Expert System for Land Suitability Evaluation using Data mining's Classification Techniques: a Comparative Study

C.Parthiban[1], M.Balakrishnan[2]
[1]Assistant Professor, Department of Computer Applications, JNRM, A & N Islands, India
(Research Scholar, PRIST University, Tamil Nadu, India)
[2]Sr.Scientist, National Agriculture Academy Research Management (NAARM), Hyderabad-500030, AP, India

Abstract: *Data mining involves the extraction of implicit, "interesting" information from a database. Classification is an important Data mining's "machine learning" technique which is used to predict data instances from dataset. It involves the order wise analysis of large amount of information sets. Data mining applications are used in various areas such as health care, insurance, medicines, Agriculture, banking and soil management. In soil region the Data mining mainly used to classify the soil and predicting the land suitability for the crop and fertilizer recommendation. The purpose of this study is to predict the land suitability for the crop using classification algorithms namely Naive Bayes and J48. This work focused on find out the best classification algorithm based on accuracy measure, performance measure, error rate and execution time using the soil dataset. From the experimental result using WEKA tool it is observed that the performance of the J48 is better than the Naive Bayes algorithm.*

Keywords: WEKA, Data Mining, Naïve Bayes, J48, Soil Dataset, Classification Algorithm.

I). INTRODUCTION

Data Mining refers to extracting or mining the knowledge or information from the large amount of data in the database or Knowledgebase. The other terminologies of the Data Mining are Knowledge mining from database, Knowledge extraction or Knowledge Discovery in Database (KDD).KDD process, in progressive order include data cleaning, data integration, data selection, data transformation, pattern evaluation, and knowledge presentation[1].The elements of Data Mining are extracting, transforming and loading transaction data on the data warehouse system, store and manage data in multidimensional database system [2]. The main application of the Data Mining is Web Mining. Data Mining has five functions which are Classification, Clustering, Association, Sequencing and Forecasting. Machine learning algorithms typically used in data mining have been applied to learn rules for an expert system based on examples provided by experts [3].Agriculture soil profiles are used in research for completeness of soils classification. Data mining techniques when applied to an agricultural soil profile, may improve the verification of valid soil profile, may improve verification of valid patterns and profile classification when compared to standard statistical analysis techniques [16].Expert systems have gained importance for data collection, organization, transmission, and recommendation [15].Machine learning algorithms typically used in data mining have been applied to learn rules for an expert system based on examples provided by experts [17]. The knowledge representation used by the Expert system is enriched to include explicit "strategic" knowledge, i.e. knowledge about how to reason, and domain-specific knowledge. From this knowledge, the rules used by the expert system are compiled, and this knowledge is also used to provide more abstract explanations of the system's reasoning [18]. Expert system applications areas are: Agricultural, Accounting and Finance, Business, Chemical, Computer, Construction, Engineering, Insurance, Medical and many other areas. COMAX is a Crop management expert system for cotton which can predict crop growth and yield in response to external weather variables, soil physical parameters, soil fertility, and pest damage [19].The Classification is the one of the major role in Data mining. Classification algorithms typically contain two phases which are Training Phase and Testing Phase. The most common methods used in data classification are decision trees, rule-based methods, probabilistic methods, SVM methods, instance-based methods, and neural networks [20].

The main objective of this work is carried out to find the land suitability for the particular crop. Using the soil dataset of Andaman, we developed an Expert System for soil classification which gives the recommendation to the farmers/end-user that the soil is unsuitable, highly suitable or moderately suitable for the crop. Here we carried out a comparative study of accuracy, performance

measure, time taken for execution and error rate of classification algorithms of Naive Bayes and J48 algorithm to find out an efficient algorithm for this soil dataset with the help of WEKA Data Mining Tool. The rest of this paper is organized as section 2 describes the proposed materials and methods, section 3 explains the experimental result and discussion and conclusions and future works are presented in section 4.

II. MATERIALS AND METHODS

A. Study area and dataset collection

Information of the soil is very much essential for the proper land use like which soil is suitable for cultivation and which crop is suitable for the particular soil, this is very useful to the farmers and all the persons which are helping to do good cultivation. Land use suitability study is the process of finding the suitability of a given land area for a certain type of use agriculture purpose, and the level of suitability [4]. During the systematic soil survey of the revenue area of Andaman and Nicobar islands, it was observed that a great variety of soils (8 soil series) occur in these islands [5]. We have collected the dataset from CIARI (Central Island Agricultural Research Institute, Port Blair, and it contains information about eight soil series in the islands (School Line, Garacharma, Dhanikhari, Rangachang, Tushnabad, Pahargoan, Wandoor, Little Andaman) [5]. This dataset has 11 attributes and it contains the total of 112 instances of soil sample. In Table 1 explains the attributes of the collected sample soil data.

