International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 02, February 2017
www.ijates.com
ijates
ISSN 2348 - 7550

AGRICULTURAL SOIL LIME STATUS ANALYSIS USING DATA MINING CLASSIFICATION TECHNIQUES

[1]Dr. K. Arunesh, [2]V. Rajeswari

[1]Department of Computer Science, Sri S. Ranmasamy Naidu Memorial College, (India)
[2]Department of Computer Science, G. Venkataswamy Naidu College (SFC), (India)

ABSTRACT

Data Mining is one of the emerging research fields in Agriculture soil analysis. In this paper our focus is on the applications of Data Mining Classification Techniques in agricultural field on analyzing the soil data. Soil Lime status is very important in agriculture sector because of high lime status in soil should contains the fewer amounts of other nutrients. So, prediction of lime status says about the other default nutrients levels. The level of soil lime status is analyzed using various data mining classification techniques such as, J48, Random Tree, JRip, OneR and Naïve Bayes. Various lime status levels based on the soil type, soil texture and PH value are classified in this paper and the performance of the data mining classification algorithms are also analyzed in this research work.

Keywords: Data Mining, classification, Lime Status, Naïve Bayes, Soil Data

I. INTRODUCTION

Data mining is defined as mining the knowledge from large amount of data. Here, knowledge refers to the useful information prediction from the database using with various data mining techniques. Data mining techniques are having an ability to find the relationship and pattern from existing database. India is an agriculture country and Indian economic status also depends on the agriculture. Agriculture is one of the most important research fields because variety of data is available in this field for researchers. Biological and Agricultural and research studies have been used for the purpose of various techniques of data analysis including natural trees, statistical machine learning and other analysis methods. This research aimed to assess data mining techniques and apply them to a soil science database to establish the meaningful relationships. It can be found. Different techniques were proposed for mining data over the years. A detailed and elaborated Data Mining techniques were discussed by the researchers [1].

A soil test is the analysis of a soil sample to determine nutrient content, composition and other characteristics. Tests are usually performed to measure fertility and indicate deficiencies that need to be remedied [2]. The soil testing laboratories are provided with suitable technical literature on various aspects of soil testing, including testing

methods and formulations of fertilizer recommendations [4]. It helps farmers to decide the extent of fertilizer and farm yard manure to be applied at various stages of the growth cycle of the crop.

For supervised data mining classification concept, classification is defined as a tree based structure of data mining technique. Popular classification techniques like Naïve Bayes, Decision Tree, J48, Random Tree, JRip, and OneR are considered for to achieve a high accuracy and a high generality in terms of soil lime status prediction capabilities and analysis. For this purpose, different types of Data Mining techniques were evaluated on the data set.

II. LITERATURE REVIEW

There are several applications of Data Mining techniques in the field of agriculture. Some of the data mining techniques are related to ecosystem management, weather conditions and forecasts, yield prediction, soil type analysis etc., This section discusses the various data mining techniques from the past research. Classification, clustering, association rule mining and support vector machine are the data mining techniques considered for this survey in different aplications.

Lida Xu et,al., proposed a systematic approach based on Integrated Information System for agricultural ecosystem management. For the purpose of agricultural land and ecosystem management and integration, they extract data on terrain, planting and land usages. Finally they concluded that, a systematic approach is essential in which Integrated Information Systems play a crucial role for effective management of agriculture and ecosystems.

Arvind Jaiswal, Gaurav Dubey et.al., discovers association rules and optimization using genetic algorithm for finding frequent item sets. They proposed genetic algorithm based method. The following various steps are executing to repeatedly transform the population. The fitness evaluation is an objective function which is calculated for each individual. The current population as parents to be involved in recombination and individuals are chosen from that are produced from the parents by applying genetic operators such as crossover and mutation. These New individuals often called offspring. Some of the new individuals are replaced with some individuals this process is done usually with their parents. A combination of genetic algorithms and a modified a-priori based algorithms are used to design an interactive Association Rule mining.

