2016 IEEE Eighth International Conference on Advanced Computing (ICoAC)

Crop Recommendation System for Precision Agriculture

S.Pudumalar*, E.Ramanujam*,
R.Harine Rajashree‡, C.Kavya‡, T.Kiruthika‡, J.Nisha‡.

*** Assistant professor, Department of Information Technology, Thiagarajar College of Engineering.**
[spmit , erit]@gmail.com
‡ Department of Information Technology, Thiagarajar College of Engineering,
[harinerajashree, ckavya005, kiruthigomi, jaffernisha96]@gmail.com

Abstract: ***Data mining is the practice of examining and deriving purposeful information from the data. Data mining finds its application in various fields like finance, retail, medicine, agriculture etc. Data mining in agriculture is used for analyzing the various biotic and abiotic factors. Agriculture in India plays a predominant role in economy and employment. The common problem existing among the Indian farmers are they don't choose the right crop based on their soil requirements. Due to this they face a serious setback in productivity. This problem of the farmers has been addressed through precision agriculture. Precision agriculture is a modern farming technique that uses research data of soil characteristics, soil types, crop yield data collection and suggests the farmers the right crop based on their site-specific parameters. This reduces the wrong choice on a crop and increase in productivity. In this paper, this problem is solved by proposing a recommendation system through an ensemble model with majority voting technique using Random tree, CHAID, K-Nearest Neighbor and Naive Bayes as learners to recommend a crop for the site specific parameters with high accuracy and efficiency.***

Keywords— Precision agriculture, Recommendation system, Ensemble model, Majority Voting technique, Random tree, CHAID, K-Nearest Neighbor and Naive Bayes.

I. INTRODUCTION

India is one among the oldest countries which is still practicing agriculture. But in recent times the trends in agriculture has drastically evolved due to globalization. Various factors have affected the health of agriculture in India. Many new technologies have been evolved to regain the health. One such technique is precision agriculture. Precision agriculture is budding in India .Precision agriculture is the technology of "site-specific" farming. It has provided us with the advantage of efficient input, output and better decisions regarding farming. Although precision agriculture has delivered better improvements it is still facing certain issues. There exist many systems which propose the inputs for a particular farming land. Systems propose crops, fertilizers and even farming techniques. Recommendation of crops is one major domain in precision agriculture. Recommendation of crops is dependent on various parameters. Precision agriculture aims in identifying these parameters in a site-specific manner in order to resolve issues regarding crop selection. The "site-specific" technique has improved the results yet there is a need to supervise the results of such systems. Not all precision agriculture systems provide accurate results. But in agriculture it is important that the recommendations made are accurate and precise because incase of errors it may lead to heavy material and capital loss. Many research works is being carried out, in order to attain an accurate and efficient model for crop prediction. Ensembling is one such technique that is included in such research works. Among these various machine learning techniques that are being used in this field; this paper proposes a system that uses the voting method to build an efficient and accurate model.

II. LITERATURE SURVEY

The paper [1] states the requirements and planning needed for developing a software model for precision farming is discussed. It deeply analyses the basics of precision farming. The author's start from the basics of precision farming and move towards developing a model that would support it. This paper describes a model that applies Precision Agriculture (PA) principles to small, open farms at the individual farmer and crop level, to affect a degree of control over variability. The comprehensive objective of the model is to deliver direct advisory services to even the smallest farmer at the level of his/her smallest plot of crop, using the most accessible technologies such as SMS and email. This model has been designed for the scenario in Kerala State where the average holding size is much lower than most of India. Hence this model can be deployed elsewhere in India only with minor modifications. The paper [2] makes a comparative study of classification algorithms and their performance in yield prediction in precision agriculture. These algorithms are implemented in a data set collected for several years in yield prediction on soya bean crop. The algorithms used for yield prediction in this paper are Support Vector Machine, Random Forest,

978-1-5090-5888-4/16/$31.00@2016 IEEE

Neural Network, REPTree, Bagging, and Bayes. The conclusion drawn at the end is that bagging is the best algorithm for yield prediction among the above stated algorithms since the error deviation in bagging is minimum with a mean absolute error of 18985. 7864. The paper [3] states the necessity for crop yield prediction and its help in a nation's strategic policy making in agriculture. A framework eXtensible Crop Yield Prediction Framework (XCYPF) is developed. It facilitates flexible inclusion of various techniques towards crop yield prediction. A tool was also developed that would help people to predict crop yield for various crops with dependant and independent variables.

