Here is the description of the kernel used and their performance. Table 4 radial basis Kernel gives best performance.

Kernel type	Prediction percentage	Support Vectors	Training error
Polynomial	69.00	767	.32
Radial Basis	80.00	714	.15
Hyperbolic tangents	44.00	751	.62

Table 4: Description of SVM result for different Kernels

Technique 2 for SVM: For SVM, we have also used the technique of one vs all for classification and kernels like polydot, vanilladot, rbfdot, splinedot for this dataset which are given in the code. For each label for the class accuracy is calculated. Splinedot gives the best accuracy rate. Below is the accuracy, precision, recall, F-measure of the SVM using splinedot for 3 different labels of the class. we have used f-measure to calculate the performance.

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

Table 5: Accuracy, precision, recall and F-measure for SVM

4) KNN

QUESTION 3: Which properties do not vary with respect to fertility of soil?

After applying all the classification methods on Soil data ANN, SVM and DT, then we applied kNN classification method on class labeled type, to check classification accuracy of the algorithm. In this algorithm, we used on the labeled class, which contains three categorical values.
In the data, values of the variables are different, so we used normalization technique to transform the data into common scale. After that, data divided into two sets training and testing with 70 :30 ratio. And then used kNN function on the dataset with the value K= 17, which played significant role to determine the efficiency of the model, and value calculated by the square root of the test observation, and plot the confusion matrix on the class label to check the accuracy, Kappa statics. And plotted the histogram to check the accuracy of the dataset.

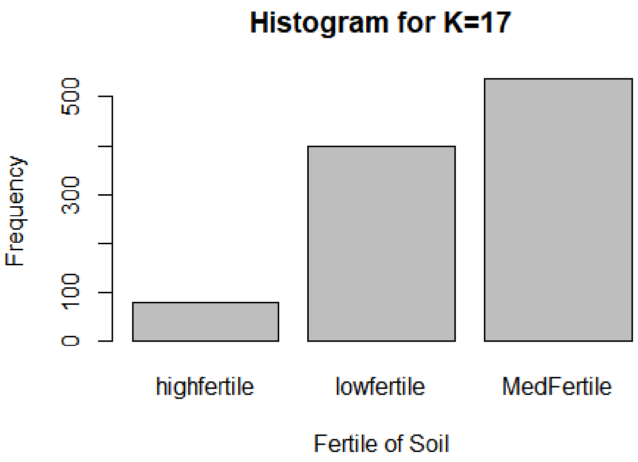


Fig:8 Frequency of soil fertility

ANSWER 3: To answer this question we have performed ANOVA of each property according to soil type in R and results are as follows

Property	F-Statistic	P Value
Potassium	0.42	0.51
WHC	352.88	2.2e-16***
Phosphorous	16.387	5.473e-05***
Nitrogen	30.074	5.004e-08***
Carbon	8.5978	0.003
pH	3.84	0.0502

Table 6: ANOVA output

From the Table 6 we can state that Pottasium, Carbin and pH are not statistically significant. So, WHC, Nitrogen and Phosphorous does not vary with respect to the fertility o soil.

QUESTION 4: Which Classifier performs best for prediction of soil data?

ANSWER 4: In this question, we have applied four important techniques of classification which are decision tree, ANN, SVM, kNN. For each algorithm data is divided in test and train. The performance of classifier is compared on fact that how each of them has classified data correctly.

Techniques	**Accuracy**
Decision Tree	63.48 %
ANN	55 %
SVM	80%
kNN	70%

Table 7: Comparison of accuracy

From the above Table 7 we can conclude that SVM performs best among all the technique.

5) Deployment

This is 6th and final phase of the life cycle in which accuracy is monitored for all the results and final report has been prepared.

IV. Conclusion and Future Work

In this paper, we have implemented and executed different classification methods such as Decision Tree, ANN, SVM and kNN. The result of SVM out performs among all the techniques. CRISP-DM methodology has been used with the help of R tool.
In Future, we can collect more data from different parts of the country and soil recommendation system can be built for the commercial use which can help grow agriculture industry.

References

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

[2] 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]

[3] 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]

[4] 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.

[5] 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.

[6] 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..

[7] 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.

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] Ravindra, M., Lokesha, V., Kumara, P., & Ranjan, A.

Study and Analysis of Decision Tree Based Irrigation Methods
in Agriculture System.

[13] Herman,Robinson T,Giovanni P,Alvaro N. A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change.

[14] 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.