Neelamadhab Padhy et al., gives a complete introduction about data mining. Authors explained about the data mining classification tasks, data mining life cycle and other applications. Variety of data mining techniques, approaches and crucial research areas in data mining which have been marked as the important field of data mining technologies for research are focused. applications of data mining with proposed feature directions for some of data mining applications like data mining techniques in Healthcare, Education and Market Basket Analysis are focused.

Vijayarani, et.al., discussed about the data mining classification techniques to predict heart diseases. They applied data mining classification rule algorithms namely, Decision table, OneR, JRip and Part. From this experimental

result of accuracy measure, they observed that the Decision Table classification rule technique contains more accuracy than other algorithms. So, it is the best classifier for heart disease prediction. The OneR and Decision Table classification algorithm contain least error rates based on two outcomes analyzed by the authors.

C. Lakshmi Devasena, et.al., studied the effectiveness of Rule-Based classifiers for classification. They taking a sample data set and comparing with different rule-based classifiers such as, Conjunctive Rule Classifier, Decision Table Classifier, DTNB classifier, JRip classifier, OneR Classifier, NNGE Classifier, RIDOR Classifier, PART Classifier, and ZeroR Classifier. They taken the Iris Data set to evaluate the performance of these classifiers experiment have been done using an open source Machine Learning Tool. The following parameters like Accuracy, RMSE, MAE and Confusion Matrix are measured to analyze the performances of these various classification algorithms. NNGE Classifier gives better result.

Dr. Medhat Mohamed Ahmed Abdelaal and Muhamed Farouq, they extract the mammographic mass features using with the classification technique as Support Vector Machine with Tree Boost and Tree Forest in analyzing the DDSM dataset. They extract these features along with age. These features discriminates true and false cases. Here, SVM techniques show results in high accuracy of classifying. The largest area under the ROC curve is compared to values for tree boost and tree forest.

S. Anupama Kumar and M. N. Vijayalakshmi explained that various data mining techniques like classification, clustering. These techniques are applied on the student's data base. This study can be used to enable the learner and teaching community to increase the performance. To increase the capacity of the model these techniques can also be combined with them. In this paper explained many data mining techniques according to design a new environment Result to Educational Data. Then education system can enhanced their performance by using these data mining techniques. In this paper, every method has its own key area which is performed accurate.

Umamaheswari and S. Niraimathi illustrate the various data mining techniques which are applied to analysis the student records in order to categorize the students into their grade order in all their education studies and it helps in interview situation. It predicts that which techniques help to categorize students in rank order to arrange for the recruitment process. From this experimental, we can easily discover the eligible student using with this technique and it also reduces the short listings. From result of this paper, efficient data mining techniques are described to manage the performance level of students. Classification is one of the data mining techniques which are used to accurately classify the data for categorizing student based on the levels. Clustering is an important data mining function to analysis and discovers data sources distribution of information and, the cluster analysis is an important research topic. This way is help to how define the recruitment process in an easier manner.

Shiv Pratap Singh Kushwah, et.al., describe the comparisons of data mining clustering algorithms for clustering. In this paper also covered classification, clustering techniques. Data mining techniques are used in every field for the analysis of large volumes of data. The K mean approach is use to predict the solution less sensitive to initialization