TABLE I: ATTRIBUTES EXPLANATION

Attribute	Description
D (cm)	Soil Depth in (cm)
Sd texture	Sand texture %
St texture	Silt texture %
Cl texture	Clay texture %
pH	Potential hydrogen ion concentration value of soil
Org C	Organic Carbon %
EC (dsm^{-1})	Electrical Conductivity, decimen per meter
E.B	Exch. Bases value
P	Available Phosphorus
K	Available Potassium
Sl	Slope %

B. Training set

Table 2 shows the training dataset of the soil which contains the different attributes, with this attributes the expert system giving the recommendation for the crop whether it is unsuitable, suitable or moderately suitable. Like below we have created 112 instances to finding the land suitability for the crop, here we have taken three crops namely Arecanut, Coconut and Black Pepper.

C. Expert system for Land Resource Management

To discover the Soil site suitability for the crop with the help of an Expert System, it is very important that to classifying the Soil to identify the soil attributes. An Expert System is a powerful tool to give recommendation for the soil site suitability for the crop properly. An expert system has increase importance for data collection, organization, transmission, and recommendation [6]. We have developed an Expert System for Land Resource Management as a prototype which is classifying the soil and giving the soil site suitability recommendation for the crop. Here we have developed the rule engine for soil site suitability and the rule was collected from the domain expert. The soil training dataset instance were categorize into the site suitability class which labeled as unsuitable, suitable and moderately suitable for the crop. The example of rules given below that how it is classifying and giving the recommendation.

```
Rule1()
    {
      if (DEPTHS >=val && DEPTHS <= val
&& SAND <= val && SILT <= val && CLAY
      <= val && AVAILK <= val &&
SLOPEID >= val && SLOPEID <= val)
      {
       if ( PH> val ||PH< val || EC > val || EC <
       val || ORGC > val || ORGC < val ||
       EXCHBS >
       val || EXCHBS < val || AVAILP > val ||
       AVAILP < val) { }
        else
        {
           LIMITID = limitclass;
           SUITID = suitclass;
        } }
         if (DEPTHS >= val && DEPTHS <=
       val && SAND <= val && SILT <= val
       &&
                CLAY <= val && AVAILK <=
       val && SLOPEID >= val && SLOPEID
       <= val)
       {
         if (PH> val ||PH< val || EC > val || EC <
       val || ORGC > val || ORGC < val ||
       EXCHBS >
```

```
                val || EXCHBS < val ||
        AVAILP > val || AVAILP < val) { }
          else
          {
            LIMITID = LIMITCLASS;
            SUITID = SUITCLASS;
          } }
      if (DEPTHS >= val && DEPTHS <= val
&& SAND <= val && SILT <= val && CLAY
        <= val && AVAILK <= val &&
SLOPEID >= val && SLOPEID <= val)
      {
        if ( PH> val ||PH< val ||EC > val || EC <
val || ORGC > val || ORGC <val || EXCHBS >
        val || EXCHBS <val || AVAILP >val ||
AVAILP <val) { }
          else
          { LIMITID = LIMITCLASS;
            SUITID = SUITCLASS;
          } } }
```

TABLE 2: TRAINING DATASET

depth	sand	silt	clay	ph	oc	ec	exba	ap	ak	slope	Suit
A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	Unsuitable
A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	Highly suitable
A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	C1	Moderately suitable

The above rule is framed with the help of soil attributes and the Expert System is giving the suitability class and limitation class recommendation for the crop whether this soil is unsuitable, suitable or moderately suitable and also giving the limitation like poor drainage, no limitation, severe erosion and drought, moderate erosion etc. and this class have been used further for comparative study of classification technique using Data Mining Tool.