The paper [4] states the usage of agricultural data with data mining and visual data mining techniques are depicted. This paper reduces the high dimensional agricultural data to smaller size to acquire useful knowledge related to yield, input application(like fertilizers).The techniques used is Self-organizing maps and multi-dimensional scaling techniques (Sammon's mapping) to reduce the data. The conclusion derived is that Self-organizing maps is suitable when dataset is large and Sammon's maping is suitable when data set is small. The paper [5] depicts the importance of crop selection and the factors deciding the crop selection like production rate, market price and government policies are discussed. This paper proposes a Crop Selection Method (CSM) which solves the crop selection problem and improves net yield rate of the crop. It suggests a series of crop to be selected over a season considering factors like weather, soil type, water density, crop type. The predicted value of influential parameters determines the accuracy of CSM. Hence there exists a need to include a prediction method with improved accuracy and performance. Data mining techniques in paper [6] are used to estimate the crop yield for cereal crops in major districts of Bangladesh. The methodology comprises of two parts namely Clustering (for creating district clusters) and Classification using k-NN (k-nearest neighbor), Linear Regression, (ANN) artificial neural network in rapid miner tool. The accuracy of prediction lies in the range of 90-95. The data set included 5 environmental variables, 3 biotic variables and 2 area related variables to determine the crop yield in different districts. The paper proposed a future work of geospatial analysis to improve accuracy.

The paper [7] aims to solve the crucial problem of selecting the classifiers for the ensemble learning. A method to select a best classifier set from a pool of classifiers has been proposed. The proposal aims to achieve higher accuracy and performance. A method called SAD was proposed based on accuracy and classification performance. Using Q statistics, the dependency between most relevant and accurate classifiers is identified. The classifiers which were not chosen were combined to form the ensemble. This measure is supposed to ensure higher performance and diversity of the ensemble. Various methods such as SA (Selection by Accuracy), SAD (Selection by accuracy and Diversity) and NS (No selection) algorithm were identified. Finally it is inferred that SAD works better than others. The paper [8] proposes various classification methods to classify the liver disease data set. The paper emphasizes the need for accuracy because it depends on the dataset and the learning algorithm. Classification algorithms such as J48, Naive Bayes, ANN, ZeroR, 1BK and VFI were used to classify these diseases and compare the effectiveness, correction rate among them. The performance of the models where compared with accuracy and computational time. It was concluded that all the classifiers except naive bayes showed improved predictive performance. Multilayer perceptron show the highest accuracy among the proposed algorithms. The paper [9] tries to solve the problem of food insecurity in Egypt. It proposes a framework which would predict the production, and import for that particular year. It uses Artificial Neural Networks along with Multi-layer perceptron in WEKA to build the prediction. At the end of the process we would be able to visualize the amount of production import, need and availability. Therefore it would help to make decisions on whether food has to be further imported or not. The soil datasets in paper [10] are analyzed and a category is predicted. From the predicted soil category the crop yield is identified as a Classification rule. Naïve Bayes and k-Nearest Neighbor algorithms are used for crop yield prediction. The future work stated is to create efficient models using various classification techniques such as support vector machine, principal component analysis.

III. METHODOLOGY

3.1 Dataset Collection

The dataset comprising the soil specific attributes which are collected for Madurai district tested at soil testing lab, Madurai, Tamil Nadu, India. In addition, similar online sources of general crop data were also used. The crops considered in our model include millet, groundnut, pulses, cotton, vegetables, banana, paddy, sorghum, sugarcane, coriander. Figure 1 gives an analysis of the dataset. The number of instances of each crop available in the training dataset is depicted. The attributes considered where Depth, Texture, Ph, Soil Color, Permeability, Drainage, Water holding and Erosion.

