Crop Recommendation System using Machine Learning

[1]Shubham Pathrikar, [2]Pranjali Gurav, [3]Kshitij Agrawal, [4]Govind. S. Pole
[1]Student, [2]Student, [3]Student, [4]Professor
[1,2,3,4]Computer Department, MES College of Engineering, SPPU, Pune, India

Abstract: The agricultural sector in India has gone through many changes, such as the green revolution which helped in revolutionizing the yield per hectare and many more such plans implemented by the government. Even though such developments have taken place there is a strong need to increase the production of crops to satisfy and sustain the Indian population. To increase the yield per hectare we propose making use of Machine Learning and Data Science to develop a system which would recommend proper crop taking into account various factors. The factors which will be taken into account will be soil pH, soil texture, soil NPK levels, soil's electrical conductivity and many more. A lot of data is being generated due to the progress of IOT in agriculture, all that data can be used to develop machine learning models. These machine learning models will then help us to recommend proper crop suited for particular soil factors.

Keywords **- Machine Learning, Agriculture, Recommender Systems.**

I. INTRODUCTION

Agriculture is one of the most ancient occupations invented by human beings and still remains the main source of survival. Before the advent of agriculture humans survived through foraging but due to changes in their environment agriculture was adopted. Agriculture started about 11,000 years ago in Paleolithic age and thus lead to the development of agrarian civilizations around the world. Two major factors responsible for the development of such civilizations are collective learning and technological development. Collective learning helped humans to learn different things faster by sharing information with each other and thus increasing their chances of survival against hazards in their environment. Technological developments made a lot of difficult tasks easy for the man. Humans started living in groups which lead to the development of villages, towns, and cities. Agriculture was able to support more people as compared to foraging as it provided more energy and more resources, this rapid change in the adoption of agriculture lead to the growth of population and thus the pace of historical change increased. Over the thousands of years of life on earth, humans have grown the most, about 11,000 years ago the population of humans was in thousands and it has grown to be more than 7 billion till now.

Post Agrarian phase was marked with the unprecedented growth of human intellect and technology. Fast forwarding to the industrial revolution which leads to a drastic increase in the production capacity of goods and resources. The Industrial Revolution was the transition to new manufacturing processes in the period from about 1760 to sometime between 1820 and 1840. This transition included going from hand production methods to machines, new chemical manufacturing, and iron production processes, the increasing use of steam power, the development of machine tools and the rise of the factory system. The Industrial Revolution also led to an unprecedented rise in the rate of population growth.

India, the name comes from the great Indus Valley civilization which is one of the ancient civilization to have lived on earth. As the world progressed India also went through many changes. India at the core is an agrarian country with around 60% of the population directly or indirectly depending on agriculture. The Indian agriculture sector accounts for 18 percent of India's gross domestic product (GDP) and provides employment to 50% of the countries workforce. Throughout the history of Indian democracy, the government has taken many initiatives to support agriculture in a country like the Green Revolution, Pradhan Mantri Fasal Bima Yojana, Paramparagat Krishi Vikas Yojana, etc. But irrespective of such initiatives the growth which was expected has not yet been seen. Most importantly these plans fail to reach each and every person who is somehow involved in agriculture. Farmers suffer the most not only due to a failure of these initiatives but also due to other factors such as climatic changes and market price fluctuation which ultimately affects the crop yield. More than 75% of the farmers depend on traditional methods of farming which are proving to be obsolete in today's fast-paced world.

To solve the about issue we are proposing a recommendation system which uses machine learning techniques to predict suitable crop for soil parameters. Soil parameters included are, for example, soil pH, NPK level, water holding capacity (WHC), the quantity of micronutrients, etc. Machine learning is a sub-branch of Artificial intelligence which deals with algorithms which learn for the information and data. In machine learning computers don't have to be explicitly programmed but can change and improve themselves; this phenomena of machine learning makes it optimal for developing a system which learns from the huge amount of historical and learns in much more efficient way than human beings.

II. LITERATURE SURVEY

[1] Data mining has gained importance in recent years as it allows to process huge amounts of data comfortably and generate priceless information. Agriculture plays a very important role in Indian economy and employment. Farmers in India use traditional methods to select a crop to be farmed which sometimes might turn out to be a wrong choice thus reducing profits for farmers and

has a setback to the economy. Data mining can be helpful in solving this problem. Recommender systems developed with different algorithms such as Naïve Bayes, CHAID, Random Tree and k-Nearest Neighbor can be used.