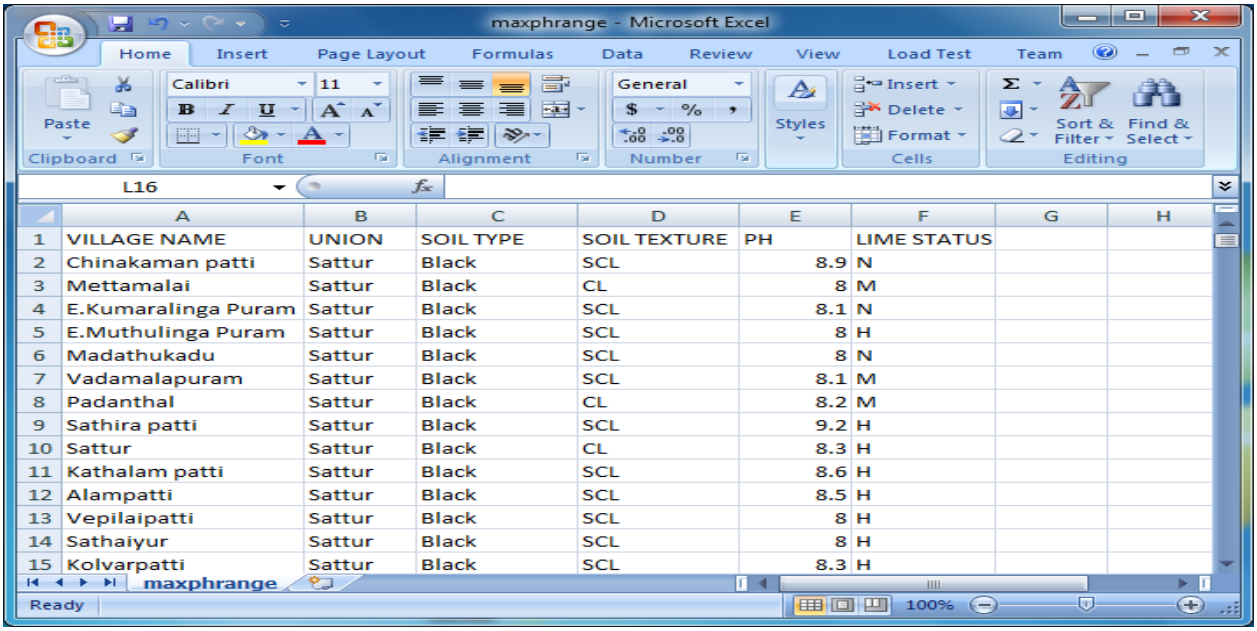
and provides results at multiple resolutions, and K-mean algorithm is also sensitive to the presence of outliers. KNN classification is an easy to implement and easy to understand.

Madhuri V. Joseph, Lipsa Sadath and Vanaja Rajan has compared the various data mining techniques. To exploring the important information from large amount of database using with various data mining techniques it can deals with different data type. In this paper, they explained and compared some common data mining techniques which are mainly used in our business environment and daily life. Every data mining techniques and its functionality is an important role in the business environment. So, there is no any one model to plays the entire roll in business environment.

III. DATASET DESCRIPTION

The data carried out in this paper is obtained for the years 2013 to 2015 from the Agricultural soil testing laboratory, Virudhunagar District, Tamilnadu, India. Primary data for the soil survey are acquired by field sampling. Dataset has 6 attributes such as, Village Name, Soil Type, Color, Soil Texture, PH and Lime Status, and 203 data instances. Here, two types of soil type are considered Black and Red. Soil texture contains three categories such as, SCL (Sandy Clay Loam), CL (Clay Loam) and SL (Sandy Loam). Class label Lime Status contains three categories like High, Medium and Nil.

Figure.1 shows the soil data set which contains the attributes such as, Village Name, Union, Soil Type, Soil Texture, PH value and Lime Status. This data set organized in Excel Sheet with saves as type .CSV extension. Here the entire dataset is considered as training set.



