# PREDICTIVE DATA MINING ALGORITHMS FOR OPTIMIZED BEST CROP IN SOIL DATA CLASSIFICATION

G. Divya[1], G. Bharathi Mohan[2]

[1]PG student, [2]Assistant Professor, Jaya Engineering College, Thiruninravur, Chennai-602 024.

divyaganesan8682@gmail.com

***Abstract***

Agricultural research has strengthened the optimized economical profit, internationally and is very vast and important field to gain more benefits.

In future agriculture is the only scope for all the people. But today number of people having land, but they don't know how to yield the crops.

So many of people are doing useless agriculture by cultivating the crop on improper soil. To implement the application to identify the types of soil,water source of that land whether that land is based on rain or bore water. And suggest what of crop is suitable for that soil. So through this application provide application for the people to know about the agriculture. There is no any application to know about the cultivation. However, it can be enhanced by the use of different technological resources, tool, and procedures. Predict the type of crop which one is suitable for that particular soil, weather condition, temperature and so on. So for, using machine learning with the set of data set are identified the crop for the corresponding soil.

*Keyword*-Data Mining, Soil Testing Agriculture, Analysis, Artificial Intelligence.

## I. INTRODUCTION

Data mining is a vital area of modern research world for processing, analyzing and evaluating large data sets; to identify associations, classifications, and clustering, etc. Between different attributes and predict the best results with relevant patterns[1][2].Significantly, these methods can be used in the field of agriculture and can produce extraordinary significant benefits and predictions that can be used for commercial and scientific purposes. Traditionally, Agriculture decision- making is based on experts' judgments and these judgments may not apply to classify the soil suitability and may lead the lower crop yield. The explicit data set management by the data mining techniques and algorithms contain the huge analytical potential for accurate and valid results and these can help to automate the classification process, depending on the predefined parameters developed by Agriculture research centers. Decision tree, Naïve Bayes algorithm, Rule- Based classification, Neural Networks, Support Vector Machine and Genetic Algorithm etc. are very well-known algorithm for data classification and further for knowledge discovery.

Traditionally, Agriculture decision-making is based on experts' judgments [3] and these judgments may not apply to classify the soil suitability and may lead the lower crop yield. Decision tree [4], Naïve Bayes algorithm [5][6][7], Rule-Based classification, Neural Networks[5][6], Support Vector Machine [8] and Genetic Algorithm [9] etc., are very well-known algorithm for data classification and further for knowledge discovery.

In this research, we intended to understand the related domain, analyzed the behavior of different data mining classification algorithms on the soil data set and evaluating the most predictive and accurate algorithm. The data set has been accumulated from different soil surveys that were conducted at numerous agricultural areas located in Tamil Nadu District and Andhra.

## II. PROBLEMSTATEMENT

The soil is highly important and subservient organism torun the ecosystem and the importance of soil in agriculture is understandable because that is the basic bedrock of the agricultural industry. In Pakistan, the soil characterization is a basic component and has the potential to increase the yield per acre, but unfortunately due to not having any appropriate technological resources that are difficult to distinguish and classify the soil so that the suitable crops can be grown at the right location. Moreover, there are many other factors that may be affected the soil quality parameters, for example, traditional cropping system, the application of fertilizers, and irrigation, etc. Therefore, it is highly important to maintain a system that can classify the soil in adequate quantities for best practices. The primary objectives of ours study are:

o To classify the soil under different agro ecological zones in Kasur district, Punjab, Pakistan by different classification algorithm available in data mining.

o To recommend the relevant crops depending on their classification.

o To evaluate the performance of predictive algorithms for better knowledge extraction.

## III. METHODS

The rapid growth of interest in data mining is due to the (i)falling cost of large storage devices and increasing ease of collecting data over networks (ii) development of robust and efficient machine learning algorithms to process this data, and (iii) falling cost of computational power, enabling use of computationally intensive method sf or data analysis.

Though, there are lots of techniques available in the data mining, few methodologies such as Artificial Neural Networks, K nearest neighbor, K means approach, are popular currently depends on the nature of the data.