D. WEKA tool

WEKA is a data mining tool which is developed by the University of Waikato in New Zealand and thisis equipped with data mining algorithms. Data mining refers to extracting or mining the knowledge or information from the database or data warehouses. It uses machine learning, statistical and visualization techniques to discover and present the knowledge in a form, which is easily comprehensive to humans [7]. Data Mining Tools are used sophisticated, automated, algorithms to discover hidden patterns, correlations and relationships among the organizational data [2]. WEKA supports various data mining tasks, such as, data pre-processing, clustering, classification, regression, visualization, and feature selection [8], It also contains Association rule learner, Select Attributes and visualize. The algorithms can either be applied directly to a dataset or called from your own Java code. The workflow of WEKA as follows

Data ⟶ Pre-processing ⟶ Data Mining ⟶ Knowledge

Data mining steps in the knowledge discovery process are as follows:

1. Data cleaning- to remove noise and inconsistent data.
2. Data integration - multiple data sources may be combined
3. Data selection - where data relevant to the analysis task are retrieved from the database.
4. Data transformation – where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.
5. Data mining –An essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation –To identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation –Where visualization and knowledge representation techniques are used to present the mined knowledge to the user [2]. Here we are using the WEKA data mining tools classification algorithm and comparing that which technique is correctly classifying in the given soil dataset.

E. A comparison study of classification algorithm for soil suitability

Classification is an important data mining technique which also called supervised learning and it is using the train dataset here the classification of the soil is essential that to give the recommendation of land suitability for the particular crop like unsuitable, suitable or moderately suitable. In this research we have taken three classifier namely Bayes, Rules and trees in that Bayes classifier we have examined Naive Bayes classification algorithm, in rules classifier we have examined Decision Table classification algorithm and in trees classifier we have examined J48 classification algorithm. The purpose of the work is to find out the best classification algorithm among the Bayesian, Rules and trees classifier. Figure 1 show that the Classification algorithm's system architecture.

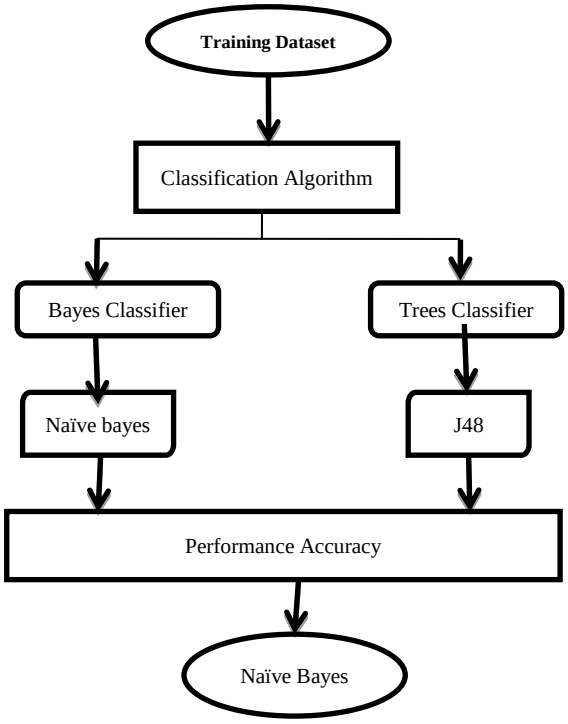

Fig 1: System Architecture

1.Naive Bayes

The Bayesian classifier is 'naive' in the sense that attributes are treated as though they are completely independent, and as if each attribute contributes equally to the model [5]. The Naive Bayes algorithm is based on conditional probabilities. All attributes of the data set are considered as independent and strong of each other [9]. An advantage of the NaiveBayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification [10]. Bayes theorem explanation given below

Bayes theorem:

1. P (C|X) = P (X|C) ·P(C) / P(X).
2. P(X) is constant for all classes.
3. P(C) = relative freq of class C samples c such that p is maximum=c Such that P (X|C) P(C) is maximum
4. Problem: computing P (X|C) is unfeasible! [11][12]

2. J48

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. It is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [13].

The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning [14].The above Rule1 shows that the extraction of classification rules from trees.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this Research, Two classification algorithm namely Naive Bayes and J48 were used and using WAIKA data mining tool we evaluated and compared on the basis of Time Accuracy, All Error rate, True positive Rate, False Positive Rate, Precision, Recall, F Measure, Receiver Operating Characteristics (ROC) Area and Kappa Statistics. Tenfold cross-validation was used in the experiment. The following tables show the accuracy measure of classification techniques.

