



JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Comparison of Data Mining Classification Algorithms using Soil Dataset

R. Sudha Abirami¹, M. S. Irfan Ahmed²

¹Assistant Professor, Department of Computer Science, sudha.tbakc@gmail.com

²Professor, Department of Computer Science, drmsirfan@gmail.com

Thassim Beevi Abdul Kader College for Women, Kilakarai

Ramanathapuram District, Tamil Nadu, India

ABSTRACT

Data Mining is one of the emerging research fields in Agriculture soil analysis. In this paper, the focus is on the applications of Data Mining Classification Techniques in agricultural field on analyzing the soil data. This paper presents the application of a random Forest algorithm for classification of Soil data to classify soil. e. The database has physical and chemical properties. The physical and chemical properties of the soil always play an important role in farming. The Classification algorithms are used for discovering the rules that classify the data into separate groups. Categorizing the soil with to the soil nutrients which is present in the soil, is more useful to the farmers and they predict which crop can be cultivated in a particular soil type. In this paper, the proposed method classifies the soil according to the macro nutrients and micro nutrients.

Key words : *Classification, Feature extraction, Discretization, Data mining*

I. INTRODUCTION

Indian economy is highly depending on agriculture. Agriculture is the major source of income for the most of the population. So, farmers are regularly interested about yield prediction. Many factors are important like soil, weather, rain, fertilizers and pesticides are used to increase the crop production. In agriculture field, Data mining plays a main role in crop yielding. There is a need to transform the large data into technologies and make them available to the farmers. It is can be very useful for farmers to take efficient and effective decision. Soil is one of the parameters which is used to increase crop

production is considered.

Data mining is the process to find interesting knowledge from large amounts of data [1]. The aim of the data mining process is to extract knowledge from an existing data set and transform it into a human understandable formation for advance use. It is the process of analyzing data from different view and encapsulates it into useful information. There is no constraint to the type of data that can be analyzed by using data mining algorithms. It analyzes data hold in a relational database, a data warehouse, a web server log or a simple text file. Analysis of data in successful way requires understanding of appropriate techniques of data mining. This paper is to gives the details about different data mining techniques in view of agriculture domain for soil classification. Data Mining is important to determine the agricultural related information such as soil fertility, yield prediction and soil erosion. Soil prediction helps for soil remedy and crop management.

Soil is a important natural resource and nonrenewable resource for agricultural development. It gives plants vital nutrients such as minerals, water, and air, which support in their physical production, strong growth, survival, and flourishing. Fertile soil is indeed a good foundation for growing stable and nutritious crops [2]. It performs a variety of productive functions while causing no deterioration or harm to the environment [3]. Soil fertility is specifically affected by its intrinsic physical, physiological, biological, and mineralogical properties [4]. The measurement and valuation of soil properties are normally performed by chemical methods of manually collected soil samples. Since the technique

is complex and time-consuming methods for evaluating or estimating some of the properties utilizing previously specified features are needed. The soil must first be divided into distinct similar classes before it can be defined. Without a proper rating, soil analysis is equivalent to conducting field experiments with green plants or laboratory experiments with the minimum of soil nutrients [5]. As a consequence, soil classification has been an essential aspect of soil science.

II. LITERATURE REVIEW

The classification method divides the soil data into separate groups based on certain predefined criteria. It also oversees the formal classification of soils based on distinct characteristics, as well as the parameters that describe the choices and options [2,6]. Furthermore, it assists in predicting the action and ability of the land for crop production, soil reduction mitigating environmental degradation, and increasing productivity. The description of soil increases information, comprehension, and coordination [7,8]. The implementation of a classification model that classifies soils based on soil properties as health indicators will increase fertilizer use and farmland reuse for different crop types.

The author V.Bhuyar et al represented the classification of soil fertility rate using J48, Naïve Bayes, and Random forest algorithm in the paper. The author concludes that J48 algorithm gives better result than other algorithms. The J48 set of rules facilitates the farmer and decision makers to perceive the soil fertility rate and at the assist of nutrients observed in the soil sample exclusive fertilizers may be recommended [9]. It is useful for researchers to gain information of current framework of data mining techniques and applications to classify soil fertility [10]. Author R.Ramesh Vamanan et al aimed to evaluate the various classification techniques of data mining and apply them to a soil science database to establish if meaningful relationships can be found. The application of data mining techniques has been organized for Tamil Nadu soil data sets. This paper compares the different classifiers and the outcome of the research could improve the management and systems of soil uses throughout a large number of fields that include agriculture, horticulture, environmental and land use management [11]. The Author V.Rajeswari et al Narrate the comparative analysis of three algorithms like Naïve Bayes, JRip and J48 is projected. Finally, the author concludes that JRip classification algorithm gives good result of this soil dataset when compared to other classification algorithms and also is correctly classified into maximum number of instances comparing with the other two classification algorithms. JRip can be

recommended to predict soil types [12]. The author Sofianita et al demonstrated the application of SOM in soil classification to identify the type of soil. The results are compared with k-means algorithm. The proposed algorithm has shown its ability in classifying the soil with a 91.8% of accuracy [13].