***Artificial Neural Network:*** Artificial Neural Networks (ANN) is systems inspired by the research on human brain (Hammerstrom, 1993). Artificial Neural Networks (ANN) networks in which each node represents a neuron and each link represents the way two neurons interact. Each neuron performs very simple tasks, while the network representing of the work of all its neurons is able to perform the more complex task. A neural network is an interconnected set of input/output units where each connection has a weight associated with it. The network learns by fine-tuning the weights so as able to predict the call label of input samples during testing phase. Artificial neural network is a new technique used in flood forecast. The advantage of ANN approach in modeling the rain fall and run off relationship over the conventional techniques flood forecast. Neural network has several advantages over conventional method in computing. Any problem having more time for getting solution, ANN is highly suitable states that the neural network method successfully predicts the pest attack incidences for one week in advance. Pedo transfer functions(PTFs) provide an alternative by estimating soil parameters from more readily available soil data. The two common methods used to develop PTFs are multiple-linear regression method and ANN. Multiple linear regression and neural network model (feed-forward back propagation network) were employed to develop a pedo transfer function for predicting soil parameters using easily measurable characteristics of clay, sand, silt, SP, Bd and organic carbon. Artificial Neural Networks have been successful in the classification of other soil properties, such as dry land salinity (Spencer et al. 2004). Due to their ability to solve complex or noisy problems, Artificial Neural Networks are considered to be a suitable tool for a difficult problem such as the estimation of organic carbon in soil.

***Support Vector Machines:*** Support Vector Machines (SVM) is binary classifiers (Burges, 1998; Cortes and Vapnik,1995). SVM is able to classify data samples in two disjoint classes. The basic idea behind is classifying the sample data into linearly separable. Support Vector Machines (SVMs) are a set of related supervised learning methods used for classification and regression. In simple words given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.SVM is used to assess the spatiotemporal characteristics of the soil moisture products.

***Decision trees:*** The decision tree is one of the popular classification algorithms in current use in Data Mining and Machine Learning. Decision tree is a new field of machine learning which is involving the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees and discrimination nets or product ion rules. Application of data mining techniques on drought related data for drought risk management shows the success on Advanced Geospatial Decision Support System (GDSS). Leisa J Armstrong states that data mining approach is one of the approaches used for crop decision-making.

Research has been conducted in Australia to estimate a range of soil properties, including organic carbon (Hendersonetal. 2001). The nation-wide database had 11,483 soil points available to predict organic carbon in the soil. An enhanced decision trees tool (Cubist), catering for continuous outputs was used for this study. A correlation of up to 0.64 was obtained between the predicted and actual organic carbon levels.

***K nearest neighbor**:* K nearest neighbor techniques is one of the classification techniques in data mining. It does not have any learning phase because it uses the training set every time a classification performed. The Nearest Neighbor search(NN) also known as proximity search, similarity search or closest point search is an optimization problem for finding the closest points in metric spaces. K nearest neighbor is applied for simulating daily precipitation and other weather variables (Rajagopalan and Lall,1999).

***Bayesian networks*: **A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over fitting of data Development of a data mining application for agriculture based on Bayesian networks were studied by Huang et al. (2008). According to him, Bayesian network isa
powerful tool for dealing uncertainties and widely used inagriculture data sets. He developed the model for agriculture application based on the Bayesian network learning method.The results indicate that Bayesian Networks are a feasible and efficient. Support Vector Machines Support Vector Machines

## IV. RELATEDWORK

### *A. History of agricultural systems*

Agricultural system's science generates knowledge that allows researchers to consider complex problems or take informed agricultural decisions. Modeling, an essential tool in agricultural system's science, has been accomplished by sci- entists from a wide range of disciplines, who have contributed concepts and tools over more than six decades. As agricultural scientists now consider the" next generation" models, data, and knowledge products needed to meet the increasingly complex systems problems faced by society, it is important to take stock of this history and its lessons to ensure that avoid re- invention and strive to consider all dimensions of associated challenges. To this end, we summarize here the history of agricultural systems modeling and identify lessons learned that can help guide the design and development of next generation of agricultural system tools and methods.

Recent trends in broader collaboration across institutions, across disciplines, and between the public and private sectors suggest that the stage is set for the major advances in agricultural system's science that are needed for the next generation of models, databases, knowledge products and decision support systems.