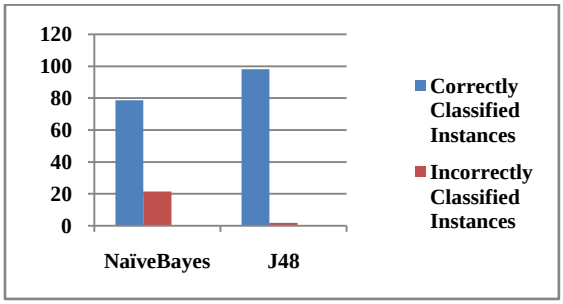

Fig 2: Accuracy measure for Classification Algorithms

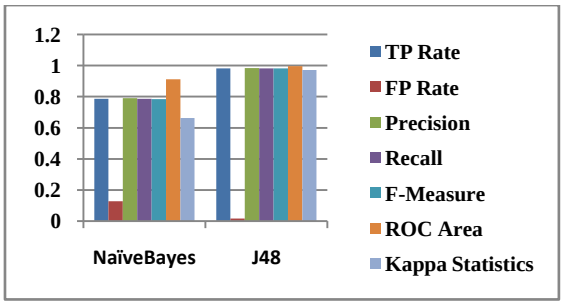

Fig 3: Performance measure for Classification Algorithms

Figure 2 shows that the accuracy measures and figure 3 shows that the performance measures of classification algorithms of Naïve Bayes and J48, the experiment performed on soil dataset by using WEKA Tool the best accuracy and performance of classification algorithm for this soil dataset is J48. This chart represent as given in table 3 which shows the correctly and incorrectly classified instances, TP Rate (true positive), FP Rate (false positive), Precision, Recall, F-Measure, ROC (Receiver Operating Characteristics) and Kappa statistics. J48 correctly classified is 98.2143% and Naïve Bayes correctly classified is 78.5714% hence J48 gives more classification accuracy.

TABLE 3: ACCURACY MEASURE AND COMPARISON OF NAÏVE BAYES AND J48 CLASSIFIER

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	TP Rate	FP Rate	Precision	Recall	F- Measure	ROC Area	Kappa Statistics
Naïve Bayes	88(78.5714 %)	24(21.4286 %)	0.786	0.128	0.791	0.786	0.785	0.911	0.6618
J48	110 (98.2143%)	2 (1.7857 %)	0.982	0.017	0.983	0.982	0.982	0.996	0.9711

TABLE 4: COMPARISON OF ERROR RATE, TIME TAKEN AND CONFUSION MATRIX OF NAÏVE BAYES AND J48 CLASSIFIER.

Algorithm	MAE	RMSE	RAE	RRSR	Time Taken	Confusion Matrix
Naïve Bayes	0.163	0.3333	39.3037	73.2254	0.01 Sec.	a b c 19 2 0 | 8 41 6 | 0 8 28 |
J48	0.0182	0.1158	4.3789	25.4481	0.03 Sec.	a b c 21 0 0| 0 55 0| 0 2 34|

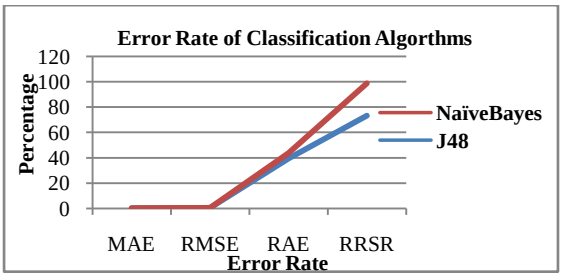

Fig 4: Error Rate of Classification Algorithms

Figure 4 represents the Mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error rate, with the help of the graph, it is observed that NaïveBayes algorithm reached the graph more error rate than the J48 classification algorithm. J48 algorithm performs well and it shows minimum error rate than the Naïve Bayes. This graph represent as given in table 4.

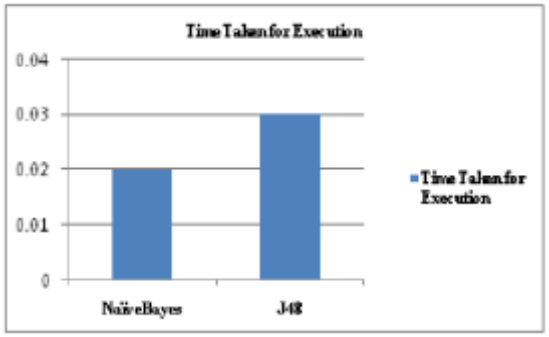

Fig 5: Execution Time of Classification Algorithms

Figure 5 shows that the time taken for execution process of NaïveBayes and J48, here Naïve Bayes performs with minimum period of time for execution than the J48 algorithms. But naïve Bayes accuracy measure and performance measure poor than the J48 and also error rate higher than the J48. This chart represented as given in table 4.