[2] Different approaches to building a recommendation system can be used which recommend suitable crop in different places to farmers. The system can use user's location and work with different agro-ecological and agro-climatic data in subdistrict level to calculate the similarity between sub-districts using Pearson co-relation similarity algorithm. After that, it selects top n similar sub-districts. It then utilizes the crop production rates of each crop and seasonal information of the similar sub-districts, it recommends top-k crops to a user of a sub-districts.

[3] Agriculture is one of the major factors of the Indian economy but in certain unfamiliar circumstances like natural calamities, it faces major losses. Here they suggest building a system which will predict the crop yield in advance which would help farmers and government to make decisions on selling, storing, fixing minimum support price, importing, exporting, etc. The system applies different learning algorithms for predicting yields of different crops. Algorithms used are: Multiple Linear Regression, Decision tree analysis and ID3, Support Vector Regression model, APAR, SEBAL, Carnegie Institution Stanford model, Neural Networks, C4.5 algorithm and decision tree, Harmonic Analysis of NDVI TimeSeries algorithm, Gaussian Processes, Relational cluster Bee Hive algorithm, KMeans algorithm for clustering and for classification Linear Regression,k-NN, ANN model, J48, LADTree, K-Nearest Neighbor (KNN) and Naive Bayes (NB).

[4] The paper talks about a system which predicts a class according to the soil parameters like pH, Texture, EC and K. A class contains crops suitable for those parameters. It has used mining algorithms such as J48, Random Forest, Naïve Bayes, and BF Tree. Weka tool was used to calculate the accuracy and to build a model.

[5] Machine learning has seen unprecedented growth in its applications. Different machine learning techniques like Classification, Clustering, Regression, Ranking, and Dimensionality Reduction are used to deal with different kinds of data. Kinds of machine learning algorithms are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

[6] Variations in the size of training and test dataset affect the accuracy of the model built. The model predicts weather parameters using Gaussian Process with RBF kernel.

[7] Agriculture accounts for 13.7% of India's GDP even though it is one of the major sectors of employment. Paper aims to build a recommendation system which collects raw data from soil and weather data with the help of a sensor network. The data is sent to the cloud for further processing and then the results are stored and shared with the user. The result can be sent to users through SMS or Voice Response.

III. METHODOLOGY

3.1 Data Pre-processing

Initially, the dataset contained 46 features with 10 crop classes. Missing values were replaced by the mean of the particular feature. Feature selection reduced the dataset to 8 features which affect the kind of crop to be grown on that particular soil. These features affect the crops ability to absorb nutrients and water from the soil. After feature selection data was normalized using the min-max method which scales the data between 0 and 1. Normalization makes sure that all the features have equal influence on the outcome; normalized data is mainly used with parametric classification algorithms. After normalization data was split into training and testing with the splitting ratio equal to 70:30; with 70 % as training data and 30% as testing data.

3.2 Selecting Classifiers:

Eight different classifiers from sci-kit learn library of python were applied on the pre-processed data and the following results were achieved:

Classification Algorithm	Accuracy of Prediction(%)
MLP Classifier	81.0
Naïve Bayes	85.67
Decision Tree Classifier	84.0
Random Forest Classifier	72.67
Ada Boost Classifier	41.67
Quadratic Discriminant Analysis	85.33
Gaussian Process Classifier	96.0
K-Nearest Neighbour	92.67

Table.1 Results of different classifier

Above table shows the accuracy associated with different classifiers. Following 3 classifiers selected for implementation because of their high accuracy of prediction:

- Gaussian Process Classifier (GPC)
- K-Nearest Neighbor (KNN)
- Naïve Bayes (NB)

These three classifiers are used to build an ensemble system in which all are used to recommend a crop and the crop with most frequency is recommended. Ensemble algorithm is as follows:

```
1. Take the three results (R1, R2, R3) of the trained model as input.
2. Calculate the majority by
        (a) if(R1==R2 and R1==R3) then
                FinalResult = R1;
        (b) else if(R1 == R2 or R1 == R3)then
                FinalResult = R1;
        (c) else if(R2 == R3) then
                FinalResult = R2;
        (d) else FinalResult = R1;   //none of them same

3. Send FinalResult to Database and display it to the user.
```

Here classification accuracy of first classifier (GPC) is highest therefore if none of the results match the result of classifier one is selected by default.

