AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms

Zeel Doshi
Department of Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
zeelrdoshi@gmail.com

Subhash Nadkarni
Department of Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
sub8896@gmail.com

Rashi Agrawal
Department of Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
agrawal.rashi169@gmail.com

Prof. Neepa Shah
Head of Information Technology Department
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
neepa.shah@djsce.ac.in

***Abstract*— Agriculture is a major contributor to the Indian economy. The mainstream Indian population depends either explicitly or implicitly on agriculture for their livelihood. It is, thus, irrefutable that agriculture plays a vital role in the country. A vast majority of the Indian farmers believe in depending on their intuition to decide which crop to sow in a particular season. They find comfort in simply following the ancestral farming patterns and norms without realizing the fact that crop output is circumstantial, depending heavily on the present-day weather and soil conditions. However, a single farmer cannot be expected to take into account all the innumerable factors that contribute to crop growth before reaching a consensus about which one to grow. A single misguided or imprudent decision by the farmer can have undesirable ramifications on both himself as well as the agricultural economy of the region. A combination of Big Data Analytics and Machine Learning can effectively help alleviate this issue. In this paper, we present an intelligent system, called AgroConsultant, which intends to assist the Indian farmers in making an informed decision about which crop to grow depending on the sowing season, his farm's geographical location, soil characteristics as well as environmental factors such as temperature and rainfall.**

Keywords*— *crop prediction; machine learning; crop recommendation system; smart farming; multi-label classification

I. INTRODUCTION

Maharashtra underwent several fluctuations last year with respect to the retail price of onions. The price increased from Rs. 26 per kilo in the first half of the year to a whopping Rs. 50 per kilo in August [1]. Observing the shoot in the price, many of the farmers in the state decided to grow onions on their farm, in the hope of making exorbitant profits. While this resulted in abundant supply in certain regions of Maharashtra, many other regions suffered failed crop output due to unfavorable conditions for growing onions. A subsequent shortage again in the following months had harsh ramifications on the lives of common man, as middleclass households could no longer afford onion- a frequently used commodity in their kitchen.

This example just goes on to show that a farmer's decision about which crop to grow is generally clouded by his intuition and other irrelevant factors like making instant profits, lack of awareness about market demand, overestimating a soil's potential to support a particular crop, and so on. A very misguided decision on the part of the farmer could place a significant strain on his family's financial condition. Perhaps this could be one of the many reasons contributing to the countless suicide cases of farmers that we hear from media on a daily basis. In a country like India, where agriculture and related sector contributes to approximately 20.4 per cent of its Gross Value Added (GVA) [2], such an erroneous judgment would have negative implications on not just the farmer's family, but the entire economy of a region. For this reason, we have identified a farmer's dilemma about which crop to grow during a particular season, as a very grave one.

The need of the hour is to design a system that could provide predictive insights to the Indian farmers, thereby helping them make an informed decision about which crop to grow. With this in mind, we propose AgroConsultant- an intelligent system that would consider environmental parameters (temperature, rainfall, farm's latitude, longitude, altitude and distance from the sea) and soil characteristics (pH value, soil type and thickness of aquifer and topsoil) before recommending the most suitable crop to the user. This model would take input from another recommendation system, called Rainfall Predictor, which would predict the month-wise rainfall of the next twelve months for the particular user's district.

II. LITERATURE REVIEW

More and more researchers have begun to identify this problem in Indian agriculture and are increasingly dedicating their time and efforts to help alleviate the issue. In [3], the

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

authors make use of Regularized Greedy Forest to determine an appropriate crop sequence at a given time stamp.

The authors of [4] have proposed a model that makes use of historical records of meteorological data as training set. Model is trained to identify weather conditions that are deterrent for the production of apples. It then efficiently predicts the yield of apples on the basis of monthly weather patterns. The effect of temperature on the sugar content of apples is also taken into account to detect potential amount of damaged yield.

The use of several algorithms like Artificial Neural Network, K-Nearest Neighbors, and Regularized Greedy Forest is demonstrated in [5] to select a crop based on the prediction yield rate, which, in turn, is influenced by multiple parameters. Additional features included in the system are pesticide prediction and online trading based on agricultural commodities.

