IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Classification Techniques in Analysis of Salem District Soil condition for Cultivation of Sunflower

N. Hemageetha[1*], N. Nagalakshmi[2]

[1]Department of Computer Science, Govt. Arts College for Women, Salem, Tamil Nadu, India
[2]Department of Computer Science, Govt. Arts College for Women, Salem, Tamil Nadu, India

[*]*Corresponding Author:geekani2010@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 18/Aug/2017, Published: 31/Aug/2017

***Abstract*—** Agriculture is the backbone of Indian economy. Sunflower is one of the most important oil seed crops grown in temperate countries and India is one of the largest producers of oil seeds in the world. The farmers can determine which type of crops to be cultivated in a particular place with the help of the soil condition analysis. The valuable knowledge is extracted from the agricultural data set with the help of data mining techniques. The farmers can make use of the technology and the right techniques; they can make agriculture a profitable enterprise. Salem district is one of the largest districts in Tamil Nadu, India and it is famous for mango cultivation This paper analyzes whether the Salem district soil is suitable for the cultivation of sunflower crop with the help of data mining classification techniques.

***Keywords*—** Agriculture Soil, Bayes Net, Random Forest, J48

I. INTRODUCTION

Interesting patterns and knowledge are extracted from large amount of data is called as Data Mining. Data Mining in agriculture is a very interesting field. Plenty of researches are going with base of Data mining [1][2][3]. The main aim of this paper is to investigate data mining techniques, which are suitable for solving the problems in the agricultural sector. Salem district is the largest district in Tamil Nadu, India. It is surrounded by Shervroy hills and Kalvarayan hills, Salem is famous for mango cultivation, silver ornaments, textile, sago industries and steel production [4]. Corn, maize, paddy and ground nut are the major crops grown in the Salem district. Other crops like sugarcane, turmeric, chrysanthemums are also grown in this district [4].

Sunflower oil is considered a premium one and is preferred the world over due the health factor. Andhra Pradesh, Maharashtra, Bihar and Orissa are major sunflower producing states in India. Sunflower cultivation is getting more and more attractive due to the low cost of cultivation high revenue generation because the sunflower seeds are in continuous demand by the edible oil industry. This paper analyzes, with the help of data mining classification techniques based on the soil dataset and find out whether Salem district soil is suitable for sunflower cultivation.

Section II presents related work, Section III presents a few Data mining Classification algorithms; Section IV presents results and discussions finally Section V presents Conclusion and Future Scope.

II. RELATED WORK

The previous literatures about the related work were collected. Soil is one of the important natural resource for the cultivation of the crops. There are several nutrients are presents in the soil in the form of organic and mineral. Soil testing is an important component of nutrient management in agriculture. These nutrients are classified as macro and micro nutrients. Nitrogen (N), Phosphorus (P), Potassium (K) are macro nutrients, they are very important for the growth of the plants. The pH value of the soil will affect the availability of macronutrients and micronutrients [5].

pH value is used to find out whether the soil is acidic or alkaline. So it acts as a limiting factor of the growth of the plant. The nutrient rating of pH, EC, N, P and K [5] for plants growth are given in the table 1.

The result of [6] shows that pH level is high, EC level is low in salem district. The Nitrogen level is low and Phosphorus level is high to medium and potassium level is high to medium in Salem district which is given in the paper [7]. Coming to sunflower cultivation, Sunflower is used for

making oil, paint and cosmetic. America is the origin of the sunflower.

Table 1: Nutrient rating of pH, EC, N, P and K.

Parameter	Low	Medium	High
Ph	<6.5	6.5 - 7.5	>7.5
Ec(ds m^{-1})	<1.0	1.0 -3.0	>3.0
N(Kg ha^{-1})	<280	280-450	>450
P(Kg ha^{-1})	<11	11-22	>22
K(Kg ha^{-1})	<118	118-280	>280

In[8] the authors analyzed and found that the soil in Thanjavur district is suitable for the sunflower cultivation. The required range of nutrients for the growth of the sunflower is given in the Table 2 [8]. This paper analyzes whether the Salem district soil is suitable for sunflower cultivation using data mining classification techniques based on the Table 2.