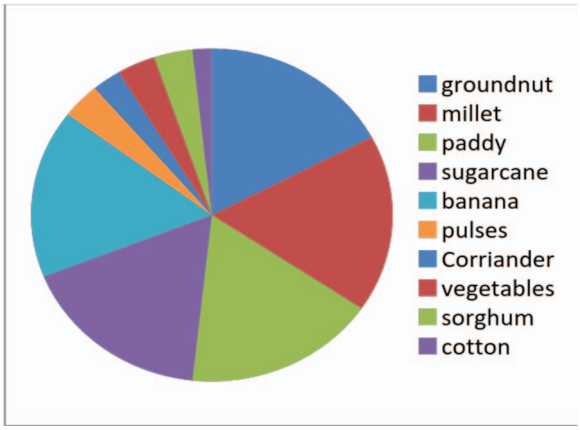

Fig 1 Analysis of dataset with respect to crops(Training data)

The above stated parameters of soil play a major role in the crop's ability to extract water and nutrients from the soil. For crop growth to their fullest potential, the soil must provide a satisfactory environment for it. Soil is the anchor of the roots. The water holding capacity determines the crop's ability to absorb nutrients and other nutrients that are changed into ions, which is the form that the plant can use. Texture determines how porous the soil is and the comfort of air and water movement which is essential to prevent the plants from becoming waterlogged. Soil texture which affects the soil's ability to hold onto nutrients. The level of acidity or alkalinity (Ph) is a master variable which affects the availability of soil nutrients. The activity of micro-organisms present in the soil and also the level of exchangeable aluminum can be affected by PH. The water holding and drainage determine the penetration of roots. Hence for the following reasons the above stated parameters are considered for choosing a crop.

3.2 Crop Prediction using ensembling technique:

Ensemble is a data mining model also known as the Committee Methods or Model Combiners, that combine the power of multiple models to acquire greater prediction, efficiency than any of its models could achieve alone. In our system, we use one of the most familiar ensembling technique called Majority Voting technique .In the voting technique any number of base learners can be used. There has to be at least two base learners. The learners are chosen in a way that they are competent to each other yet being complimentary also. Higher the competition higher is the chance of better prediction. But it is necessary for the learners to be complimentary because when one or few members make an error, the probability of the remaining members correcting this error would be high. Each learner builds itself into a model. The model gets trained using the training data set provided. When a new sample has to be classified, each model predicts the class on its own. Finally, the class which is predicted by majority of the learners is voted to be the class label of the new sample. This method is implemented in Rapid miner tool (figure 2, 3, 4, 5) depicts the process implemented in rapid miner.

3.3 Learners Used in the Model:

3.3.1 RANDOM TREE:

Random tree [11] is similar to that of a decision tree .But it differs from random tree in a way that for each split only a random subset of attributes is available. Random trees can be built for both nominal and numerical data. The Random Tree is similar to C4.5 or CART but it varies in the fact that before it is applied for training it selects only a random subset of attributes. At each node it considers K randomly chosen attributes. The subset ratio parameter specifies the size of the subset .

3.3.2 CHAID:

CHAID [13] is a type of decision tree technique, which is based upon adjusted significance testing. CHAID stands for **Ch**i-squared **A**utomatic **I**nteraction **D**etection. CHAID is also similar to decision tree but instead of information gain or gain ratio it uses a chi-squared based criterion. CHAID has many advantages.Such advantages include highly visual and interpretable results because it uses multiway splits by default.

3.3.3 K-NEAREST NEIGHBOR:

K-Nearest Neighbor [15] can be used for both classification and regression. K-Nearest Neighbors is a non-complex algorithm which stores all the available cases and classifies new cases based on some similarity measure .The sample set is classified based upon the "closeness" that is the distance measure such as Euclidean distance or Manhattan distance.

3.3.4NAÏVE BAYES:

Naive Bayes [14] classifier is a simple probabilistic classifier which works based on applying Bayes' theorem (from Bayesian statistics) with strong naive independence assumptions. Naive Bayes is a technique for constructing classifier models which assign class labels to problem instances which are represented as vectors of feature values, where the class labels are drawn from some finite set. It is not just a single algorithm for training such classifiers, but a family of algorithms based on a common principle.All naive Bayes classifiers assumes that the value of a particular feature is independent of the value of any other feature, given the class variable.