3.3 Selected Classifiers

3.3.1 Gaussian Process Classifier (GPC)

Gaussian Process Classifier is based on Gaussian process which is a stochastic process whose kernel is a Gaussian normal distribution. A stochastic process is a group of random variables over space and time. Points in the stochastic process are related to the underlying probability distribution governing the process as well as to the other points earlier in the process. Brownian motion is an example of a stochastic process.

Gaussian Process Classifier is a non-parametric classification method based on Bayesian methodology. It is a very effective classifier and was developed in geo-statistics in the seventies. GPC assumes prior distribution on the underlying probability densities that ensures some smoothness properties. The final classification is determined as the one that provides a good fit for the observed data, and at the same time ensuring smoothness.

GPC focuses on modeling the posterior probabilities, by defining certain latent variables: Ki is the latent variable for the pattern i.

Consider a two-class case: Ki is a measure of the degree of membership of class A1, meaning:

- If Ki is positive and large then pattern i belongs to class A1 with large probability.
- If Ki is negative and large in magnitude then pattern i belongs to class A2 with large probability.
- If Ki is close to zero, class membership is less certain.

3.3.2 K-Nearest Neighbor (KNN)

KNN is one of the simplest and easy to implement algorithm. It can be used for regression problems as well as classification problems. Results of KNN are also easy to interpret as compared to other algorithms.

In KNN, k indicated the number of nearest neighbor we wish to take votes from (i.e for a particular data point we will find the k nearest points and take their class into consideration). The class to which the majority of the k points belong becomes the class of the data point.

The measure of distance used to calculate the nearest k point usually is Euclidean Distance given by:

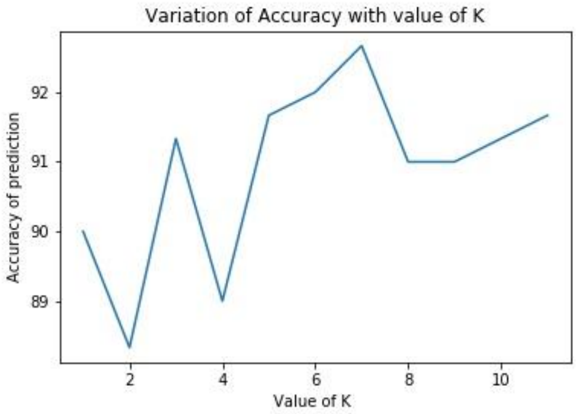

Figure 3.1 KNN

$$Euclidean\ Distance = \sqrt[2]{(x1-x2)^2+(y1-y2)^2}$$

For selecting the value of k, we varied the value of k from 1 to 11 and selected k for which accuracy of the model was the greatest. Following graph show the results

3.3.3 Naïve Bayes

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayes theorem provides a way of calculating posterior probability P(c—x) from P(c), P(x) and P(x—c) as given below: The equation states that probability of event c occurring if event x has occurred depends upon event x occurs given c occurred, class (event c) probability and predictor (independent attribute, event x) probability.

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right) * P(c)}{P(x)}$$

$$P\left(\frac{c}{x}\right) = Posterior\ Probability,\ P\left(\frac{x}{c}\right) = Likelihood\ Probability,\ P(c) = Class\ Prior\ Probability,$$
$$P(x) = Predictor\ Prior\ Probability$$

There are three different types of Naïve Bayes algorithms: Gaussian Naïve Bayes, Binomial Naïve Bayes, and Multinomial Naïve Bayes.

Gaussian NB is used when working with continuous values. Its probabilities are modeled using Gaussian distribution given as:

$$P(x) = \frac{(e^{(x-\mu)^2/2\sigma^2})}{\sqrt[2]{2\pi\sigma^2}}$$

Where µ and σ are the mean and standard deviation by class respectively.

Algorithm:

- Convert the data set into a frequency table.
- Create Likelihood table by finding the probabilities.
- Now, use the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction.

3.4 Ranking Crops

Along with recommending the best crop, other crops are also ranked according to their likelihood (Probabilities). Likelihood/ Probabilities are obtained for Naïve Bayes and then are sorted in descending order to obtain the ranking of crops (i.e. the one with the second highest probability is ranked 2 and so on).