Table 2 : Required Range of Sunflower cultivation

Parameter	Required Range
pH	6.0 – 7.2
Ec(ds m^{-1})	< 4.8
N(Kg /acre)	16 – 36
P(Kg/acre)	40 – 60
K(Kg/acre)	10 – 16

III. AGRICULTURE DATA MINING CLASSIFICATION TECHNIQUES METHODOLOGY

There are several classifiers are available for analyse the soil dataset. The best classifier is selected based on the ranker tester option in WEKA tool. For this work the whole data set is given as the input for the option and several classifiers are selected for analysis. The ranker tester screen shot of WEKA is shown in Figure 1.

The ranker tester has given the experimental accuracy for various selected classifiers. This is shown in the Table 3. Out of five classifiers the J48, Naïve Bayes and Random forest are having high accuracy, so for this work the first three classifiers are selected for the data analysis

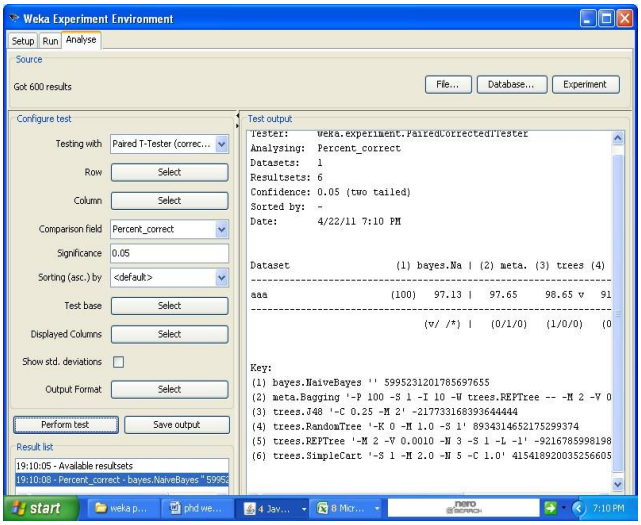


Figure 1. The screen shot of WEKA ranker tester

. Table 3: Results of ranker tester

S.No.	Classifier Tool	Experimenter Accuracy (%)
1	J48	99.97
2	Random Forest	97.02
3	Naïve Bayes	89.43
4	Jrip	89.06
5	ZeroR	49.36

Naïve Bayes, J48 and Random forest classifiers are discussed below

A. J48 (C4.5)

J48 is one of the Decision tree classification algorithm. Java implementation of C4.5 algorithm is J48. It can handle the missing attributes. The decision trees are generated from a set of labelled training dataset. It uses the fact that each data attribute can be used to make the decision by splitting into small subsets

B. Random forest

The Random Forest algorithm is mainly used to classify large amount of data with high accuracy. Its main concept is grouping up the "weak learner" to form a "strong learner". Number of decision trees is constructed randomly during the training time. Finding the root node and for splitting the feature node is based on random concept instead of using information gain. Combinations of tree predictors form the random forest. It calculates votes for each predicted target and considers the high votes predicted target as the final prediction.

C. *Naïve Bayes*

Bayes theorem is developed by the British minister Bayes [8]. Independent attributes and dependent attributes are the two types of attributes in Bayes theorem. The Naïve Bayes theorem is

$$P(Y_j \mid X) = [P(X \mid Y_j)\, P(Y_j)] / P(X)$$

$P(Y_j \mid X)$ is the probability of the object X belonging to class Y_j ,$P(X \mid Y_j)$ is the probability of obtaining attribute values X if we know that it belongs to class Y_j. $P(Y_j)$ is the probability of any object belonging to classY_j ,$P(X)$ is the probability of obtaining the attribute values of X .

IV. RESULTS AND DISCUSSION

A. *Data Collection*

Soil Dataset was collected from Krishi Vigyan Kendra, Tamil Nadu Agricultural University, Santhiyur, Salem. Soil samples are taken from 11 blocks in Salem District. Dataset has the attributes like Sample no., Block no, PH value, Electric conductivity(EC), Organic Carbon(OC), Phosphorous (P), Potassium(K),Nitrogen (N) [9][10].

