Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique

Nidhi H Kulkarni	Dr. G N Srinivasan	Dr. B M Sagar	Dr.N K Cauvery
MTech IT, Dept. of ISE	*Professor, Dept. of ISE*	*HOD, Dept. of ISE*	*Professor, Dept. of ISE*
RV College of Engineering	*RV College of Engineering*	*RV College of Engineering*	*RV College of Engineering*
Bangalore, India	*Bangalore, India*	*Bangalore India*	*Bangalore, India*
[1]nidhihk.sit17@rvce.edu.in	srinivasangn@rvce.edu.in	sagarbm@rvce.edu.in	cauverynk@rvce.edu.in

***Abstract* - Agriculture plays a predominant role in the economic growth and development of the country. The major and serious setback in the crop productivity is that the farmers do not choose the right crop for cultivation. In order to improve the crop productivity, a crop recommendation system is to be developed that uses the ensembling technique of machine learning. The ensembling technique is used to build a model that combines the predictions of multiple machine learning models together to recommend the right crop based on the soil specific type and characteristics with high accuracy. The independent base learners used in the ensemble model are Random Forest, Naive Bayes, and Linear SVM. Each classifier provides its own set of class labels with an acceptable accuracy. The class labels of individual base learners are combined using the majority voting technique. The crop recommendation system classifies the input soil dataset into the recommendable crop type, Kharif and Rabi. The dataset comprises of the soil specific physical and chemical characteristics in addition to the climatic conditions such as average rainfall and the surface temperature samples. The average classification accuracy obtained by combining the independent base learners is 99.91%.**

***Keywords* - Ensemble, Majority-Voting, Naive-Bayes, soil, crop-recommendation**

I. OBJECTIVES

- Design a recommendation system for accurate crop selection based on the various soil, rainfall and surface temperature parameters.
- To improve crop productivity by providing predictions of high accuracy and efficiency through the ensembling technique
- To reduce the wrong choice on a crop by application of principles of precision agriculture.

II. INTRODUCTION

India is one of the established nations that has agriculture as its primary source of income. Agriculture is one such domain that contributes only around 14% to the GDP but has a considerate amount of impact on the Indian economy. The conventional agricultural practices and techniques are posing a lot of issues in terms of efficiency, cost-effectiveness and resource utilization. There is a necessity of better techniques that can improve the standard of living of the farmers too. Over the years due to globalization, agriculture has evolved by adapting the latest technologies and techniques for a better standard of living. Among the technologies and techniques , precision agriculture is one budding technology in the field of agriculture. Precision agriculture mainly focuses on site-specific farming [1]. Crop recommendation is a prima arena in precision agriculture. Crop recommendation relies on multiple parameters, for which precision agriculture practices help in identifying the parameters thereby facilitating better crop selection.

Agricultural domain has imbibed the machine-learning algorithm to produce efficient, cost-effective solutions to the difficulties faced by the farmers. Researchers can utilize PC simulations to lead early tests to assess how an assortment may perform when looked with changed sub-atmosphere, soil composes, climate designs, and different variables. Researchers in present day agribusiness are trying their speculations at a more prominent scale and making considerably more precise, ongoing forecasts..

III. LITERATURE SURVEY

The paper [2] mainly throws light on the implementation of a crop prediction system based on sensor networks that has been developed using IoT. Soil testing labs take a considerable amount of time in providing the results of the submitted soil samples. Hence the system claims that it helps the farmers to get a better crop prediction without any delay in the waiting period [2]. In the paper, the authors have mainly focused on analyzing the N(Nitrogen), P(Phosphorus), K(Potassium) contents in the soil sample collected for survey. The proposed method in the paper efficiently estimates the soil nutrients based on the data fetched by the sensor network. This enables in predicting the apt crop for that soil under test. The farmers need to enlist their NPK sensor with the fundamental server. The NPK extract the supplement level from the soil sample and refresh this information to the primary server through the raspberry pi unit. In view of the readings got from the calculation makes predictions on the basis of the recorded information. The major shortfalls in this implementation were the inefficiency of the crop prediction algorithm and major focus on the data collection through NPK sensor which has high range of fluctuations.