Another intelligent model, presented in [6], allows for the prediction of soil attributes such as phosphorous content. Here, the authors make use of different classification techniques like Naive Bayes, C4.5, Linear Regression and Least Median Square to achieve high prediction accuracy. This system can be very beneficial for farmers to determine the suitability of the soil to support a particular crop.

One shortcoming that we identified in all these notable published works was that the authors of each paper concentrated on a single parameter (either weather or soil) for predicting the suitability of crop growth. However, in our opinion, both these factors should be taken together into consideration concomitantly for the best and most accurate prediction. This is because, a particular soil type may be fit for supporting one type of crop, but if the weather conditions of the region are not suitable for that crop type, then the yield will suffer. Similarly, there may be a case where the weather conditions are favorable but soil characteristics are not.

III. Proposed Approach

To eliminate the aforementioned drawbacks, we propose AgroConsultant- which takes into consideration all the appropriate parameters, including temperature, rainfall, location and soil condition, to predict crop suitability. Fig. 1 illustrates the system architecture of AgroConsultant.

A. Sub-system 1: Crop Suitability Predictor

This sub-system is fundamentally concerned with performing the primary function of AgroConsultant, which is, providing crop recommendations to farmers. The steps involved in this sub-system are:

a) Acquisition of Training Dataset

The accuracy of any machine learning algorithm depends on the number of parameters and the correctness of the training dataset. For the first sub-system, we have made use of 'India Agriculture and Climate Data Set' [7].

This dataset encompasses historical records of soil and meteorological parameters, which were accumulated over the thirty-year period (from 1957-58 to 1986-87). It covers more than two hundred and seventy Indian districts, thereby constituting 13 major states of the country. The aforementioned parameters are provided for five major (bajra, jowar, maize, rice and wheat) and fifteen minor (barley, cotton, groundnut, gram, jute, other pulses, potato, ragi, tur, rapeseed and mustard, sesame, soybean, sugarcane, sunflower, tobacco) crops.

The schema of the training dataset is as follows:

- Soil Type: not used, laterite, red and yellow, gray brown, desert, tarai, shallow black, medium black, deep black, saline and deltaic, red, red and gravely, mixed red and black, coastal alluvial, skeletal, deltaic alluvial, calcerous, black, saline and alkaline, alluvial river.
- Aquifer thickness: DMAQ3 (value=1 if aquifer is >150 meters thick), DMAQ2 (value=1 if aquifer is 100-150 meters thick), DMAQ1 (value=1 if aquifer is <100 meters thick).
- Soil pH: DMPH5 (4.5<pH<5.5), DMPH6 (5.5<pH<6.5), DMPH7 (6.5<pH<7.5), DMPH8 (7.5<pH<8.5).
- Thickness of topsoil: DMTS1 (value= 1 if topsoil is 0 - 25 cm. thick), DMTS2 (value= 1 if topsoil is 25-50 cm. thick), DMTS3 (value= 1 if topsoil is 50 - 100 cm. thick), DMTS4 (value = 1 if topsoil is 100 - 300cm. thick), DMTS5 (value = 1 if topsoil is > 300 cm. thick).