### *B. Agricultural Systems on global food production and consumption*

Over the next decade's mankind will demand more food from fewer land and water resources. This study quantifies the food production impacts of four alternative development scenarios from the Millennium Ecosystem Assessment and the Special Report on Emission Scenarios. Partially and jointly considered are land and water supply impacts from population growth, and technical change, as well as forest and agricultural commodity demand shifts from population growth and economic development. The income impacts on food demand are computed with dynamic elasticity's. Simulations with a global, partial equilibrium model of the agricultural and forest sectors show that per capita food levels increase in all examined development scenarios with minor impacts on food prices.

Global agricultural land increases by up to 14% between 2010 and 2030. Deforestation restrictions strongly impact the price of land and water resources but have little consequences for the global level of food production and food prices. While projected income changes have the highest partial impact on per capita food consumption levels, population growth leads to the highest increase in total food production. The impact of technical change is amplified or mitigated by adaptations of land management intensities.

### *C. Predicting farmer uptake of new agricultural practices: A tool for research, extension and policy, Agricultural systems.*

There is much existing knowledge about the factors that influence adoption of new practices in agriculture but few attempts have been made to construct predictive quantitative models of adoption for use by those planning agricultural research, development, extension and policy. ADOPT (Adoption and Diffusion Outcome Prediction Tool) is the result of such an attempt, providing predictions of a practice's likely rate and peak level of adoption as well as estimating the importance of various factors influencing adoption. It employs a conceptual framework that incorporates a range of variables, including variables related to economics, risk, environmental outcomes, and farmer networks, characteristics of the farm and the farmer, and the ease and convenience of the new practice. The ability to learn about the relative advantage of the practice, as influenced by characteristics of both the practice and the potential adopters ,plays a central role.

ADOPT provides a prediction of the diffusion curve of the practice and sensitivity analyses of the factors influencing the speed and peak level of adoption. In this paper the model is described and its ability to predict the diffusion of agricultural practices is demonstrated using examples of new crop types, new cropping technology and grazing options. As well as providing predictions, ADOPT is designed to increase the conceptual understanding and consideration of the adoption process by those involved in agricultural research, development, extension and policy.

## V. SYSTEMDESGIN

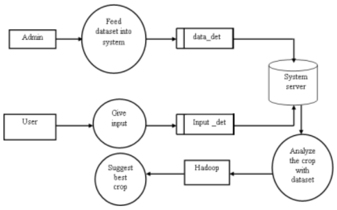


Fig. 1. Architecture Diagram

### *A. USER INTERFACEDESIGN*

User interface design, in this module user will the soil type on UI. To develop our application we use net beans as an IDE and MYSQL as a back end. All inputs and output will put and get through this IDE only. Creating a user registration to get their information. After registration user will get login id. In this two different login one for admin and another for user. Because admin can only add the crop details initially on the training set.

### *B. MAINTAINING TRAININGDATASET*

The Server will monitor the entire data set (set of crop) information in their database and verify them if required. Also, the Server has to establish the connection to communicate with the Users. The Server will update each soil and input details into its database. The soil and crop data sets are the main input of the user. Based on that system will compare and predict the best crop to the user. The Server will store the entire information in their database. Admin will feed location, weather condition of that location; water source of that location, crop type etc., and those details will be stored on training set.

### *C. SOILESTIMATION*

In this module, have to analyze the soil types. Soil type usually refers to the different sizes of mineral particles in a particular sample. Soil is made up in part of finely ground rock. Hard surface of base is called hard strata soil particles, grouped according to size as sand and silt in addition to clay, organic material such as decomposed plant matter. We have to feed different types of soil and their features on dataset.

### *D. WATER SOURCE AND WEATHERANALYSIS*

Here have to gather the information about water and temperature land of particular area. Because based on weather and water facility only the best crop will cultivate. So the source of weather that land is depending on well or rain fall. Through this can easily predict the crop type.

### *E. BEST CROPRECOMMENDATION*

In this module system will compare the new input with existing training set data. Here it will generate a new set output for given input. User will get the output based on input. If user gives soil as input they will output as type of crop which is to be cultivated on that land. If they give crop name, output will become soil as an output. Output will be like, in which sand those crops will cultivate.