The paper [3] presents a vivid representation of a Crop Selection Method which aims to solve the crop selection issue and enhances the net yield of the harvest[3]. The authors have proposed a strategy that proposes a scope of crops to be chosen over a season by keeping into thought the essential elements like the climate, soil composition, water density, crop category. The estimated value of the factors that are highly influential determine the precision of Crop Selection Method. The technique taken into account in the paper is the method of crop sequencing. A categorization of the crops is done in four divisions namely seasonal, whole year, short-time plantation, and long-time. The grouping of the crops from each category is selected in a sequence for the crop cultivation. Hence there is a necessity for a prediction technique with upgraded precision and performance. In addition to this, there is a compulsion of selecting atleast one crop from the category which serves as a major setback.

The paper presents a comprehensive analysis of the soil characteristics and behaviour, and using this the authors have introduced a method of foreseeing the crop utilization using the information mining approach [4]. The paper highlights the focus on improvising the yield of the crops rather than the crop selection technique. The soil datasets are taken as inputs and analyzed. Based on the thorough analysis, the soil is arranged into low, medium, and high classes by making best use of the procedures that are used in data mining. As a result of such categorization, the crop yield has been predicted using Naive-Bayes and k-Nearest algorithm. Here the crop yield anticipated is formalized as a classification rule. Although there are many positive highlights with respect to improvising the crop yield, there are a few drawbacks. The major drawback is that problem eradication at the source is not focussed, rather more focus is only on increasing the crop yield. The further setbacks are that the method has been experimented on very less data mining algorithms. There is a necessity to include multiple data mining algorithms. The dataset size used is very small as stated by the authors themselves, as huge datasets invite large amount of complexities.

The paper provides clear details regarding the new framework designed by the authors namely eXtensible Crop Yield Prediction Framework (XCYPF)[5]. This framework is mainly designed for predicting the crop yield. The framework claims to provide crop selection, selection of dependent and independent variables ,datasets required for crop yield forecast. The framework has been tested for rice and sugarcane crops. The framework professes to have been coordinated with a management information system providing expertise in the field of precision agriculture. The framework that has been designed in the paper works exclusively for rice and sugarcane crops.

A special concern has always been shown in case of how to increase the productivity of the crops. There have been various methods designed and other improvised techniques that are used to boost the yield of the crops. Solving the problem at the source immediately eradicates the issue. Hence deciding the perfect solution for the crop to be cultivated will lead to better crop productivity, and in turn boosting the economy of the country.

IV. METHODOLOGY

A brief step by step procedure of designing the crop recommendation system is explained as follows:

Step 1: Input

The input dataset is a comma separated values file containing the soil dataset, which has to be subjected to preprocessing.

Step 2: Preprocessing of input data

Input dataset is subject to various preprocessing techniques such as filling of missing values, encoding of categorical data and scaling of values in the appropriate range

Step 3: Splitting into training and testing dataset

The preprocessed dataset is then split into training and testing dataset based on the specified split ratio. The split ratio considered in the proposed work is 75:25, which means 75% of the dataset is used for the training the ensemble model and the rest 25% is used as test dataset.

Step 4: Building individual classifiers on the training dataset

The training dataset is fed to each of the independent base learners and the individual classifiers are built using the training dataset.

Step 5: Testing the data on each of the classifiers

The testing dataset is applied on each of the classifiers, and the individual class labels are obtained.

Step 6: Ensembling the individual classifier output using Majority Voting Technique.

The class labels obtained from the individual classifiers is subjected to the majority voting technique to get an ensembled class label as the final prediction

A. Ensemble Framework

The ensemble framework is of utmost importance. The ensemble framework is explained as follows.
Before diving into the details of the ensemble framework, the actual meaning of ensembling and the reason for its usage.

Ensembling is a technique of building a prescient model by incorporating multiple models. The main reason for using an ensemble framework is that it provides a classifier that outperforms each of the individual classifiers.

Ensembling uses two frameworks, dependent framework and independent framework: In the dependent framework, the yield of one classifier is utilized in the development of the following classifier. The second method involves independent method, that is each classifier produces a class label in an independent fashion. All the classifiers work in a parallelized manner. The output of one classifier is independent of the other. The independent method has been used in the proposed work since it reduces the execution time.

The ensemble framework comprises of the basic components that are explained as follows:

a) Training set – A labelled set of instances that is utilized in training the ensemble model. Each example in the training set is potrayed as attribute-value vectors.