	A	B	C	D	E	F	G	H
1	VILLAGE NAME	UNION	SOIL TYPE	SOIL TEXTURE	PH	LIME STATUS		
2	Chinakaman patti	Sattur	Black	SCL	8.9	N		
3	Mettamalai	Sattur	Black	CL	8	M		
4	E.Kumaralinga Puram	Sattur	Black	SCL	8.1	N		
5	E.Muthulinga Puram	Sattur	Black	SCL	8	H		
6	Madathukadu	Sattur	Black	SCL	8	N		
7	Vadamalapuram	Sattur	Black	SCL	8.1	M		
8	Padanthal	Sattur	Black	CL	8.2	M		
9	Sathira patti	Sattur	Black	SCL	9.2	H		
10	Sattur	Sattur	Black	CL	8.3	H		
11	Kathalam patti	Sattur	Black	SCL	8.6	H		
12	Alampatti	Sattur	Black	SCL	8.5	H		
13	Vepilaipatti	Sattur	Black	SCL	8	H		
14	Sathaiyur	Sattur	Black	SCL	8	H		
15	Kolvarpatti	Sattur	Black	SCL	8.3	H		

Figure.1. Soil Dataset

IV. RESULT AND ANALYSIS

Soil lime status classification system is essential for the identification of soil properties. Expert system can be a very powerful tool in identifying the types of soils and other properties quickly and accurately. Traditional classification systems include use of tables and flow-charts. This type of manual approach takes a lot of time, hence quick, reliable automated system for soil classification is needed to make better utilization of time and effort. WEKA is one the popular data mining tool and the tool is used in this work to predict and analyze the status of the lime in soils. In this work, J48, Random Tree, JRip, OneR and Naïve Bayes are evaluated and applied to predict the lime status of soil using WEKA. The evaluated results are compared on basis of time, accuracy, Error Rate, True Positive Rate and False Positive Rate.

