SOIL ANALYSIS USING MACHINE LEARNING

1.Vasu Aggarwal, 2. Abhishek Jawla, 3. Chetan Singh Tomer

StudentsDepartment of Computer Science & Engineering, Raj Kumar Goel Institute of Technology

Abstract –The title of the project is **"SOIL ANALYSIS USING MACHINE LEARNING"** In India, the greatest agriculture 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 the Agriculture industry to make India better quality and quantity producer in the 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 the Agriculture industry to improve efficiency in this sector. We have discussed the application of Machine learning techniques in Agriculture sector to analyze the fertility of the soil. The agriculture industry has been always one of the interesting areas of research. This study venture to analyze soil data depending upon various factors, classify it and improve the efficiency of each model using different combinations.

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

I. INTRODUCTION

Around 51% of the population in India has been employed by the Agriculture industry. Contradictory to that it is accountable for only 18% of annual GDP. The mismatch is due to the lack of research and less use of technologies. The agriculture sector in India is less automated as compared to western countries. Western countries using various techniques to do predictions in Agriculture. The Indian agriculture sector is lacking behind in this area. Productivity and growth are very much less due to the lack of technology. The fertility of soil the limiting factor in Agricultural industry India.Soil fertility defines the growth ofplants when other environmental factors like light, water, temperature are favourable. 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 for decreasing the quality of soil in India. Deficient soil type leads to less agricultural production and ultimately higher cost of food products. Different soil types are used to analyze the fertility of the soil. The ultimate goal of applying technology in Agricultural with minimal impact infertility of soil and quality of food product. Machine Learning techniques are useful in abstracting patterns and establishing relationships between varied data sets and predicting reasonable outputs. The agriculture industry in India is the greatest sector considered for employment and has been part of the research. Machine learning techniques can be efficiently applied in the Agriculture industry to improve research. In the current scope of the project, we have developed a model for the fertility of soil based on different soil type. After receiving fertility various soil machine learning techniques such as ANN, linear regression, SVM and Decision tree is carried out.

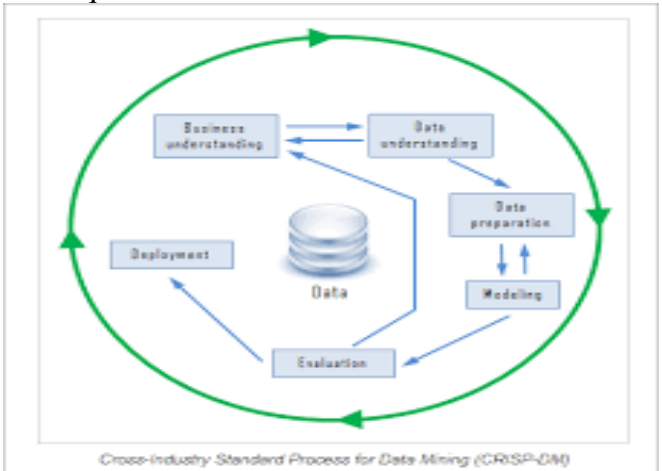
II. RELATED WORK

Agriculture is one of the hot topics of research among academic as well as scientific researchers. A lot of previous research has been done in the Agriculture industry and fertility of the soil by using different data mining, classifications and statistical techniques. Various factors affect the fertility of the soil. Among which Sulphur, Water and Zinc are the most influencing factors constituting the fertility of the soil. An agriculture case study carried out on 3622 soil samples from different districts in India. It is concluded in the study that water or moisture level in the soil is the most influencing factor. Soil fertility varies depending on the moisture level in the soil. It used the segmentation algorithm to divide signals and features. It will be used boundary method and then classifiers divided into classes; they used Decision Trees, and ANN, SVM to classify surface soil data.

It is developed hierarchical neural network models to predict water retention and hydraulic conductivity. It is developed regression andartificial neural network to check the waterretention with the help of texture and bulk density. It is predicted soil moisture using SVM for data sensing remotely. It is implemented linear regression technique for the forecasting of soil data along with Naïve Bayes, J48 classification. It is implemented cluster analysis on soil data collected by the food department of Australia. It is used different classification techniques on soil texture and found Bayesian classification is more accurate and performance is also good. It is implemented artificial neural network and digital terrain- analysis for high-quality soil maps. It is implemented decision tree to the fields to help the farmers in making the decision to select a pump for the irrigation and it depends on irrigation types, total area coverage of the field, the capacity of the motor, and the height. Several machine learning techniques, such as J48, Naïve Bayes, and random forest algorithm to classify the fertility of the soil. J48 gives a better result than other algorithms. Rub implemented multiple regression techniques on soli data and concluded SVM generated a better model for the prediction.