Thetemplateisusedtoformatyourpaperandstylethetext.Allmargins,columnwidths,linespaces,andtextfontsareprescribed;pleasedonotalterthem.Youmaynotepeculiarities.Forexample,theheadmargininthistemplatemeasuresproportionatelymorethaniscustomary.This measurementandothersaredeliberate,usingspecificationsthatanticipateyourpaperasonepartoftheentireproceedings,andnotasanindependentdocument.Pleasedonotreviseanyofthecurrentdesignations.

IV. IMPLEMENTATION

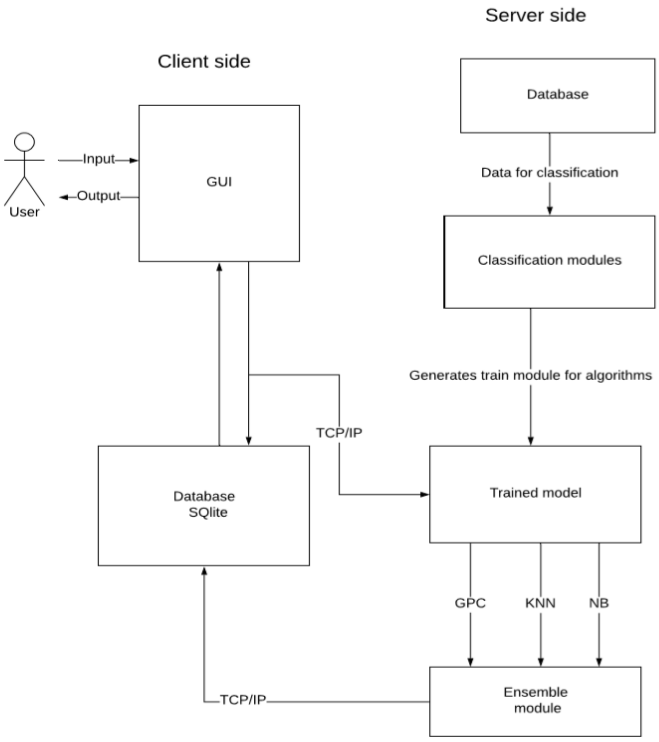

Fig 4.1 System Architecture

The diagram above illustrated the flow of the system implemented. The user enters the value of different parameters (i.e. features) manually and they are sent to a remote server which is responsible for classification. The server returns the value of recommended and ranking of other crops. And finally, the result is displayed to the user.

Front end (Client side) of the system is built on android studio and is compatible for android versions 15 to 28. The backend of the system (Server side) is built in python which is responsible for the major processing of the data. Client and Server communicate using TCP/IP

V. CONCLUSION

The prosperity of the farmers prospers the nation. The objectives and focus of this content-based recommendation system planned to be built on the predictive system are to assist farmers and agricultural sector to overcome the problems faced in cultivation due to improper or incomplete knowledge. We have developed a recommendation system to help farmers in choosing appropriate crops. From the experimental evaluation, we found that the developed system can recommend appropriate crops to a satisfactory level.

VI. FUTURE WORK

The system can be extended to market applications to help the farmers. The parameters like climatic factors can be used to predict the yield of the crop or recommend suitable crop. Collection of more valid details of soil class, latitude, longitude and, the suitable crop can greatly accelerate the efficiency of work. The pre-processing unit could hence be improved and a lot more features can be extended, thus significantly contributing towards the agricultural welfare worldwide.

VII. REFERENCES

[1] S.Pudumalar, E.Ramanujam, R.Harine Rajashreeń, C.Kavyań, T.Kiruthikań, J.Nisha, “Crop Recommendation System for Precision Agriculture”, 2016.

[2] Miftahul Jannat Mokarrama, Mohammad Shamsul Arefin, “RSF: A Recommendation System for Farmers”, 2017.

[3] Yogesh Gandge, Sandhya, “A Study on Various Data Mining Techniques for Crop Yield Prediction”, 2017.

[4] Ansif Arooj, Mohsin Riaz, Malik Naeem Akram,“ Evaluation of Predictive Data Mining Algorithms in Soil Data Classification for Optimized Crop Recommendation”, 2016

[5] Agata Nawrocka, Andrzej Kot, Marcin Nawrocki, “Application of machine learning in recommendation systems”, 2018

[6] Ramesh Medar, Vijay S. Rajpurohit, Rashmi B., “Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning”, 2017

[7] Avinash Jain, Kiran Kumar P S, “Application of Recommendation Engines in Agriculture, 2017.