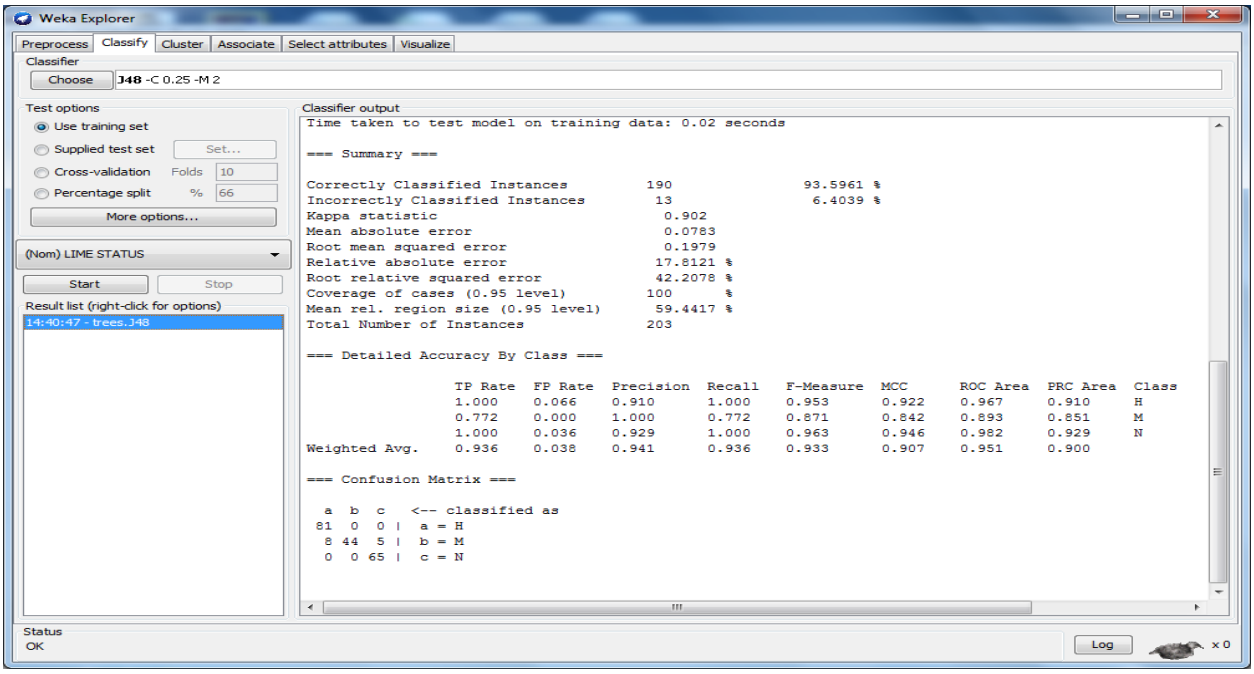


Figure 2. Soil Status Classification Using J48

J48 performance analysis is shown in Figue.2. from 203 data instances, J48 correctly classified 190 instances and incorrectly classified 13 instances. Here, Kappa Statistics become nearest 1. Confusion Matrix is formed based on the correctly and incorrectly classified instances. The error rate, true positive rate, false positive rate, precision, ROC area and recall are also described in this result.

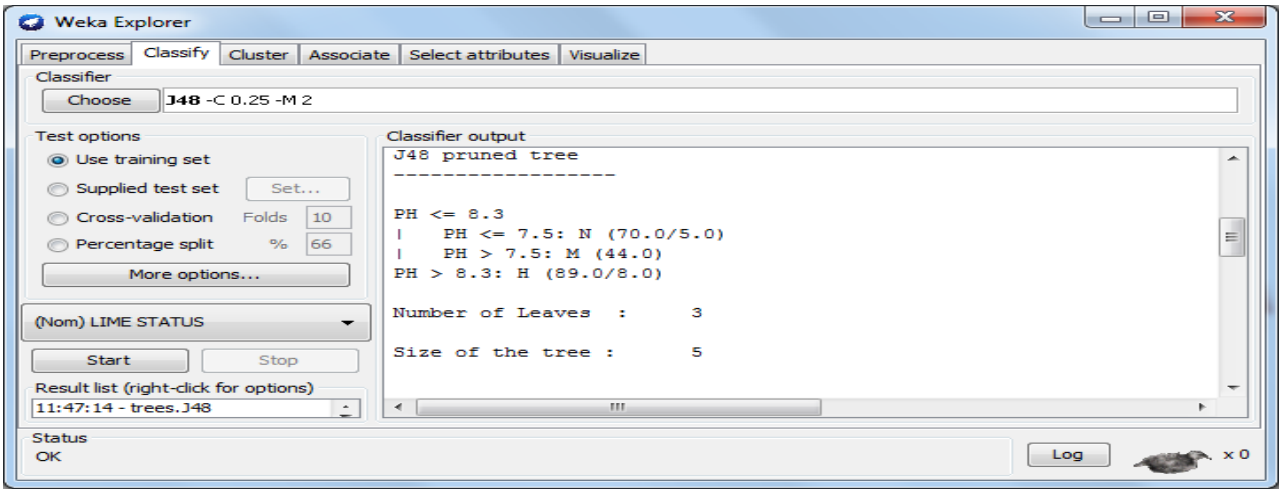


Figure 3. Soil status classification J48 Rules

The rules generated by the J48 are given in Figure.3. PH value is divided into two categories which are PH value greater than and less than or equal to 8.3. It is predicted that When PH value is greater than 8.3, lime status also become high otherwise medium or nil.

Naïve Bayes classifier performance results are shown in Figure.4. It is observed that the error rate is very low when it compared with J48.