IV. CONCLUSION

In this research work classification algorithms two classifiers were used namely Bayes classifier's NaïveBayes and trees classifier's J48 algorithm and in this work focuses on to finding the best algorithm between two classifiers. Here the algorithm used to classify the soil dataset. From the result J48 given good accuracy, performance measure and minimum error rate than the NaïveBayes, but NaïveBayes classifies the data with minimum execution time. By analyzing the overall experimentation result of this soil dataset, it is concluded that J48 algorithm has produced the best classification performance then the NaïveBayes. Performance of the algorithm varies depending on the dataset. In future we could use the clustering technique in the same soil data set.

REFERENCES:

1. F. N. Afrati, A. Gionis, and H. Mannila. "Approximatinga Collection of Frequent Sets". *In Proc. 2004 ACMSIGKDD Int. Conf. Knowledge Discovery in Databases(KDD'04)*, pp. 12-19, Seattle, WA, Aug. 2004.
2. BalakrishnanM., Application of Data Mining Techniques in Agriculture,Training Manual, National Academy of Agricultural Research Management, Hyderabad.pp1-?
3. S. Muggleton, Inductive Acquisition of Expert Knowledge, Addison-Wesley, Reading, Mass, USA, 1990.
4. Halil Akıncı , Ays_e Yavuz Ozalp , Bulent Turgut, Agricultural land use suitability analysis using GIS and AHP technique, Computers and Electronics in Agriculture,vol.97, pp.71-82.
5. AN.Ganeshamurthy, R.Dinesh, N.Ravisankar, AK.Nair, SPS.Ahalwat, Land Resources of Andaman and Nicobar Islands, Central Agricultural Research Institute, (ICAR).
6. Say, N.P., Yucel, M., and Yilmazer, M., "A Computer-based System for Environmental Impact Assessment (EIA) Applications to Energy Power Stations in Turkey: CEDINFO", Journal of Energy Policy, Vol. 35, pp.6395-6401, 2007.
7. Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, 2nd ed., Morgan Kaufmann publishers, SanFrancisco, 2006.
8. Sunita B Aher, Lobo LMRJ, Data Mining in Educational System using Weka, International Conference on Emerging Technology Trends (ICETT), Proceedings published by International Journal of Computer Applications (IJCA) Number 3, 2011, pp-20-25.
9. D. Pedro and M. Pazzani "On the optimality of the simple Bayesian classifier under zero-one loss". *Machine Learning*, 29:103–137, 1997.
10. S. Vijayarani, Mr.S.Dhayanand, Data mining classification algorithms for kidney disease prediction, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015, pp.13-25.
11. Uffe B. Kjærulff, Anders L. Madsen, Probabilistic Networks — an Introduction to Bayesian Networks and Influence Diagrams, May 2005.
12. Zhang H.; Su J.; (2004) "Naive Bayesian classifiers for ranking". Paper appeared in ECML2004 15th European Conference on Machine Learning, Pisa, Italy
13. http://www.c4.5-Wikipedia, the free encyclopedia.htm accessed on 16/12/2010.
14. J. R. Quinlan "C4.5: programs for machine learning" *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA, 1993.
15. Say, N.P., Yucel, M., and Yilmazer, M., 2007,"A Computer-based System for Environmental Impact Assessment (EIA) Applications to Energy Power Stations in Turkey: CEDINFO", Journal of Energy Policy, Vol. 35, pp.6395-6401.
16. Ramesh Vamanan., K.Ramar. (2011), Classification of Agricultural Land Soils of Data Mining Approach, International Science on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol.3 No. 1 Jan 2011, pp. 379-383.
17. S. Muggleton, Inductive Acquisition of Expert Knowledge, Addison-Wesley, Reading, Mass, USA, 1990.
18. Swartout W., and Moore J. (1993), Explanation in Second Generation Expert Systems. In David J., Krivine, J-P., and Simmons R., Editors, Second Generation Expert Systems, pp. 543-585. Springer Verlag.
19. Lemmon, H. (1986), COMAX : An Expert System for Cotton Crop Management, Science 233 (4759), pp. 29-33.
20. Charu C. Agarwal, Data Classification Algorithm and Applications (An Introduction to Data Classification),IBM T. J. Watson Research CenterYorktown Heights, New York, USA.