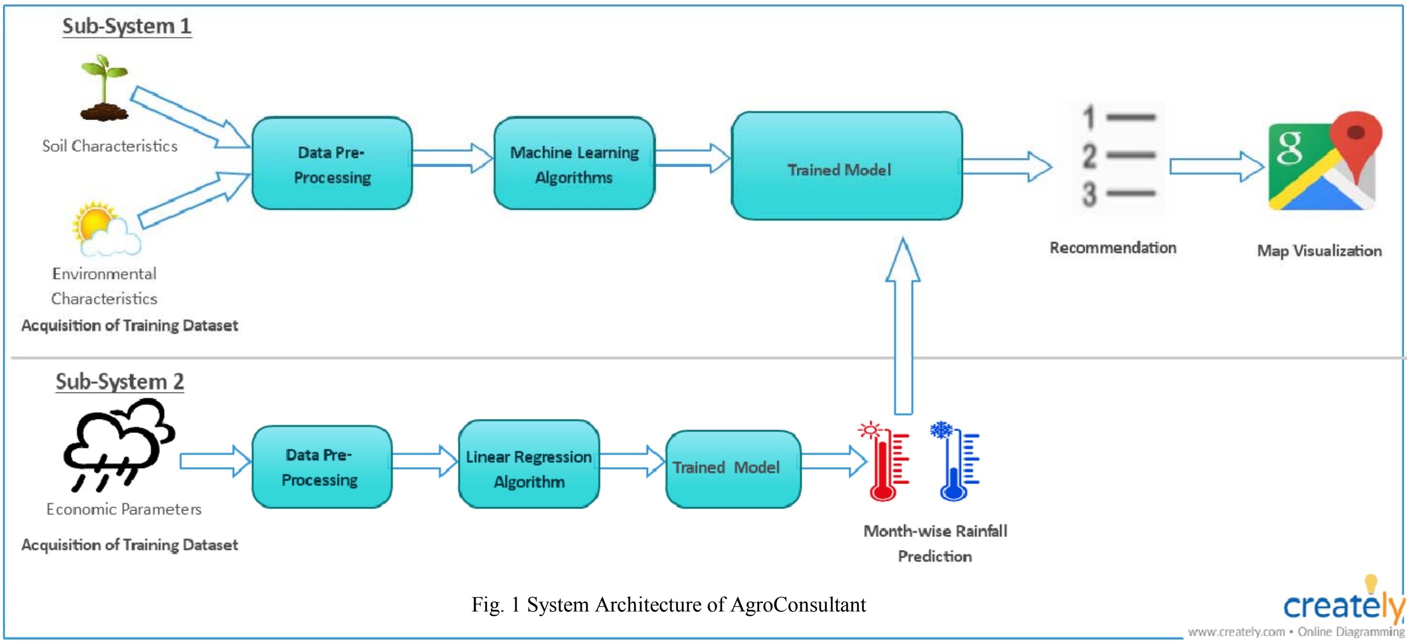

Sub-System 1
Soil Characteristics
Environmental Characteristics
Acquisition of Training Dataset
Data Pre-Processing
Machine Learning Algorithms
Trained Model
1
2
3
Recommendation
Map Visualization
Sub-System 2
Economic Parameters
Acquisition of Training Dataset
Data Pre-Processing
Linear Regression Algorithm
Trained Model
Month-wise Rainfall Prediction
creately
www.creately.com • Online Diagramming

Fig. 1 System Architecture of AgroConsultant

- Precipitation: Month-wise rainfall (in mm).
- Temperature: Month-wise temperature (in °C).
- Location parameters: Includes the latitude, longitude, altitude and distance from the sea of the farm.

b) Data Preprocessing

This is a two-step process. The first step is to remove the missing values which were represented by a dot ('.') in the original dataset. The presence of these missing values deteriorates the value of the data and subsequently hampers the performance of machine learning models. Hence, in order to deal with these missing values, we replace them with large negative values, which the trained model can easily treat as outliers.

The second step before the data is ready to be applied to machine learning algorithms is to generate class labels. Since we intend to use supervised learning, class labels are necessary. The original dataset did not come with labels, and hence we had to create them during the data preprocessing phase. The required labels were generated using production (in tons) and area under cultivation (in hectares) for each crop. Those whose production ÷ area value was greater than 0 were given label 1. In all other cases, a class label of 0 was assigned.

c) Machine Learning Algorithms

Since in the proposed model, more than one class can be assigned to a single instance, Multi-label classification (MLC) would be the ideal choice. Decision Tree, K Nearest Neighbor (K-NN), Random Forest and Neural Network are four machine learning algorithms that have in-built support for MLC.

Decision Tree

It is a supervised learning algorithm where attributes and class labels are represented using a tree. Here, root attributes are compared with the record's attribute and subsequently, depending upon the comparison, a new node is reached. This comparison is continued until a leaf node with a predicted class value is reached. Therefore, a modeled decision tree is very efficient for prediction purposes. [8]

K-NN

It is a non-parametric method used for making predictions. In this, the predicted value is a class membership. The first step of the K-NN algorithm is to identify the k nearest neighbors for each incoming new instance. The instance is classified by a majority vote of these neighbors. In the second step, depending on the label sets of the k neighbors, a label is predicted for the new instance. [9]

