International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 3, Issue 2, March – April 2014 ISSN 2278-6856

Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District

Vrushali Bhuyar

Assistant Professor, Dept. of MCA, Marathwada Institute of Technology,
Aurangabad, Maharashtra, INDIA

Abstract: ***Indian economy is highly depending on agriculture. Agriculture is the main source of income for most of the population. So farmers are always curious about yield prediction. To increase yield production many factors are responsible like soil, weather, rain, fertilizers and pesticides. Now a days Data mining plays an important role in in agriculture. The large amounts of data that is available with agriculture universities are mainly restricted to labs and research centers. There is a need to transform this huge data into technologies and make them available to the farmers. It can be possible with data mining. This huge amount of data can be utilized to mine nuggets of knowledge that can be useful for farmers and decision makers to take efficient, effective and prompt decision. In this paper one of the parameter which is used to increase yield production is considered; that is soil. Different classification algorithms are are applied to soil data set to predict its fertility. This paper focuses on classification of soil fertility rate using J48, Naïve Bayes, and Random forest algorithm. J48 algorithm gives better result than other algorithms. Decision tree form by J48 algorithm helps the farmer and decision makers to identify the the soil fertility rate and on the basis of nutrients found in the soil sample different fertilizers can be recommended.***

Keywords: Data Mining, J48, Naïve Bayesian, Random forest, Soil fertility.

1. INTRODUCTION

Enhancement of food technology is today's need. India is living in the era of huge population wherein the ratio and proportion of food and humans has no toning, resulting in high rates of inflation. Agriculture is totally dependent on the soil quality but as time passes more and more agricultural production results in the loss of nutrients present in the soil. We require identifying techniques that will slow down this elimination of nutrients and also will return the required nutrients with the soil, so that we keep getting high quality and good quantity crop productions. [1]

In agriculture, good soil health means capability of soil to posses physical, chemical and biological activities for consistent productivity of high crop yield. Good quality of soil assures us for retention and release of water and nutrients, enhancement and consistent root growth whilst maintaining biotic environment, providing the expected result and resist filth. Most of the agricultural soil in India is deficient in primary nutrients (Nitrogen, Phosphorous and Potassium).[2] Therefore the researchers are always thinking, about how to increase productivity of crops. The more the production of food material, the cheaper will be the cost of food products. The ultimate aim of any technology with respect to agriculture is to make the food production cheap and at the same time to give farmers many immediate and sustainable benefits [3]. Therefore there is need to transform huge amount of data that are available in lab and agriculture university into information. This can be possible with data mining.

Data mining is a process of discovering previously unknown patterns that are used for strategic decision making. [4] . There are different steps under this process such as

- ➢ Data Collection
- ➢ Data cleaning
- ➢ Data transformation
- ➢ Applying Data mining algorithms
- ➢ Model construction and pattern evaluation
- ➢ Knowledge gain used for decision making

The proposed work investigates application of J48, Naïve Bayes and Random Forest Classifier and compares these algorithms performance based on fertility index.

This paper tried to bring into picture the work done by different researchers with reference to enhance agricultural yields using data mining.

S. S. Bhaskar et al. [5] made a comparative study for soil classification of naïve bayes, JRip and J48. They found J48 to be the best method. They also used regression technique like linear regression and least square Median. They found least median squares regression produce better results for prediction than the classical linear regression technique.

Ravindra M et al. [6] uses decision tree in selecting the best suited pump for irrigation. D Ramesh et al. [7] uses Multiple Linear Regression for predicting rice yield.

S. Veenadhari et al. [8] uses decision tree induction technique to analyze the influence of climatic parameter on soybean productivity. Georg Ruß [9] evaluates four regression techniques on agriculture data. He found support vector regression can serve as a better model for yield prediction among MLP, RBF and RegTree.

Jay Gholap [10] uses J48 algorithm for predicting soil fertility class. Also for performance tuning of J48 algorithm he uses attribute selection and boosting techniques.