III. METHODOLOGY

In the Data Mining Technology the among various techniquessuch as



Cross-Industry Standard Process for Data Mining (CRISP-DM)

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

B. Understanding the soil data: Data set selected had 60 attributes. The final output of the dataset contains 1300 records and 10 different attributes.

C. Data Preparation :
Removed unnecessary columns, null values, extra blank spaces in dataset using R programming language and basic excel functionality. Some packages like the reader, tidy have been used. The following step has been performed in the data cleaning process.

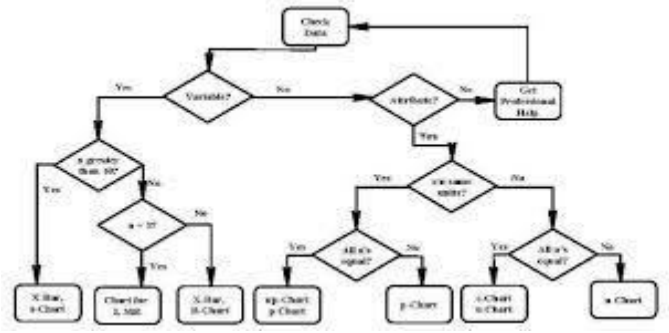
1. Null Values handling: There were very few null records in the dataset. Null records replaced with NA values to remove inconsistency in the data set. 2. Removing irrelevant columns: Data set to contain some columns which were irrelevant for our research, removed those unnecessary columns, only relevant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

4.Unnecessary Spaces handling:

Used 'TRIM' function in excel to remove unnecessary spaces in column values. D. Data Modeling and Evaluation Sequence for data modelling and evaluation in selecting the technique, generate, test design, building a model, model assessment, evaluation of the result.

1) Decision Tree
In our project, we used a decision tree as a classification model. There are three labels in the class in dataset depending upon the fertility of the soil which is 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 the data frame. The shuffled data frame is then divided into train and test, 70:30 per cent ratio is maintained for train and test. The decision tree is applied to 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 a decision tree as input on the test data set. A graph isplotted using the predictionmodel**.**



Below is the list of factors which are arranged in order of significance of role in classification 1) Conductivity 2) WHC 3) Potassium 4) Nitrogen 5) pH 6) Phosphorous

2). ANN

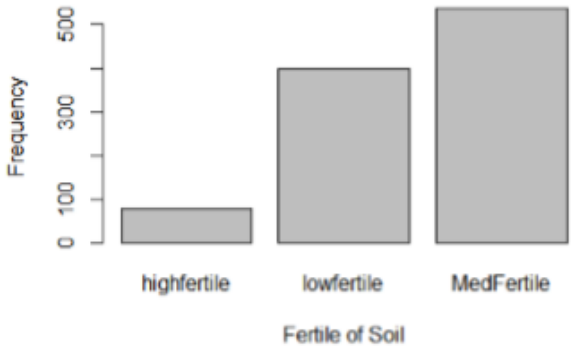
We have used ANN as it gives a good result for classification. For ANN, we have changed the class as numeric as ANN does not take the string as an input. Using the min-max formula. Data has been trained using neural net function by using the neural net package for different hidden nodes. Using cos function percentage of correctly classified data in ANN has been identified between class and type of the soil.

3) SVM

For support vector machine, we do not need to convert the final string class into numeric data. Our Class label contains 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-dimensional 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. This dataset can't be classified using simple SVM. We have used different kernels like Polynomial, Radial Basis, Hyperbolic Tangent.

Technique 2 for SVM: For SVM, we have also used the technique of one vs all for classification and kernels like poly dot, vanilla dot, RBF dot, spline dot for this dataset which are given in thecode. 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 spline dot for 3 different labels of the class. we have used f measure to calculate the performance.

4) KNN



In this algorithm, we used on the labelled class, which contains three categorical values. In the data, the values of the variables are different, so we used the normalization technique to transform the data into a common scale. After that, data divided into two sets of training and testing with 70:30 ratio. And then used kNN function on the dataset with the value K= 17, which played a significant role to determine the efficiency of the model, and the 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.

5) Deployment

This is 6th and final phase of the life cycle in which accuracy is monitored for all the results and the 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 outperforms 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 the 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). A machine learning approach. Advances in Water Resources, 33(1), 69-80. [5] Armstrong, L. J., Diepeveen, D., & Maddern, R. (2007, December).