III. RESEARCH METHODOLOGY

Classification algorithms involve finding rules that partition the data into disjoint groups. A set of classification methods are used by a classification process, which can be used to classify soil data. Classification algorithms such as Naive Bayesian classifier, J48 decision tree classifier and JRip classifier are used.

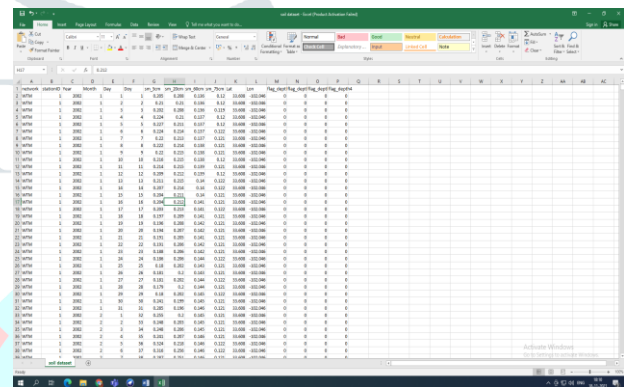


Figure 1 : Sample data set

This dataset has 13 attributes: CY, SN, SL, PH, CaCl₂, OC, N, Ca, P, Mg, K, Na, and EC. The USCS classifies soils into three types: coarse-grained soils (such as sands and gravels), fine-grained soils (such as silts and clays), and highly organic soils (referred to as "peat"). Table 1 displays the attribute description.

S.No.	Feature	Particulars
1.	CY	Clay Content of the soil
2.	SL	Salinity Of the soil
3.	SN	Quantity Of sand of the soil
4.	PH	PH value of the soil
5.	CaCl ₂	Calcium Chloride content of the soil
6.	OC	Organic Carbon
7.	N	Nitrogen Content Of the soil
8.	P	Phosphorus Content of the soil
9.	Ca	Calcium Content of the soil
10.	Mg	Magnesium content of the soil
11.	K	Potassium content of the soil
12.	Na	Sodium content of the soil
13.	EC	Electrical conductivity of the soil

Table 1 Data Set Description

For clarity, the USCS splits the three major soil classes into subgroups. Soil color, moisture content, weight, and slightly more detail about the nutrients of the soil will be included in a full geotechnical engineering soil specification [14].

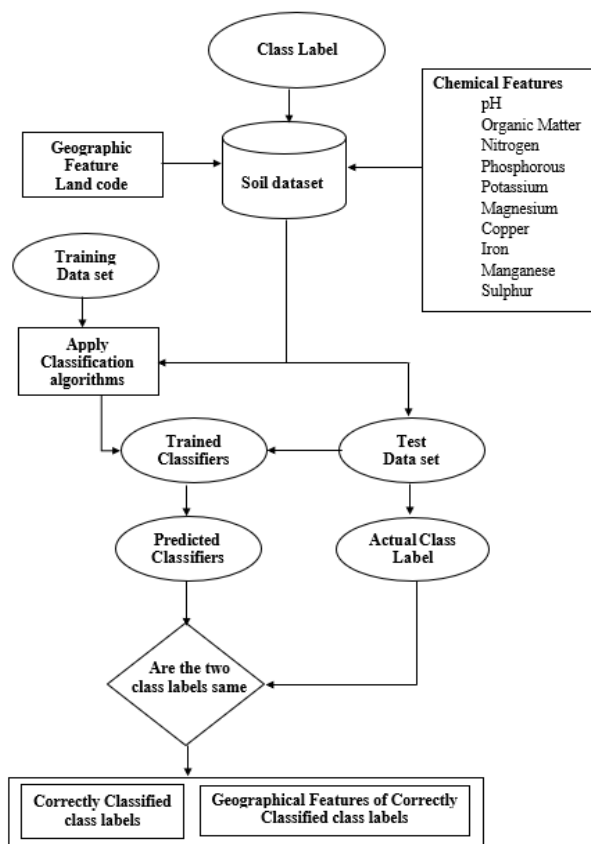


Figure : 2 : Implementation process flow

i. Naive Bayes Classifier :