Random Forest

It is an ensemble method of learning that is commonly used for both classification and regression. In order to train the model to perform prediction using this algorithm, the test features must be passed through the rules of each randomly created tree. As a result of this, a different target will be predicted by each random forest for the same test feature. Then, votes are calculated on the basis of each predicted target. The final prediction of the algorithm is the highest votes predicted target. The fact that random forest algorithm can efficiently handle missing values and that the classifier can never over-fit the model are huge benefits for using this algorithm. [10]

Neural Network

Neural Network systems progressively improve their performance by learning from examples. They are based on a collection of connected nodes called neurons. Signals are then be transmitted between these neurons using connections. The neurons and connections have a weight associated with them, which is updated and adjusted as learning proceeds. [11]

In order to ensure that AgroConsultant has the highest possible accuracy, we implemented all the four above-mentioned algorithms individually. The performances of the four were then compared, and the one with the highest accuracy was selected for the model.

d) Trained Model and Crop Recommendations

After applying the data to different machine learning algorithms, we obtain trained models of the crop recommendation system. The weights of this model can then be saved, and the farmers can easily avail crop recommendations by giving their farm's soil type, aquifer characteristics, top soil thickness and pH as the input to the system. The rainfall predicted by sub-system 2 is also given as the input to this trained model.

e) Map Visualization

A particular crop may be the most suitable for given soil and weather conditions. When all the farmers of one region use AgroConsultant for the same season, they are bound to get the same recommendations. However, we know that if all the farmers from the region will grow the same crop, it will result in surplus. To avoid such a condition, we present the Map View feature, where the farmers can view the sow decisions made by his neighboring farmers using a pop-up marker on the map. Accordingly, he can make decisions about his own crops.

To implement the Map Visualization feature in AgroConsultant, we make use of a JavaScript library called Leaflet.js [12]. It is used to produce and display interactive maps on HTML webpages. The advantage of using this library is that it creates maps of any desired tile type, enables us to zoom in and out as well as pan across the map to reach a desired location.

We also make use of Flask [13], which is a powerful Python micro-framework that allows building efficient web applications by providing various libraries, tools and technologies. It is a great way of running light-weight web services on hosts.

B. *Sub-system 2: Rainfall Predictor*

Each and every crop has its own rainfall requirement. If this requirement is not met, the crop yield will suffer. On the other hand, if surplus rainfall is available, the yield may again undergo negative consequences. Hence, rainfall is a very important parameter for the growth of any crop. However, farmers cannot be expected to predict the expected rainfall during the months between the sow and harvest season. For this reason, we decided to implement this sub-system, which predicts the rainfall (in mm) for each of the 12 months of the current year, depending on the location of the user's farm. The predicted output of this sub-system can then be fed to sub-system 1 for prediction of crop suitability. The steps involved are:

a) Acquisition of Training Dataset

For this sub-system, we used the meteorological dataset provided by [14]. This training dataset consists of 117 years (from 1900 to 2017) of month-wise rainfall of all the 29 states in India.

b) Data Preprocessing

Similar to the data pre-processing step done for sub-system 1, here the missing values are eliminated by replaced with large negative values (-9999).

c) Linear Regression Algorithm

Linear Regression is a supervised learning approach that is used to predict a quantitative response (y) from a predictor variable (x) by making use of statistical approach [15]. Given the nature of the training dataset, such a linear relation can be easily predicted between Indian states and the monthly precipitation values.

d) Trained Model and Rainfall Prediction

Once the training dataset is fitted to the linear regression algorithm, we get a trained rainfall predictor model.

When an Indian state is given as the input, this model gives 12 float values, corresponding to the rainfall (in mm) of the twelve months in that state.

IV. Experiments, Results and Discussions

All implementations are carried out on Windows 10 having hardware configuration of Intel core i5 processor with 8GB internal RAM and 1.60GHz of CPU speed.

A. *Rainfall Predictor*

An accuracy of 71% was obtained from the rainfall predictor model. Fig. 2 shows the predictor page of AgroConsultant. When the 'Get Rainfall Prediction' button is clicked, the state of the registered user is taken as the input and applied to the trained model. Fig. 3 shows the predictor output, where the rainfall values are shown underneath the button.