## VI. RESEARCHMETHODOLOGIES

### *A. UNDERSTANDING OFDATASET*

Importantly, the use of Information Technology is now the basic part of our lives and is increasing day by day in almost every industry to accomplish important tasks in every business organization. Emerging technologies have a greater impact on our lives in different ways.

Technology is being implemented in almost every section of our lives and business structures. Specifically, in agriculture new applications, technologies and methods are developed to get the efficient results; to cut down the time and to increase the crop productivity. But in agriculture, the collection of such big-data is not an easy task. Not having any computerized system is making it worse in Pakistan and in the past decades, expert opinions were taken into consideration to identify the soil properties and recommendation for crops and the better fertility.

In this research, the soil samples that are being used were collected from different fields and the surrounding of Tamil Nadu district and Andra. We have acquired the test center data from Soil Fertility Department, Tamil Nadu in the form of unstructured and manual format. The data was collected by surveying different locations on different dates and containing the test samples of soilf or different properties. After the acquisition, the digitization of record has been made to convert data into the structured format for further processing. This digitized dataset included different attributes that are defined here under: Soil-Basic, included the basic information, specifically the village and district name and the date of the sample was collected; this information will be a good resource for getting the date wise soil properties and the change in properties for different seasons.

Soil-Basic {GridID, VillageID, Sample Date, Village Name, DistrictName}

Soil-Location, contained the GPS (Global Positioning System) record; Zone Number and GridID for the unique identification of locality. ELocation and N-Location are used to identify the spatial coordinates of East and North physical location of the respective field.

Soil-Location {GridID, ZoneNumber, ELocation,N-Location}

Soil-Analytical, This part of the dataset is the most important and essential part of our research based on soil properties. The soil consists of different physical(i.e., texture, weight and density, etc.), chemical(Organic and inorganic matter, i.e. magnesium (MN),copper (Cu), zinc (Zn), Phosphorous (P), Potassium(K), Iron (Fe) etc.) And biological properties(microbial and faunal activities in the soil) these characteristics describe the productivity and fertility ratios of higher yield in crops. [12] However, thereare many other important factors, i.e. pH level, Soil Electricity Conductivity (EC) and temperature of the soil also have significant importance. These properties are the part of our dataset as well. The major attributes arehereunder:

Soil-Analytical {GridID, VillageID, pH, OM,AvgK(ppm), EC(uS/cm),Zn(ppm),Fe(ppm),Mn(ppm),Cu(ppm),texture}

In the first phase of our research, we have understood the soil dataset, this dataset included more than 800 instances of soil samples from different regions in Tamil Nadu and Andra. In order to optimized prediction, we have to clean and prune the dataset as the preprocessing and selection has the greater influence on the computational efficiency and predictive accuracy. Incomplete and inconsistent information have the significant impact on analysis and may lead the worst prediction.

### *B. PREPROCESSING OFDATASET*

Data preprocessing has the significant and substantial role in data mining tasks for better results. Physiognomies of soil data sets may include multiple noisy, incomplete, inconsistent and irrelevant features that should be addressed. Importantly, the selection of proper dataset for classification has the considerable impact on prediction accuracy. Waikato Environment for Knowledge Analysis (WEKA) is an open source machine learning to tool, consists of different data mining algorithm including, classifying data algorithms. In this research, we are using WEKA for data mining functions and methodologies to extract and construct the rules. More than 1000 data entries are collected for this research. These entries are then converted to the ARFF format, a suitable format for WEKA. The filters available for preprocessing in WEKA i.e. Remove R- 1, Replace Missing Values and Discretize has helped to convert this data into coherent and noise free state. The resulted dataset consists of 760 data entries of different attributes as described above. Soil Dataset consists of different attributes that have complex relationships between dynamic variables. Therefore, before implementing any algorithm the soil distinguished properties must be encountered. We have split the dataset into two data sets, (i) training and (ii) test data. (i) Training data, 40% of the dataset (304 instances), will be used as the tuning and validation of our data model and this will formulate the association between the predictive lasses.(ii)Test data

(456 Instances) will be used for evaluating the strength of our classification model. Moreover, we have to reduce the number of parameters for construction of soil classification model for efficiency and accuracy. The Level of pH, Texture, Electricity Conductivity (EC) and average Potassium (K) will be used for this internment as the targeted considerations.