B. *Data Formatting*

WEKA is a machine learning algorithm and data pre-processing data mining tool. It is a open source software. The collected data set is in the form of look up table. So the data are formatted into an Excel format based on the Blocks. All excel sheets are converted into a single sheet, which are converted into CSV format, because in WEKA the standard format is CSV format. The process flow of the format conversion is given in Figure 2.

.

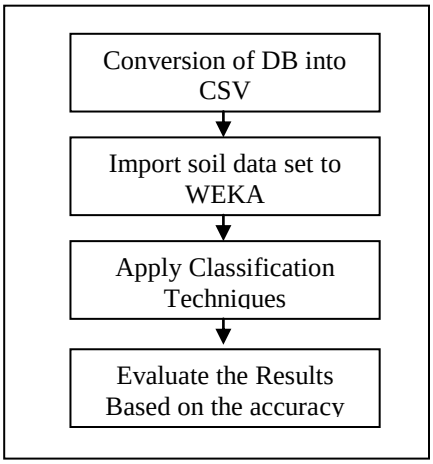
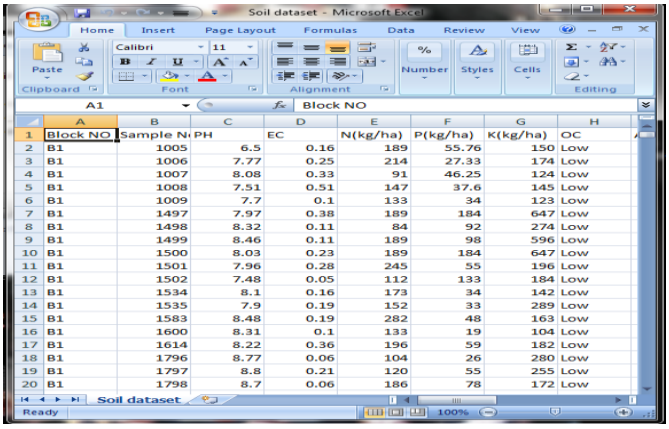


Figure 2. Process Flow

From the collected dataset, out of 792 instances of soil dataset, 701 instances have been considered for this work. Instances with missing attribute values and noisy data are filtered with the help of WEKA Filters. The filtered proposed Soil data set is shown in the Figure 3.

Tuning of the parameters is very important one for increasing the accuracy of the classifiers. For this work, Value 0.25 is consider for confidential threshold, value 3 for folds of pruning and value 1 for the number of seeds are selected for J48 classifier. In random forest the value 100 is taken for the number of trees generated. In Naïve Bayes, the K2 algorithm is selected as a searching algorithm and AD tree as false.



Block NO	Sample No	PH	EC	N(kg/ha)	P(kg/ha)	K(kg/ha)	OC
B1	1005	6.5	0.16	189	55.76	150	Low
B1	1006	7.77	0.25	214	27.33	174	Low
B1	1007	8.08	0.33	91	46.25	124	Low
B1	1008	7.51	0.51	147	37.6	145	Low
B1	1009	7.7	0.1	133	34	123	Low
B1	1497	7.97	0.38	189	184	647	Low
B1	1498	8.32	0.11	84	92	274	Low
B1	1499	8.46	0.11	189	98	596	Low
B1	1500	8.03	0.23	189	184	647	Low
B1	1501	7.96	0.28	245	55	196	Low
B1	1502	7.48	0.05	112	133	184	Low
B1	1534	8.1	0.16	173	34	142	Low
B1	1535	7.9	0.19	152	33	289	Low
B1	1583	8.48	0.19	282	48	163	Low
B1	1600	8.31	0.1	133	19	104	Low
B1	1614	8.22	0.36	196	59	182	Low
B1	1796	8.77	0.06	104	26	280	Low
B1	1797	8.8	0.21	120	55	255	Low
B1	1798	8.7	0.06	186	78	172	Low

Figure 3. Proposed Salem District Soil Dataset

The 10-fold cross validation is used for all the classifiers. In 10-fold cross validation the whole data set is divided into 10 sub datasets. In the 1st iteration the first data set is act as a testing data set, remaining 10-1st subsets are act as training dataset. In the 2nd iteration the 2nd dataset is act as test data and remaining 10-2nd subsets act as training data. This process is repeated for 10 times, so all the sub dataset are used for both training and testing purpose. This will increase the accuracy of the classifiers. The screen shot of WEKA J48 classifier is given in the figure 4.