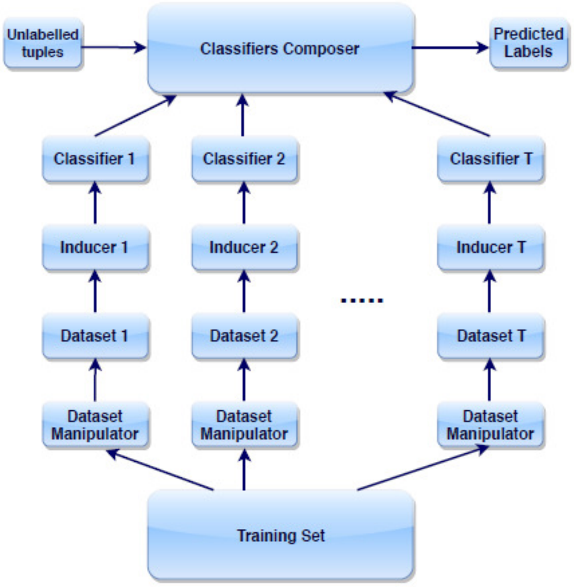
b) Base Inducers - Inducer is an inducing algorithm that produces a classifier on feeding a labelled set of instances to the inducer as input. The resulting classifier gives a distinctive potrayal of the generalized relationship between the input attribute and the target attribute.

$$\mathbf{M = I(S)}$$

M => Classifier , I => Inducer, S => Training set

c) Diversity generator – Generation of diverse classifiers.

d) Combiner – Responsible for combining the class labels obtained from the individual classifiers [21].



Advantage of using the independent framework of ensembling:

a) Enhances the prescient intensity of the classifiers.
b) Decreases the aggregate execution time.

Random Forest, Naive Bayes and Linear SVM are the three independent base learners that have been used to build the ensemble model. Each of the algorithms have been concisely explained as follows.

B. Random Forest

A Random Forest is a classifier comprising of accumulation of tree-organized classifiers where independent random vectors are disseminated indistinguishably and each tree make a unit choice for the most mainstream class at input x. A random vector is produced which is autonomous of the past arbitrary vectors with same dissemination and a tree is created by utilizing the training set. The main advantages of considering the Random Forest algorithm is that it provides better accuracy, vigorous to the outliers, quicker than bagging and boosting, basic and easy to parallelized.

C. Naive Bayes

Naive Bayes is a classification algorithm for binary and multi-class classification problems. When binary or categorically input values are provided, Naive Bayes method is very easy to understand. In Naive Bayes, a Naive Bayes classifier assumes that the presence of a particular feature in a class is not at all related to the presence of any other feature. Hence the name 'Naive Bayes'. The Naive Bayes classifier depends on the Bayes hypothesis and this technique is useful in the cases where the dimensionality of the sources of information is high. Naive Bayes has multiple applications such as for making predictions in real time, to predict the probability of multiple classes of target attribute, spam-filtering, and coupled with collaborative filtering helps to build recommendation systems.

Initially the probability of each attribute in the dataset is to be calculated which is also known as class probability.
The conditional probability gives the conditional probability of each input value given each class value.

D. Linear SVM

Linear SVM is the current machine learning algorithm that is the quickest to solve the multiclass classification problems. Linear SVM is linearly scalable, this means that the SVM model is created in a CPU time that scales linearly with the training dataset size. The main advantage of Linear SVM is that it works well with extremely large datasets along with eminent accuracy. Linear SVM also provides better performance on working with multidimensional data[5].

E. Majority Voting

Majority Voting technique is one of the techniques of combining the class labels obtained as a result of the independent classifiers. In this combining plan, a classification of an unlabelled instance is performed by the class that gets the most astounding number of votes. This technique is otherwise called the plurality vote This methodology has been utilized much of the time as a consolidating strategy for looking at recently proposed strategies. This is the most frequently used combiner. Mathematically, it can be expressed as:

$$class(x) = \underset{c_i \in dom(y)}{\arg\max} \left(\sum_k g\left(y_k(x), c_i\right) \right)$$

where y_k (x) is the classification of the kth classifier and g(y, c) is an indicator function defined as:

$$g\,(y, c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases}$$

V. DATASET DETAILS

The dataset considered for usage in the given proposed work is a soil dataset primarily comprising of soil physical and chemical properties, along with the climatic details. An open source dataset is obtained from the data repository site of the Government of India, **data.gov.in**