A Naive Bayes classifier is simple probabilistic classification techniques in machine learning. It is mainly based on the Bayes theorem with independence features. Each class labels are estimated through probability of data instance. It needs only small amount of training data to predict class label necessary for classification.

ii. J48 (C4.5) :

The J48 is one of the classification-decision tree algorithm. It can select the test as best information gain. This algorithm J48 is also referred to as a statistical classifier. J48 predicts dependent variable from existing data. It builds tree with attributes values of training values. This classifies data feature of data instances that have information gain. The importance of error tolerance is developed using pruning concept.

iii. JRip :

JRIP is also referred to as Repeated Incremental Pruning to Produce Error Reduction (RIPPER). This classification algorithm is a propositional guidelines to learners. J-Rip classifier is decision tree pruning models

with using association rules. It is an effective technique to minimize error pruning. In this algorithm, the training data is split into two sets and with pruning operators, the error is reduced on both the sets. Then, rules are formed from two sets such as Growing set and Pruning set.

IV RESULTS AND DISCUSSION

Three classification algorithms such as JRip, J48, Naive Bayes algorithms are used to classify the soil types. While applying three classifier algorithms, JRip classification algorithm concentrates the whole attributes of data set. But, J48 classifier considers only Longitude and Latitude (Location) values and tree is build based on these two attributes. But JRip classifier generates the rules efficiently and shows good performance for this soil data set. As comparing these three classification algorithms JRip gives in high accuracy. Moreover, the whole dataset considered as training set.

Based on the training data set, it is concluded that weighted average of True Positive Rate of JRip classifier is 0.982. In the case J48 and Naïve Bayes classification algorithm True Positive (TP) Rate is 0.97 and 0.86 it indicates the lowest level. So, automatically JRip classifier classified the data set with improved knowledge. pH is that below 7.0 is acidic based soil and above 7.0 is alkaline based soil. The soil dataset which contains the attributes like longitude, latitude. This data set organized in Excel Sheet as type is CSV extension.

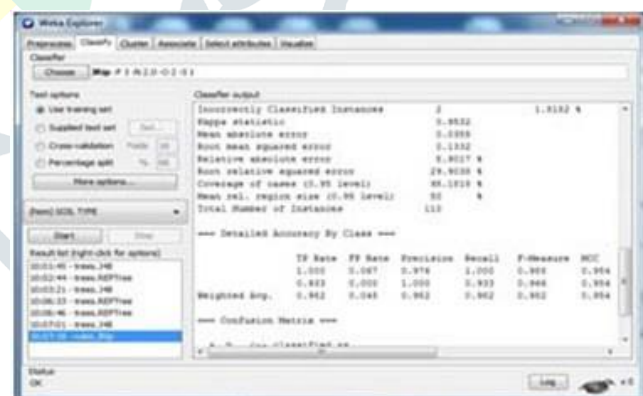


Figure 3 : JRip Classifier Result

The number of incorrectly classified instances, error rate of JRip is given in Figure 3.

The number of correctly classified instances and incorrectly classified instances are given in Figure 4. Here, JRip classified maximum number of instances.

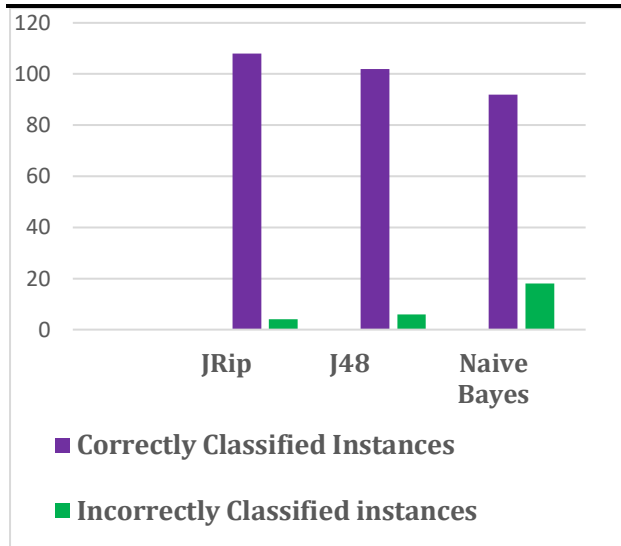


Figure 4 : Classifiers Error rate

Here JRip performed better classification to compare the other algorithms and also Kappa Statistic value becomes nearest 1.00 in JRip algorithm in Table 2.