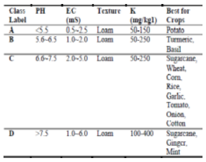
### C. CLASSIFICATION

Distinguished properties (i.e. PH level, organic and inorganic matter, texture and temperature, etc.) of soil have made the classification very critical and dynamic in nature. Therefore, we need a robust systematic categorization of soil with objectively efficient and effective algorithms and methods. Besides, the structural complexity a closer analysis is likely to lead to an improved prediction process that can be helpful in the future. Rule-Based classifiers, Bayesian Networks (BN),Decision Tree (DT), the Nearest Neighbor (NN), Artificial Neural Network (ANN), Support Vector Machine (SVM),Rough Sets, Fuzzy Logic, and Genetic Algorithms etc.

### D. RESULTS ANDANALYSIS

Specifically, in Weka, we can train data by different methods (i) Learning by examples included the nth-folds cross-validation scheme, by supplying test data or by giving the percentage of splitting. (ii) Lazy Learning which doesn't need any explicit learning model for classification also have significance, majority predictor and neural networks are the examples. (iii) Regression learning by giving numeric values to classifier is also a useful tool to plot the resultant points in linear,polynomial, single or multinomial logistic plane and the resultant outputs of these classifiers will predict the class of the specified instance.

The Table given below is the initial method of summarizing the large dataset on the basis of pH, EC,texture, and level of potassium required for different crops, so the relevant crop class can be predicted for different soil samples.



| Class Label | PH | EC (mS) | Texture | K (mg/kg1) | Best for Crops |
|---|---|---|---|---|---|
| A | <5.5 | 0.5~2.5 | Loam | 50-150 | Potato |
| B | 5.6~6.5 | 1.0~2.0 | Loam | 50-250 | Turmeric,<br>Basil |
| C | 6.6~7.5 | 2.0~5.0 | Loam | 50-250 | Sugarcane,<br>Wheat,<br>Corn,<br>Rice,<br>Garlic,<br>Tomato,<br>Onion,<br>Cotton |
| D | >7.5 | 1.0~6.0 | Loam | 100-400 | Sugarcane,<br>Ginger,<br>Mint |

Fig. 2. Soil Class Labels

Arc GIS is very well-known software for mapping and analyzing spatial data. For this study, we have used the ArcGIS tool to map the spatial data and this has helped us to visualize our results on a Thanjavur District map; figure(iii) is the visual representation of the NaïveBayes classification result.

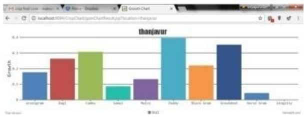


Fig. 3. Barchart for Thanjavur Crop

Thanjavur crop data is viewed with soil and growth of agriculture. This diagram shows that agriculture growth is mainly based on the soil and the crop which is used for agriculture.

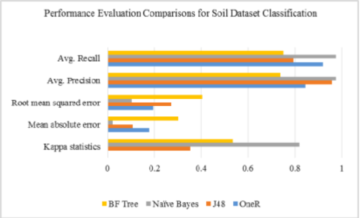


Fig. 4. Performance Evolution Comparisons

Finally, by the results of our experiment we can conclude that the Naïve Bayes classifier can predict the classified soil data for more accurate prediction as the comparison to other classifiers used in the experiment. The comparative analysis of selected classifiers is visualizing the best results of Naïve Bayes in both tables and graphical format figure (i) and these predictions will/can help the agriculturalist to find out the best crop fertilization with reference to the properties of soil without conducting traditional tests and only depending on expert opinions.

The above results can be mapped to remote sensing and Geographical Information System (GIS) Software for the better understanding of the classification. Fortunately, our soil data samples contain their Latitude and Longitude value in Soil-Location table.