The dataset size is 5MB containing 9000 rows and 15 attributes that are of prime importance.
The crops considered are Cotton, Sugarcane, Rice, Wheat. The dataset attributes that are of prime importance are

- Soil Type
- pH value of the soil
- NPK content of the soil
- Porosity of the soil
- Average rainfall
- Surface temperature
- Sowing season

VI. RESULTS

The collected data is initially subjected to preprocessing. Post dataset preprocessing, the dataset is divided into training set and test set samples. Out of the 9000 samples, 6750 samples are used as training samples, and the rest 2250 samples are used as test samples. Each of the sample is trained and tested on the Random Forest, Naive Bayes and the Linear SVM algorithms. The average accuracy of crop classification into Kharif and Rabi crops is 99.91%.

CONCLUSION

A crop recommendation system has been designed that takes into consideration the soil dataset with respect to the four crops Rice, Cotton, Sugarcane, Wheat. The soil dataset is first preprocessed and then the ensembling technique performs a critical function in the classification of the four crops. The individual base learners used in the ensemble model are Random Forest, Naive Bayes, and Linear SVM. Majority

Voting Technique has been used as the combination method to provide the best accuracy.

The accuracy obtained using the ensembling technique is 99.91%. Hence, the proposed work provides a helping hand to the farmer in the accurate selection of the crop for cultivation. This creates an exponential gain in the crop productivity which in turn boosts the economy of the country.

REFERENCES

[1] S.Pudumalar , E.Ramanujam , "*Crop Recommendation System for Precision Agriculture*", 2016, IEEE Eighth International Conference on Advanced Computing (ICoAC)

[2] Lokesh.K,Shakti.J, Sneha Wilson, Tharini.M.S, "*Automated crop prediction based on efficient soil nutrient estimation using sensor network*", July 2016,National Conference on Product Design (NCPD 2016)

[3] 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).

[4] 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.

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

[6] 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).

[7] D Ramesh , B Vishnu Vardhan, "*Data mining technique and applications to agriculture yield data*", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September2013 .

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

[9] M.Soundarya, R.Balakrishnan," *Survey on Classification Techniques in Data mining*", International Journal ofAdvanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[10] Miss. Snehal, S.Dahikar, Dr.SandeepV.Rode, "*Agricultural Crop Yield Prediction Using Artificial Neural Network Approach*". International Journal of Innovative Reasearch in Electrical, Electronic, Instrumentation and Control Engineering, Vol. 2, Issue 1, January 2014.

[11] Thoranin Sujjaviriyasup, Komkrit Pitiruek, "*Agricultural Product Fore- casting Using Machine Learning Approach*". Int. Journal of Math. Analysis, Vol. 7, no. 38, 1869 1875, 2013.

[12] Luke Bomn, James V. Zidek. (2012). "*Efficient stabilization of crop yield prediction in the Canadian Prairies*",.Elsevier , P223-232.

[13] Raorane A.A.I, Kulkarni R.V.2. (2012). "*Data Mining: An effective tool for yield estimation in the agricultural sector*". UETTCS. 1 (2), P75-79.

[14] M.C.S.Geetha," *Implementation of Association Rule Mining for different soil types in Agriculture*", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015.

[15] T.R.Lekha, "*Efficient Crop Yield and Pesticide Prediction for Improving Agricultural Economy using Data Mining Techniques*", International Journal of Modern Trends in Engineering and Science, Vol-03,Issue-10, 2016

[16] Shweta Taneja, Rashmi Arora, Savneet Kaur, "*Mining of Soil Data Using Unsupervised Learning Technique*", International Journal of

Applied Engineering Research,ISSN 0973- 4562 Vol. 7 No.11, 2012.

[17] Washington Okori, Joseph Obua, "*Machine Learning Classification Technique for Famine Prediction*". Proceedings of the World Congress on Engineering 2011 Vol II WCE 2011, July 6 - 8, London, U.K, 2011.

[18] Liying Yang (2011), '*Classifiers selection for ensemble learning based on accuracy and diversity*', Elsevier Ltd.

[19] Aymen E Khedr, Mona Kadry, GhadaWalid (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, andInformation technology (ICCMIT').

[20] 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).

[21] LiorRokach, "Ensemble-based classifiers", ArtifIntell Rev (2010) 33:1–39DOI 10.1