Suman et al. [11] cluster the data using the K-means Clustering on soil dataset then the linear regression is applied to classify the clusters.

P. Revathi et al. [12] uses Naive bayes, j48, MLP, Random forest, Random tree, they found J48 gives better result over the other algorithms for cotton seed quality.

D Ramesh et al. [13] uses MLP and K- Mean for yield prediction for East Godavari district of Andhra Pradesh.

This paper is organized into different sections. Section II includes Classification Algorithms. Section III includes experimental results using WEKA. And section IV conclusion.

2. CLASSIFICATION ALGORITHMS

Following section explain classification algorithms like J48 decision tree classifier, Naïve Bayesian classifier, Random forest.

J48 Decision tree Classifier: - J48 is extension of C4.5 classifier. It is supervised learning. A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node. During tree construction, attribute selection measures like information gain, gain ratio, gini index are used to select the attribute that best partitions the tuples into distinct classes. Model generated by decision tree helps to predict new instances of data. [4]

Naïve Bayesian classifier: - A Naive Bayesian classifier is a simple probabilistic classifier based on Bayesian theorem with strong (naive) independence assumptions. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. It uses prior probability of each category given no information about an item. Categorization produces a posterior probability distribution over the possible categories given a description of an item.[14] Bayesian formula can be written as

$$P\,(C_i / D1, D2\ldots.. D_n) = \frac{P\left(\frac{D_1 D2 \ldots\ldots Dn}{Ci}\right) P(Ci)}{P(D1, D2 \ldots\ldots Dn)}$$

Random forest: - The Random Forests algorithm is able to classify large amounts of data with accuracy. Random Forests are an ensemble learning method for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. [15]

Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers.

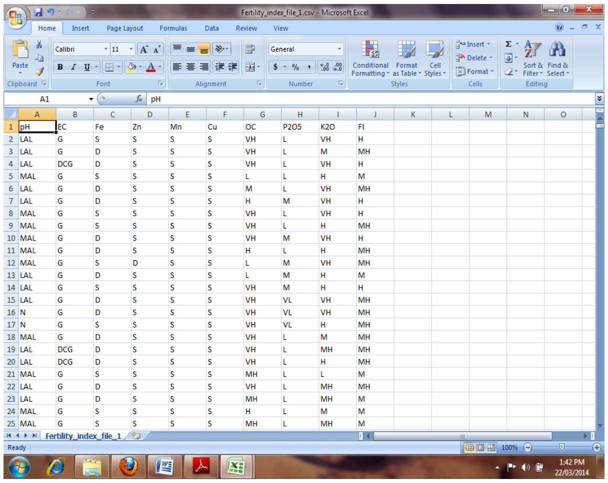
Random Forests grows many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on this m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

3. EXPERIMENT AND RESULTS

In this research, Soil dataset taken from Soil testing Laboratory, Agricultural Engineering Department, MIT Aurangabad which is recognized by department of Agriculture, Government of Maharashtra. Said laboratory prepared dataset from the project "Village wised Land fertility determination program-2010" under National agriculture development scheme funded by government of India. Dataset consist of soil data of 27 villages of Aurangabad District. Policy adopted for sample collection was 10% of the total cultivated land was taken into consideration and for every 10 hectare: 1 sample. Dataset having 10 attributes as PH – PH value of soil, EC – Electric conductivity, Fe - Iron , Cu - Copper , Zn- Zinc , OC- Organic Carbon, P_2O_5 – Phosphorous oxide , K_2O- Potassium Oxide, FI – Fertility Index. Soil Dataset consists of total 1639 instances. Fertility index is class label which categorized as VL- Very low, L- Low, M- Medium, MH- Medium High, H- High, and VH- Very High.