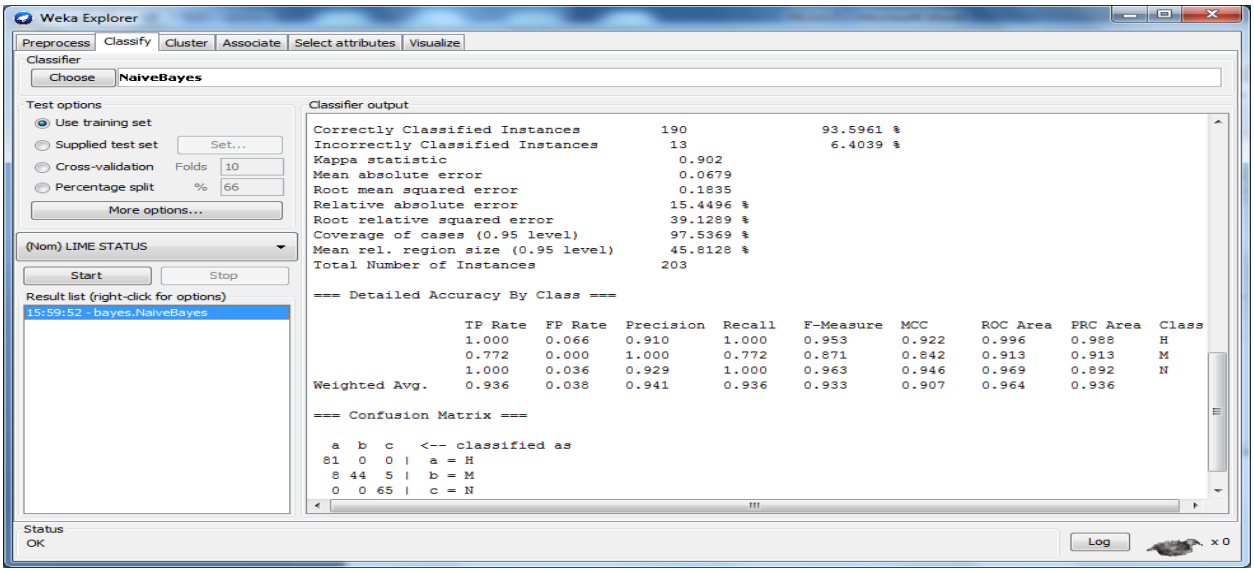


Figure 4. Soil status classification using Naïve Bayes Classifier

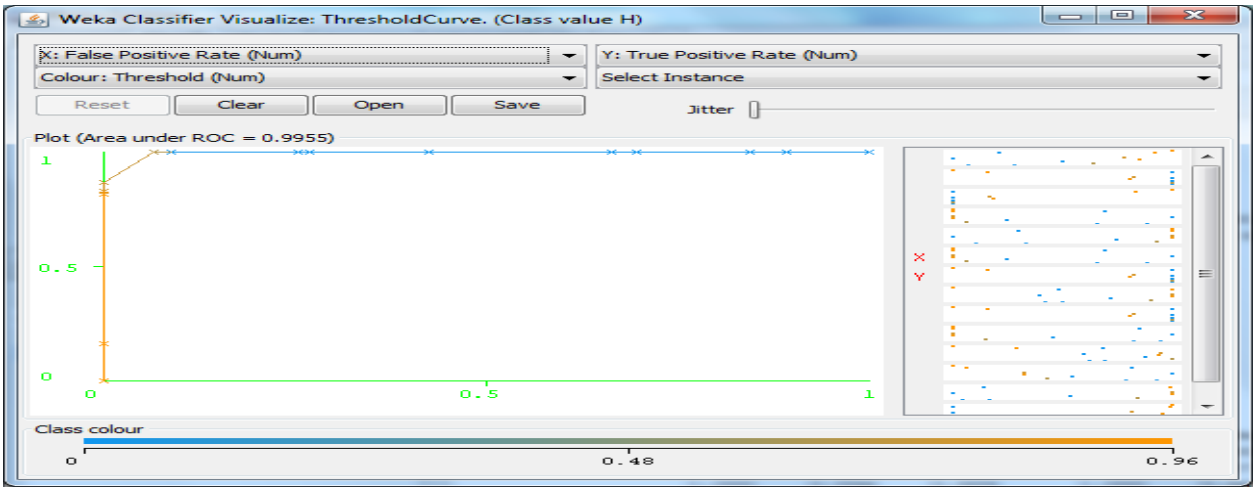


Figure 5. ROC Curve for High Lime Status

The ROC Graph for Nominal Class as High to Lime Status using Naïve Bayes classification, as shown in Figure.5. Area under ROC becomes 0.9955 which is nearest to one. From this Graph the curves on Y – axis and closer to follows the left-hand border and then the top border of the ROC space.