## VII. CONCLUSION

Thus, the paper infer that using machine learning we implement a system to predict the crop and yield for that crop. Through this app farmers and normal people can get more advantages. The experiment was conducted on data instances from Thanjavur district and Andra.We have observed the comparative analysis of these algorithms have the different level of accuracy to determine the effectiveness and efficiency of predictions. However, the benefits of the better understanding of soils classes can improve the productivity in farming, reduce de- pendence on fertilizers and create better predictive rules for the recommendation of the increase in yield. In the future, we contrive to create a Soil Management and Recommendation System, which can be utilized effectively by agriculturist and laboratories for Soil Testing. This System will help to recommend a suitable fertilizer and predict for better yield.

## VIII. REFERENCE

1. Uwe A. Schneider a, ⇑ , Petr Havlik b, Erwin Schmid c, Hugo Valin b, Aline Mosnier b, c, Michael Obersteiner b, Hannes Bottcher b, Rastislav Skalsky´d, Juraj Balkovicˇ d, Timm Sauer a, Steffen Fritz b” Impacts of population growth, economic development, and technical change on global food production and consumption” Agricultural Systems 104(2011) 204–215inelsvier.

2. Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., &Al- Shawakfa, E. M. (2011). A comparison study betweendata mining tools over some classification methods. International Journal of Advanced Computer Science and Applications, 8(2),18-26.

3. Eiben, A. E., Raue, P. E., & Ruttkay, Z. (1994, October). Genetic algorithms with multi-parent recombination. In International Conference on Parallel Problem Solving from Nature (pp. 78-87). Springer, Berlin,Heidelberg.

4. JamesW.Jonesa,âĄŐ,JohnM.Antleb,BrunoO.Basso c, Kenneth J. Boote a, Richard T. Conant d, Ian Foster e, H. Charles J. Godfray f, Mario Herrero g, Richard E. Howitt h, Sander Jansseni, Brian A. Keating g, Rafael Munoz-Carpena a, Cheryl H. Porter a, Cynthia Rosenzweig j, Tim R.Wheeler k “Brief history of agricultural systems modeling” in science direct.

5. GeoffKuehnea,RickLlewellyna,âĄŐ,DavidJ.Pannellb, Roger Wilkinsonc, Perry Dollingd, Jackie Ouzmana, Mike Ewinge, ” Predicting farmer uptake of new agricultural practices:Atoolforresearch,extensionandpolicy”inElsevier sciencedirect.

6. Zhou,S.,Ling,T.W.,Guan,J.,Hu,J.,&Zhou, A. (2003, March). Fast text classification: a training-corpus pruning based approach. In Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International Conference on (pp. 127-136). IEEE.

7. Li, Y., & Bontcheva, K. (2008). Adapting support vector machines for f-term-based classification of patents. ACM Transactions on Asian Language InformationProcessing (TALIP), 7(2),7.

8. Tubiello, F. N., Salvatore, M., Cóndor Golec, R. D., Ferrara, A., Rossi, S., Biancalani, R., ... & Flammini, A. (2014). Agriculture, forestry and other land use emissions by sourcesandremovalsbysinks.Rome,Italy..

9. Agriculture Statistics of Pakistan, Pakistan Bureau of Statistical,Retrieved10September2016byhttp://www.pbs .gov.pk/content/agriculture-statistics.

10. Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. European Journal ofOperationalResearch,173(3),781800.

11. Larose, D. T. (2014). Discovering knowledge in data: an introduction to data mining.JohnWiley&Sons..

12. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. From data mining to knowledge discovery in databases. AI magazine, 17(3),37.

13. Doran, J. W., & Parkin, T. B. (1994). Defining and assessing soil quality. Defining soil quality for a sustainable environment, (definingsoilquality),1-21.

14.Kumar, A., & Kannathasan, N. (2011). A survey on data mining and pattern recognition techniques for soil data mining. IJCSI International Journal of Computer Science Issues, 8(3),1694-0814.

15.Ramesh, V., and Ramar, K. (2011). Classification of agricultural land soils: a data mining approach. Agricultural Journal, 6(3),82-86.

16. Nevill-Manning,C.G.,Holmes,G.,andWitten,I. H. (1995,November). The development of Holte’s 1R classifier. In Artificial Neural Networks and Expert Systems, 1995.Proceedings., Second New Zealand International Two-Stream Conference on (pp. 239-242). IEEE.