These Learners predict the class label for each of the training data set. The class label that is predicted by the majority of the models is voted through the majority voting technique and the class label of the training data set is decided. From the ensembled models the rules are generated.

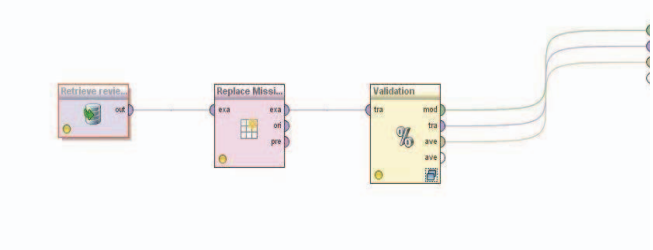

Fig 2: Illustrates the entire process work flow.

It shows three operators namely retrieve, replace missing values, Validation. The retrieve operator retrieves the dataset that is being uploaded in the tool. The replace the missing values operator replaces missing values if any. Replacement can be done by four methods namely minimum, maximum, average and zero. In order to estimate the statistical performance of a learning operator a cross-validation is performed by the validation operator.

Fig 3: Illustrates the sub-process of cross validation operator.

The training process consists of the voting operator which is the technique that we propose for better results. On the testing sub process lies the apply model and performance operators which evaluate the correctness of the model.

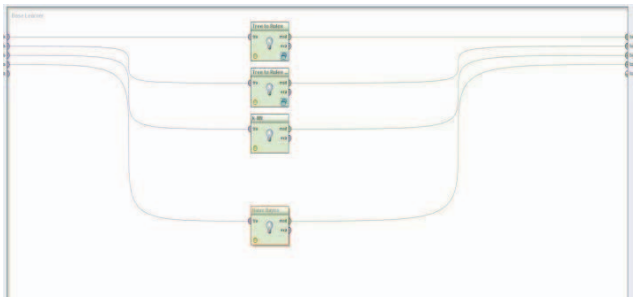

Fig 4: Illustrates the base learners which lie under the vote operator.

It consists of four machine learners namely Naïve bayes, K-Nearest Neighbor and CHAID and Random tree. The operators corresponding to each learner is positioned. The operator performs the classification correspondingly. The tree to rules operator is used to induce rules directly from the CHAID and random tree.

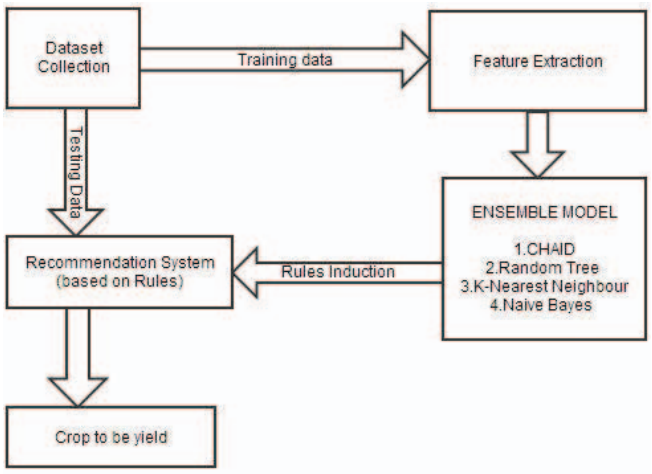

Fig 5 depicts the overall methodology of proposed system.

3.4 Rules induced from the Model:

The rule below demonstrates an example of the proposed recommendation system.
IF ph is mild alkaline
AND depth is above 100
AND water holding capacity is LOW
AND drainage is moderately well
AND erosion is moderate
THEN PADDY

The IF part of the rule states the soil specifications needed for the cultivation of the recommended crop which is specified in the THEN PART of the rule.

IV. RESULTS AND DISCUSSION

The prediction accuracy of the model accounts to 88%.