Table. 1. Performance Accuracy for different Classifiers.

ALGORITHM	ACCURACY
J48	93.46 %
Random Tree	50.66 %
JRip	93.46 %
OneR	39.92 %
ZeroR	39.92 %
Naïve Bayes	93.81 %

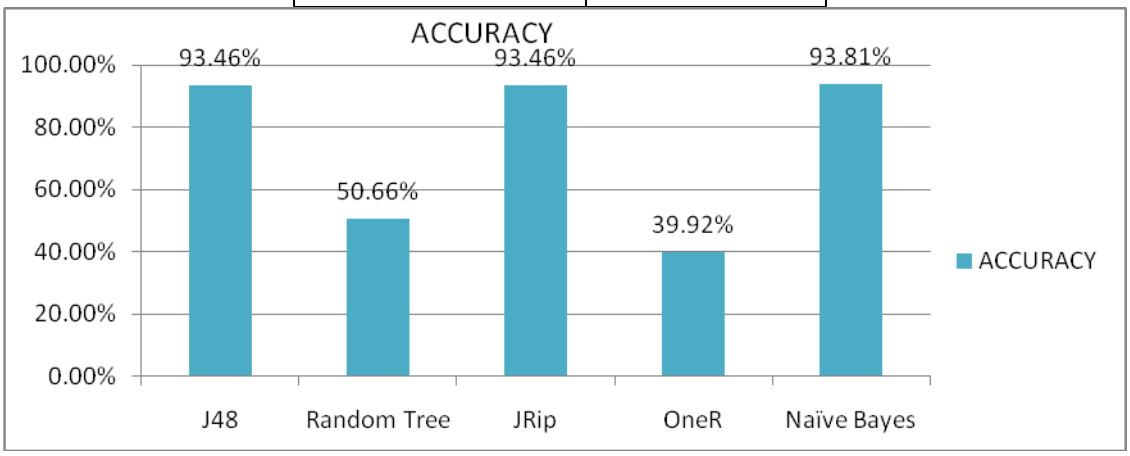


Figure. 6. Algorithm Performance Accuracy

The performance accuracy conducted in WEKA experimenter for the methods are shown in Table.1. Figue.6 shows the graphical representation of performance accuracy of the methods considered for this work. Naïve Byes gives better performance and shows high accuracy than others. Table. 2. and Figure.7. shows about the classifiers execution error rate for given dataset. Here, Random Tree has low error rate for Root Mean Squared Error. OneR algorithm has low error rate in Mean Absolute Error.

Table.2. Algorithm Error Rate

ALGORITHM	MEAN ABSOLUTE ERROR	ROOT MEAN SQUARED ERROR
J48	0.0783	0.1979
Random Tree	0.0506	0.1591
JRip	0.0783	0.1979
OneR	0.0427	0.2066
Naïve Bayes	0.0679	0.1835