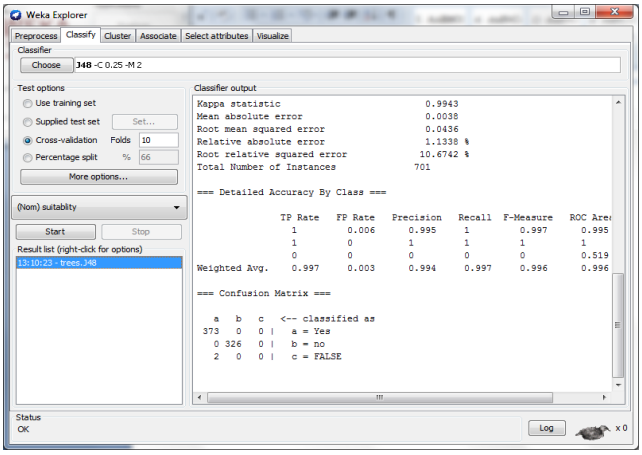


Figure 4. WEKA- J48 classifier Screen shot for soil condition

The performances of the classifiers are evaluated based on the confusion matrix, ROC, kappa statistics and the Mean

absolute Error. The confusion matrix of J48 classifier is shown in the figure 4. The instances correctly classified and incorrectly classified of all the three classifiers are given in the figure 5.

J48 classifies the dataset with 99.71% accuracy and the true positive values 372, 326 and 2 for suitable, non suitable and incorrect class. The J48 classifier's kappa statistics is 0.9943, out of 701 tuples, 699 tuples are correctly classified. The mean average error is 0.038, the ROC area is 0.995. The Random forest classifies the dataset with 99.71% accuracy and the true positive values 372, 326 and 2 for suitable, non suitable and incorrect class. The kappa statistics is 0.9943; out of 701 tuples, 699 tuples are correctly classified. The mean average error is 0.005, the ROC area is 0.999.

The Naïve Bayes classifies the dataset with 92.58% accuracy and the true positive values 337,312 and 52 for suitable, non suitable and incorrect class. The kappa statistics is 0.8524, out of 701 tuples 649 are correctly classified. The mean average error is 0.106, the ROC area is 0.96. The accuracy and Mean absolute Error Rates of all the classifiers are compared. The Correctly classified instances, Incorrectly classified instances, Accuracy and Mean Square Error are given in the Table 4.

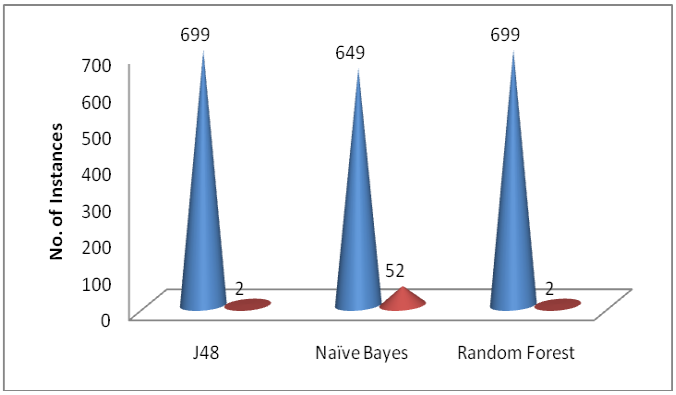


Figure 5. Instances are Correctly or incorrectly classified

Table 4: Comparison of different classifiers results

S. No	Classifiers	Classified	Time taken tbec)	Accuracy %	MAE	Kappa Statistics	ROC Area
1.	J48	699	0.05	99.71	0.003	0.9943	0.995
2	Naive Bayes	649	0.03	92.58	0.106	0.8524	0.96
3.	RF	699	0.04	99.71	0.005	0.9943	0.999

The results shows that the accuracy of J48 and Random Forest are high compared with Naive Bayes. But the Mean absolute error rate and time taken of random forest are less compared with other classifiers. The confusion matrix of all the classifiers shows that the major part of the Salem district soil is suitable for cultivation of sunflower crop. This will help the farmers to cultivate the sunflower crop in their field in Salem District.