The rules induced from the random tree model and CHAID model are shown in the figure 6, 7 respectively. The rules are generated in form of if-then rules where the then part specifies the class label. The classification of every

training set is precisely shown.

The rules (figure 7, 8) generated from the above model is used to develop a RECOMMENDATION SYSTEM. This is achieved by creating a GUI. The GUI (figure 8) is deployed as a web portal. The model which is trained with the training data set is tested with inputs from the user from the portal. The scripting done will respond to any test case suggesting a crop. If the test case doesn't match any of the predictions then a no match output is produced.

Fig 6 Rules induced from random tree model.

Fig 7 Rules induced from the CHAID model

Fig 8 User Interface for the recommendation system

V. CONCLUSION

India is a nation in which agriculture plays a prime role. In prosperity of the farmers, prospers the nation. Thus our work would help farmers in sowing the right seed based

on soil requirements to increase productivity and acquire profit out of such a technique. Thus the farmer's can plant the right crop increasing his yield and also increasing the overall productivity of the nation. Our future work is aimed at an improved data set with large number of attributes and also implements yield prediction.

VI. REFERENCES

[1] Satish Babu (2013), 'A Software Model for Precision Agriculture for Small and Marginal Farmers', at the International Centre for Free and Open Source Software (ICFOSS) Trivandrum, India.

[2] Anshal Savla, Parul Dhawan, Himtanaya Bhadada, Nivedita Israni, Alisha Mandholia , Sanya Bhardwaj (2015), 'Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture', Innovations in Information,Embedded and Communication systems (ICIIECS).

[3] Aakunuri Manjula, Dr.G .Narsimha (2015), 'XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction' , Conference on Intelligent Systems and Control (ISCO)

[4] Yash Sanghvi, Harsh Gupta, Harmish Doshi, Divya Koli, Amogh Ansh Divya Koli, Umang Gupta (2015), 'Comparison of Self Organizing Maps and Sammon's Mapping on agricultural datasets for precision agriculture', International Conference on Innovations in Information,Embedded and Communication systems (ICIIECS).

[5] Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh (2015), 'Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique', International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM).

[6] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) , 'Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh' , (SNPD) IEEE/ACIS International Conference.

[7] Liying Yang (2011), 'Classifiers selection for ensemble learning based on accuracy and diversity' Published by Elsevier Ltd. Selection and/or peer-review under responsibility of [CEIS].

[8] Tapas Ranjan Baitharua, Subhendu Kumar Panib (2016), 'Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset' International Conference on Computational Modeling and Security (CMS).

[9] Aymen E Khedr, Mona Kadry, Ghada Walid (2015), 'Proposed Framework for Implementing Data Mining Techniques to Enhance Decisions in Agriculture Sector Applied Case on Food Security Information Center Ministry of Agriculture, Egypt', International Conference on Communications, management, and Information technology (ICCMIT').

[10] Monali Paul, Santosh K. Vishwakarma, Ashok Verma (2015), 'Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach', International Conference on Computational Intelligence and Communication Networks.

[11] Bhuvana, Dr.C.Yamini (2015), 'Survey on Classification Algorithms in Data mining.' International Conference on Recent Advances in Engineering Science and Management.

[12] Roshani Ade, P.R.Deshmukh (2014), 'Efficient Knowledge Transformation System Using Pair of Classifiers for Prediction of Students Career Choice', International Conference on Information and Communication Technologies (ICICT).

[13] Ahmed Hamza Osman, Naomie Salim (2013) , 'An improvised semantic plagiarism detection scheme based on Chi- squared automatic interaction detection' , Computing,Electrical and Electronics Engineering (ICCEEE) International Conference.

[14] Saso Karakatic, Marjan Hericko and Vili Podgorelec (2015), 'Weighting and sampling data for individual classifiers and bagging with genetic algorithms' International Joint Conference and Computational Intelligence(IJCCI).

[15] Yue Zhang, Jianxia Chen, Oin Fang, Zhiwei Ye (2016), 'Fault analysis and Prediction of transmission line based on Fuzzy K-Nearest Neighbor algorithm' International Conference on Natural Computation , Fuzzy systems and Knowledge discovery(ICNC-FSKD).