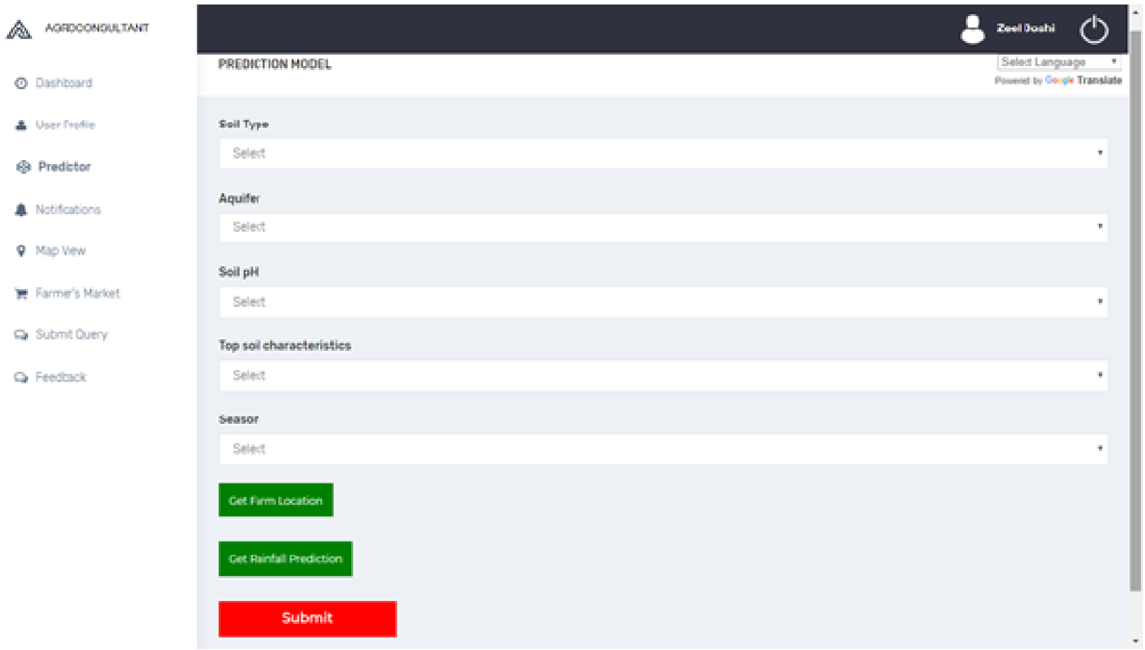
B. *Crop Suitability Predictor*

When the pre-processed training dataset of sub-system 1 was applied to four different machine learning algorithms, different accuracies were obtained. Table 1 shows a comparison between these accuracies.

ALGORITHM	ACCURACY (%)
Decision Tree	**90.20**
K-NN	**89.78**
Random Forest	**90.43**
Neural Network	**91.00**

Table 1. Comparison of the accuracies

Clearly, Neural Network provided the highest accuracy percentage. Hence, we implemented the crop recommendation model of AgroConsultant using Neural Network.