V. Conclusion and Future Scope

Data mining in agriculture is a novel research field, it will help the farmers to improve their crop productivity and get more profit. This paper analysed and find out whether the Salem district soil condition is suitable for the sunflower cultivation. The experimental results of the three classifiers are compared and found that the accuracy of Random Forest classifier is high compared with Naive bayes and J48. The confusion matrix shows that the major part of the Salem district soil is suitable for cultivation of sunflowers. This finding will help the Salem District farmers to cultivate the sunflower crop in their field in and get good profit.

Acknowledgment

The author gratefully acknowledges the support of Dr. Dr.N.Sriram, Programme Coordinator Krishi Vigyan Kendra (Farm Science Centre) Tamil Nadu Agricultural University Santhiyur, Salem,

References

[1] Mucherino.A, Petraq Papajorgji and P.M.Pardalos, "A survey of data mining techniques applied to agriculture". Published online **2009** © Springer-verlag.

[2] G.M. Nasira , N. Hemageetha, "Vegetable price prediction using data mining classification technique" Preceedings of the International Conference on pattern Recognition, Informatics and Medical Engineering (PRIME 2012), PP. **99-102** ISBN No:978-1-4673-1038-3. © **2012** IEEE.

[3] N.Hemageetha, Dr. G.M. Nasira ," Vegetable Price Prediction using Adaptive Neuro-Fuzzy Inference System", International Journal of Computer Sciences and Engineering IJCSE E- ISSN: 2347-2693 Vol-4 Issue -3 June 2017, pp 75-79.

[4] R.Santhi et at (**2014**) ,GIS based Soil map for salem district of Tamilnadu. TechInical Folder, TNAU,Coimbatore.

[5] Natesan et at(**2007**),. Technical Bulletinon "Soil test crop response based fertilizer prescription for different soils and crops in tamil nadu" ,AICRP-STCR TamilNadu Agricultural University, Coimbatore.

[6] N. Hemageetha, Dr. G.M. Nasira Analysis of Soil condition Based on pH value Using Classification Techniques IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 50-54

[7] N. Hemageetha ,Dr.G.M. Nasira , "Availability of Macro Nutrients Status in Salem District Soil using DataMining Classification Techniques **"**International Journal Of Control Theory And Applications ISSN: 0974-5572 9(40),2016, pp:57-66

[8] Durga karthik , K.vijayarekha, Simple and quick classification of soil for sunflower cultivation using data mining algorithm , International journal of chemtech research. Vol .7,No.6.,PP 2601-2605 2014-15.

[9] N.Hemageetha, Dr. G.M. Nasira , "Analysis of the Soil data Using Classification Techniques for Agricultural Purpose, International Journal of Computer Sciences and Engineering IJCSE E-ISSN: 2347-2693 Vol-4 Issue -6 June 2016

[10] N. Hemageetha ,Dr.G.M. Nasira , "Classification of Soil Type in Salem District using J48 Algorithm, "International Journal Of Control Theory And Applications ISSN: 0974-5572 9(40),2016 pp 33-41

Authors Profile

Dr. N. Hemageetha. is woking as Associate Professor, Department of Computer Science at Government Arts College for Women, Salem,Tamil Nadu. She has around 21 years of experience in teaching, at college level in various positions like Lecturer, Assistant Professor and Associate Professor. She has published so far more than 15 research papers in referred journals and conferences. Received her MCA degree from University of Madras, Chennai, the M.Phil degree in computer Science from Bharathidasan University, Trichy and PhD degree from Periyar University,Salem.

.

N. Nagalakshmi. is woking in Department of Computer Science at Government Arts College for Women, Salem,Tamil Nadu. She has around 10 years of experience in teaching at college. Received her M.Sc degree from University of Madras, and M.Phil degree in computer Science from Bharathidasan University, Trichy

.