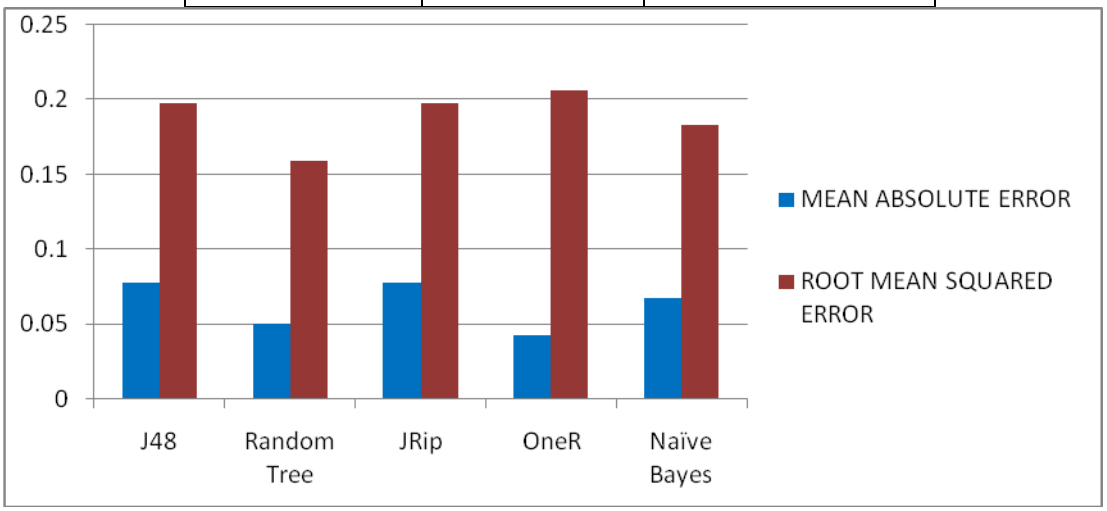


Figure .7 Algorithm Performance Error Rate

Finally, it is observed that when PH values become less than 7.85, then the soil texture will be Sandy Loam and Soil Type is Red then there is no lime status. PH value becomes greater than 8.6 then lime status becomes high. When PH value between 7.85 and 8.6 then the soil lime status is in medium level.

V. CONCLUSION

In this work, we have proposed an analysis of the soil data using various data mining classification techniques and prediction technique to predict the lime status level in soil. In spite the results that the Naïve Bayes classification algorithm gives better results. We have demonstrated comparative study of various classification algorithms with the

help of WEKA. In decision tree classifier, J48 gave the high accuracy. In future, we planned to propose Recommender System to predict and recommend appropriate fertilizer and crop suits for the lime status level in soil.

REFERENCES

[1] Lida Xu, Ning Liang, and Qiong Gao, July 2008, "An Integrated Approach for Agricultural Ecosystem Management", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, *VOL. 38, NO. 4.*

[2] Arvind Jaiswal, Gaurav Dubey., May 2013, "Identifying Best Association Rules and Their optimization using Genetic Algorithm", *IJESE, ISSN: 2319-6378, VOL.1, Issue 7.*

[3] Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi, (2009), *"The survey of data mining applications and feature scope."* arXiv preprint arXiv:*1211.5723*, 2012. [8] Phyu, Thair Nu. *"Survey of classification techniques in data mining".* Proceedings of the International MultiConference of Engineers and Computer Scientists. *Vol.1*.

[4] Vijayarani S, Sudha, February 2013, "An Effective Classification Rule Technique for Heart Disease Prediction", International Journal of Engineering Associates (IJEA), *ISSN: 2320-0804, Vol.1, Issue 4, P.No.81-85*.

[5] Devasena, C. Lakshmi, et l. "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set." Bonfring International Journal of Man Machine Interface 1. Special Issue Inaugural Special *Issue: 05-09*, 2011. a

[6] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, (2010), "Using data mining for assessing diagnosis of breast cancer," in Proc. International multi conference on computer science and information Technology, pp. 11-17.

[7] S. Anupama Kumar and M. N. Vijayalakshmi, February 2013, "Relevance of Data Mining Techniques in Edification Sector" International Journal of Machine Learning and Computing, *Vol. 3, No. 1*.

[8] Umamaheswari and S. Niraimathi "A Study on Student Data Analysis Using Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering.

[9] Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta, August 2012, "Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining" International Journal of Innovative Technology and Exploring Engineering (IJITEE) *ISSN: 2278-3075, Volume-1, Issue-3.*

[10] Madhuri V. Joseph, Lipsa Sadath and Vanaja Rajan, February 2013, "Data Mining: A Comparative Study on Various Techniques and Methods" International Journal of Advanced Research in Computer Science and Software Engineering *Volume 3, Issue 2, ISSN: 2277 128X*.