PREDICTION MODEL

Soil Type
Select

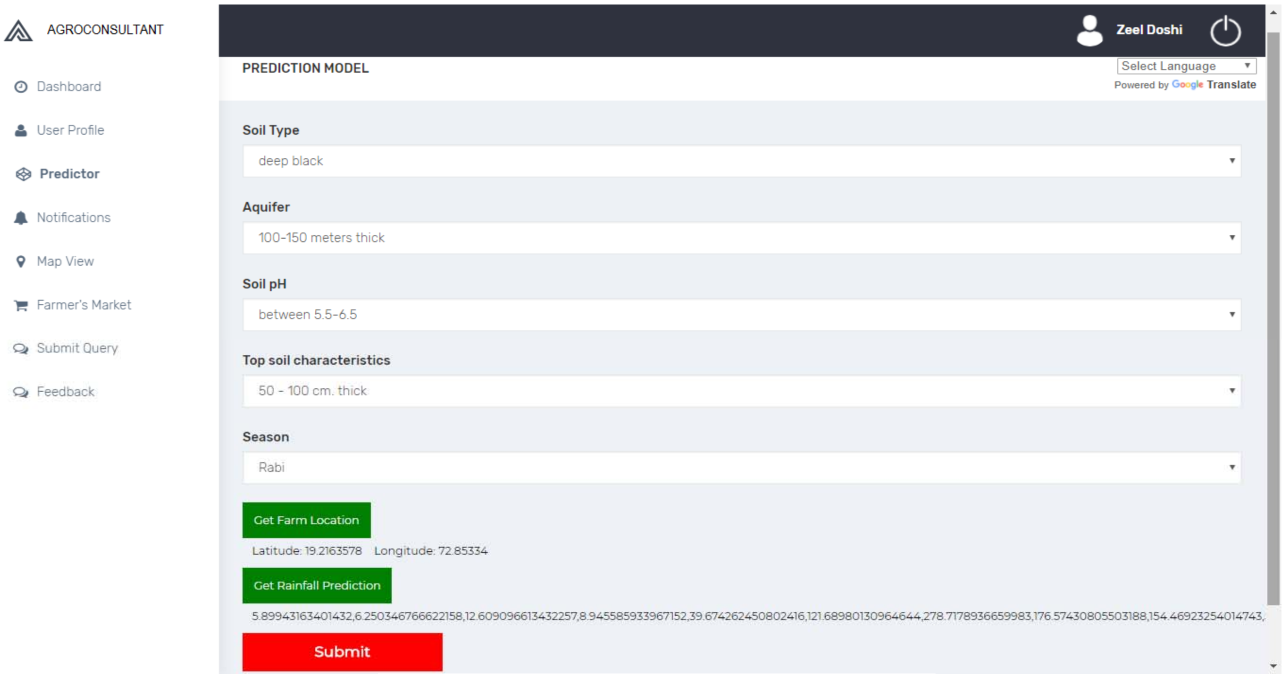
Aquifer
Select

Soil pH
Select

Top soil characteristics
Select

Season
Select

Fig. 2 Prediction Page



PREDICTION MODEL

Soil Type: deep black

Aquifer: 100-150 meters thick

Soil pH: between 5.5-6.5

Top soil characteristics: 50 - 100 cm. thick

Season: Rabi

Get Farm Location

Latitude: 19.2163578 Longitude: 72.85334

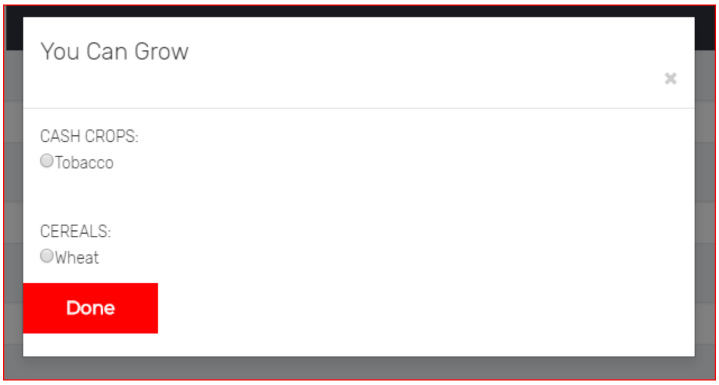
Get Rainfall Prediction

5.89943163401432,6.250346766622158,12.609096613432257,8.945585933967152,39.674262450802416,121.68980130964644,278.7178936659983,176.57430805503188,154.46923254014743,

Submit

Fig. 3 Rainfall Prediction Output

Once the farmer selects his farm parameters and gets the latitude and longitude of this farm using the 'Get Farm Location' button, he gets a list of crops that he can grow for that particular season. This result is illustrated in Fig. 4.