Evaluation Criteria	Correctly Classified Instances	Incorrectly Classified instances	Prediction Accuracy	Kappa Statistic
JRip	108	4	98.18 %	0.9532
J48	102	6	97.27 %	0.9305
Naive Bayes	92	18	86.36 %	0.5926

Table 2. Comparative analysis of classifiers

The JRip algorithm gives the high prediction accuracy is given in Figure 5. The Naive Bayes Algorithm produces low accuracy compared than J48 and JRip.

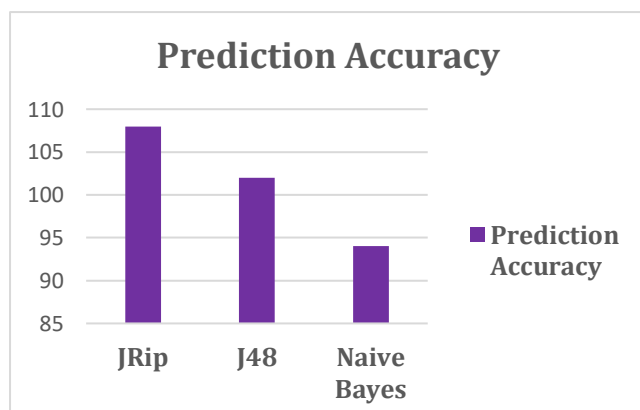


Figure 5 : Prediction Accuracy

V CONCLUSION

In this paper, the comparative analysis of three algorithms like Naïve Bayes, JRip and J48 is projected. JRip classification algorithm produces good result of this soil dataset and is correctly classified into maximum number of instances comparing with the other two classification algorithms. JRip can be used to predict soil types. In future, the soil prediction can be done using machine learning algorithms and deep learning. It may be implemented using Association rule mining for identifying suitable crops of soil.

VI REFERENCES

1. Jiawei Han, Micheline Kamber, —Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
2. Motia S, Reddy S. Ensemble classifier to support decisions on soil classification. IOP Conf. Ser.: Mater. Sci. Eng. 2021; 1022:012044. DOI: 10.1088/1757-899X/1022/1/012044.
3. El-Ramady HR, et al. Soil quality and plant nutrition. In Sustainable Agriculture Reviews 14, Springer. 2014;345–447.
4. Karlen DL, Ditzler CA, Andrews SS. Soil quality: Why and how?. Geoderma. 2003;114(3–4):145–156.
5. Hartemink AE. The use of soil classification in journal papers between 1975 and 2014. Geoderma Regional. 2015;5:127–139.
6. Brifcani A, Issa A. Intrusion detection and attack classifier based on three techniques: A comparative study. Eng. & Tech. Journal. 2011;29(2):368–412. Taher et al.; AJRCOS, 8(2): 17-28, 2021; Article no.AJRCOS.68035 27
7. Kovačević M, Bajat B, Gajić B. Soil type classification and estimation of soil properties using support vector machines. Geoderma. 2010;154(3–4):340–347.
8. Salim NO, Abdulazeez AM. Human diseases detection based on machine learning algorithms: A review. International Journal of Science and Business. 2021;5(2):102–113
9. Bhuyar V. Comparative analysis of classification techniques on soil data to predict fertility rate for Aurangabad District. International Journal of Emerging Trends and Technology in Computer Science. 2014 Mar-Apr; 3(2):200–3

10. B. Murugesakuma., Dr. K.Anandakumar., Dr. A.Bharathi., —Survey on Soil Classification Methods Using Data Mining Techniquesl. International Journal of Current Trends in Engineering & Research (IJCTER) e-ISSN 2455–1392 Volume 2 Issue 7, July 2016
11. R. Vamanan & K. Ramar, (2011), —Classification of Agricultural Land Soils A Data Mining Approachl, International Journal on Computer Science and Engineering, ISSN: 0975-3397, Vol. 3
12. V.Rajeswari. K.Arunesh., —Analysing Soil Data Mining Classification Techniquesl. Indian Journal of Science and Technology, Vol 9(19), May 2016
13. Sofianita., Jamian., —Soil Classification: An application of Self Organising Map and K-Meansl 978-1-4244-8136- 1/10/\$26.00_c 2010 IEEE
14. Pham BT, Hoang T-A, Nguyen D-M, Bui DT. Prediction of shear strength of soft soil using machine learning methods. Catena. 2018;166:181–191.