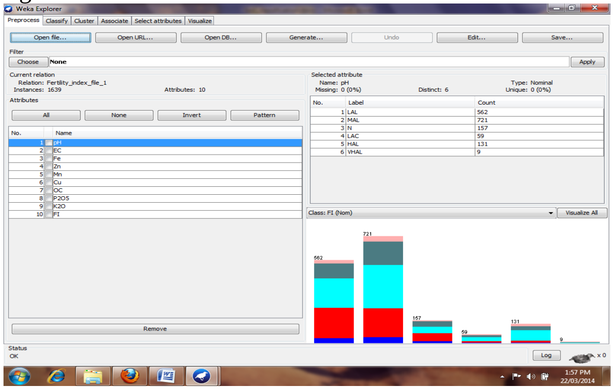
WEKA 3-6-9(Waikato Environment for Knowledge Analysis) open source data mining tool is used for experiment. This dataset prepared in Excel sheet with .CSV extension as shown in figure 1.



	A	B	C	D	E	F	G	H	I	J
1	pH	EC	Fe	Zn	Mn	Cu	OC	P2O5	K2O	FI
2	LAL	G	S	S	S	S	VH	L	VH	H
3	LAL	G	D	S	S	S	VH	L	M	MH
4	LAL	DCG	D	S	S	S	VH	L	VH	H
5	MAL	G	S	S	S	S	L	L	H	M
6	LAL	G	D	S	S	S	M	L	VH	MH
7	LAL	G	D	S	S	S	H	M	VH	H
8	MAL	G	S	S	S	S	VH	L	VH	H
9	MAL	G	S	S	S	S	VH	L	H	MH
10	MAL	G	D	S	S	S	VH	M	VH	H
11	MAL	G	D	S	S	S	H	L	H	MH
12	MAL	G	S	D	S	S	L	M	VH	MH
13	LAL	G	D	S	S	S	L	M	H	M
14	LAL	G	S	S	S	S	VH	M	H	H
15	LAL	G	D	S	S	S	VH	VL	VH	MH
16	N	G	D	S	S	S	VH	VL	VH	MH
17	N	G	S	S	S	S	VH	VL	H	MH
18	MAL	G	D	S	S	S	VH	L	M	MH
19	LAL	DCG	D	S	S	S	VH	L	MH	MH
20	LAL	DCG	D	S	S	S	VH	L	H	MH
21	MAL	G	S	S	S	S	MH	L	L	M
22	LAL	G	D	S	S	S	VH	L	MH	MH
23	LAL	G	D	S	S	S	MH	L	MH	M
24	MAL	G	S	S	S	S	H	L	M	M
25	MAL	G	S	S	S	S	MH	L	MH	M

Figure 1 : Soil Dataset in.CSV format

This soil dataset is allowed to open in WEKA as shown in figure 2.



Filter: None

Current relation
Relation: Fertility_index_file_1
Instances: 1639 Attributes: 10

No.	Name
1	pH
2	EC
3	Fe
4	Zn
5	Mn
6	Cu
7	OC
8	P2O5
9	K2O
10	FI

Selected attribute
Name: pH Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

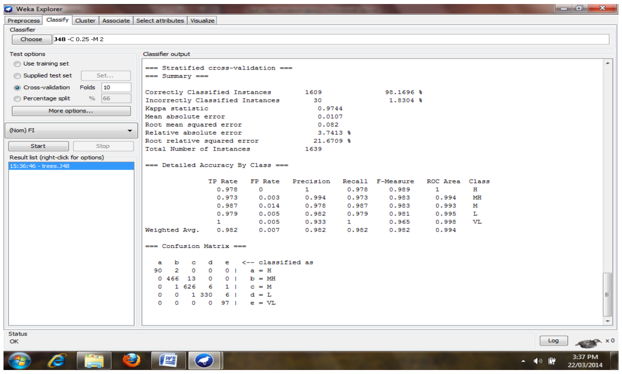
No.	Label	Count
1	LAL	562
2	MAL	721
3	N	157
4	LAC	59
5	HAL	131
6	VHAL	9

Class: FI (Nom)

Status
OK

Figure 2 : Soil Dataset open in weka

After opening soil dataset in WEKA, apply classification technique by using Classify tab and then Choose J48 algorithm. J48 Classifier generates rules as shown in fig 3.



Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) FI

15:36:46 - trees.J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1609               98.1696 %
Incorrectly Classified Instances        30                1.8304 %
Kappa statistic                          0.9744
Mean absolute error                      0.0107
Root mean squared error                  0.082
Relative absolute error                  3.7413 %
Root relative squared error             21.6709 %
Total Number of Instances             1639

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.978     0         1          0.978     0.989      1        H
                 0.973     0.003     0.994      0.973     0.983      0.994    MH
                 0.987     0.014     0.978      0.987     0.983      0.993    M
                 0.979     0.005     0.982      0.979     0.981      0.995    L
                 1         0.005     0.933      1         0.965      0.998    VL
Weighted Avg.    0.982     0.007     0.982      0.982     0.982      0.994

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  90   2   0   0   0 |   a = H
   0 466  13   0   0 |   b = MH
   0   1 626   6   1 |   c = M
   0   0   1 330   6 |   d = L
   0   0   0   0  97 |   e = VL
```

Status
OK

Figure 3 : Result of J48 algorithm on soil dataset.

Results of J48 algorithm are compared with other algorithm such as naïve Bayes and random forest. Information of all algorithms is summarized in following table. From table we can say that J48 algorithm gives better result among the others.

Table 1: Comparative result of classifier

Evaluation Criteria	J48	Naïve Bayesian	Random forest
Time to build model (in secs)	0.06	0.03	0.06
Correctly classified instances	1609	1265	1605
Incorrectly classified instances	30	374	34
Prediction Accuracy	98.17%	77.18%	97.92%

Figure 4 shows J48 classifier algorithm performs best among other classifiers on the basis of prediction accuracy.

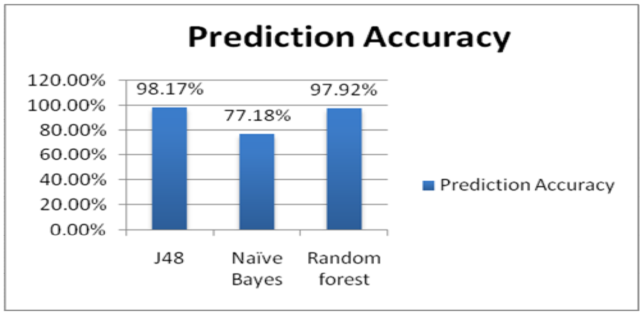


Figure 4 : Prediction Accuracy of classifier

Figure 5 shows that correctly classified instances and incorrectly classified instances for all classifiers. So J48 algorithm performs better among other classifiers and decision tree form by J48 shows that fertility rate for Aurangabad district is medium. This will help to decision maker to recommend fertilizers accordingly.

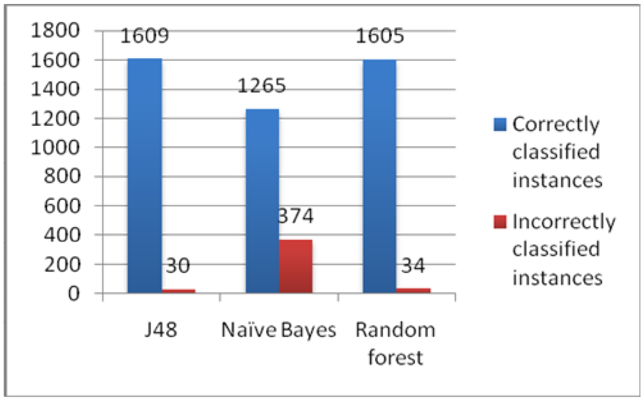


Figure 5 : Error rate of classifiers.

Figure 6 shows that the time taken to build the classifier model.

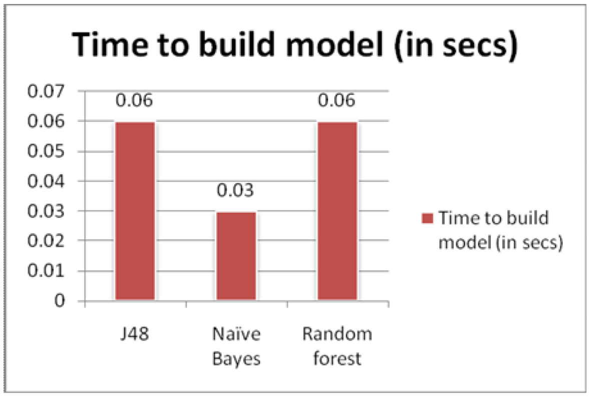


Figure 6: Time taken to build classifier model.

4. CONCLUSION

Data mining is new research area in agriculture. As agriculture is a soil-based industry, there is no way that required yield increases of the major crops can be attained without ensuring that plants have an adequate and balanced supply of nutrients. In this paper different classifier algorithms are used on soil dataset to predict fertility rate. Study shows that among the classifier J48 classifier perform better to predict fertility index. Observation also shows that fertility rate for Aurangabad district is medium. This will help to decision maker to recommend fertilizers accordingly.