You Can Grow

CASH CROPS:

○Tobacco

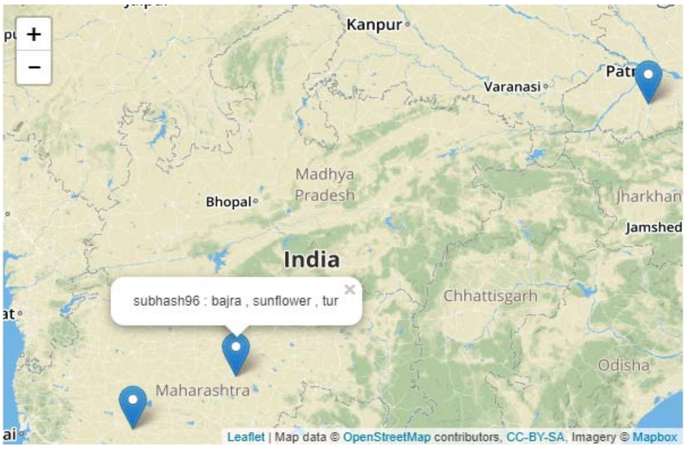
CEREALS:

○Wheat

Done

Fig. 4 Crop Suitability Predictor Output

The Map Visualization feature, where the sow-decisions of all the farmers using the AgroConsultant portal are displayed using a pop-up marker are shown in Fig. 5.



Kanpur
Varanasi
Patn
Madhya Pradesh
Bhopal
Jharkhan
Jamshed
India
Chhattisgarh
Odisha
Maharashtra
subhash96 : bajra , sunflower , tur
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Fig. 5 Map Visualization feature with Pop-Up Marker showing farmer's decision

V. Conclusion And Future Work

In this paper, we have successfully proposed and implemented an intelligent crop recommendation system, which can be easily used by farmers all over India. This system would assist the farmers in making an informed decision about which crop to grow depending on a variety of environmental and geographical factors. We have also implemented a secondary system, called Rainfall Predictor, which predicts the rainfall of the next 12 months. The high accuracies provided by both these models make them very efficient for all practical and real-time purposes.

The model proposed in this paper can be further extended in the future to incorporate a feature to predict crop rotations. This would ensure maximized yield as the decision about which crop to grow would now also depend upon which crop was harvested in the previous cycle.

Furthermore, crop demand and supply as well as other economic indicators like farm harvest prices and retail prices can also be considered as parameters to the Crop Suitability Predictor model. This would provide a holistic prediction not only on the basis of environmental and geographical factors, but also depending on the economic aspects.

References

[1] "Onion, tomato price spike: season not the only reason", available at https://www.thehindubusinessline.com/economy/agri-business/onion-tomato-price-spike-season-not-the-only-reason/article9957255.ece, visited in February 2018

[2] “Agriculture in India: Industry Overview, Market Size, Role in Development...|IBEF”, available at https://www.ibef.org/industry/agriculture-india.aspx, visited in February 2018

[3] Rakesh Kumar , M.P. Singh, Prabhat Kumar and J.P. Singh, “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, 2015

[4] Haedong Lee and Aekyung Moon, “Development of Yield Prediction System Based on Real-time Agricultural Meteorological Information”, 16th International Conference on Advanced Communication Technology, 2014

[5] T.R. Lekhaa, “Efficient Crop Yield and Pesticide Prediction for Improving Agricultural Economy using Data Mining Techniques”, International Journal of Modern Trends in Engineering and Science (IJMTES), 2016, Volume 03, Issue 10

[6] Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade and Vahida Attar, “Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction”, International Journal of Computer Science Issues, Volume 9, Issue 3

[7] “INDIA AGRICULTURE AND CLIMATE DATA SET”, available at https://ipl.econ.duke.edu/dthomas/dev_data/datafiles/india_agric_climate.htm, visited in November 2017

[8] “How Decision Tree Algorithm works”, available at http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/, visited in February 2018

[9] “k-nearest neighbors algorithm-Wikipedia”, available at https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm, visited in February 2018

[10] “How the random forest algorithm works in machine learning”, available at http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/, visited in February 2018

[11] “Artifical Neural Network”, available at https://en.wikipedia.org/wiki/Artificial_neural_network, visited in March 2018

[12] “Leaflet- a JavaScript library for interactive maps”, available at http://leafletjs.com/, visited in March 2018

[13] “Flask (A Python Microframework)” , available at http://flask.pocoo.org/, visited in March 2018

[14] “Latest Socio-Economic Statistical information & Facts About India”, available at https://www.indiastat.com, visited in February 2018

[15] “Linear Regression-Intro to Machine Learning #6-simple AI-Medium”, available at https://medium.com/simple-ai/linear-regression-intro-to-machine-learning-6-6e320dbdaf06, visited in February 2018