Acknowledgement

The Author would like to thanks Prof. Deepak Bornare, Head Soil testing Laboratory, Agricultural Engineering department MIT Aurangabad for his cooperation and providing data for this research. His guidance on analysis of result is equally important for me.

References

[1] Peter Gruhn, Francesco Goletti, and Montague Yudelman "Integrated Nutrient Management, Soil Fertility, and Sustainable Agriculture: Current Issues and Future Challenges" Food, Agriculture, and the Environment Discussion Paper 32, 2000

[2] Compendium on Soil Health, Department of Agriculture & Cooperation (INM Division), January 2012

[3] N Srihari Rao "Prospective Usage of ICT by Farmers for Agriculture" 2013

[4] Jiawei Han, Micheline Kamber, "Data Mining : Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006.

[5] S.S.Baskar L.Arockiam S.Charles "Applying Data Mining Techniques on Soil Fertility Prediction**"** International Journal of Computer Applications Technology and Research Volume 2– Issue 6, 660 - 662, 2013

[6] Ravindra M, V. Lokesha, Prasanna Kumara, Alok Ranjan "Study and Analysis of Decision Tree Based Irrigation Methods in Agriculture System" International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 2, Issue 12, December 2012

[7] D Ramesh, B Vishnu Vardhan "Region Specific Crop Yield Analysis: A Data Mining Approach" UACEE International Journal of Advances in Computer Science and its Applications – IJCSIA Volume 3 : Issue 2 [ISSN 2250 – 3765] 05 June 2013

[8] S. Veenadhari, Dr. Bharat Mishra, Dr.CD Singh "Soybean Productivity Modeling using Decision Tree Algorithms" International Journal of Computer Applications (0975 – 8887Volume 27– No.7, August 2011

[9] Georg Ruß "Data Mining of Agricultural Yield Data: A Comparison of Regression Models"

[10] Jay Gholap "Performance Tuning of J48 Algorithm for Soil Fertility"2012. Asian Journal of Computer Science and Information Technology 2: 8 (2012) 251– 252

[11] Suman, Bharat Bhushan Naib "Soil Classification and Fertilizer Recommendation using WEKA" IJCSMS International Journal of Computer Science & Management Studies, Vol. 13, Issue 05, July 2013

[12] P. Revathi, Dr.M.Hemalatha "Categorize the Quality of Cotton Seeds Based on the Different Germination of the Cotton Using Machine Knowledge Approach" International Journal of Advanced Science and Technology Vol. 36, November, 2011

[13] D Ramesh, B Vishnu Vardhan "Data Mining Techniques and Applications to Agricultural Yield Data" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013

[14] Margaret Dunham, "Data Mining: Concepts and Techniques", Morgan Kaufmann Pub.

[15] Random forest "Wikipedia"

AUTHOR

Vrushali Bhuyar is a Graduate from Shivaji College, Post Graduate from Amravati University, Amravati and having 9 years of experience in teaching. Presently working as an Assistant Professor in Marathwada Institute of Technology, College of Engineering, Aurangabad.