Machine Learning Based Crop Recommendation System

Dhruv Piyush Parikh[1], Jugal Jain[2], Tanishq Gupta[3] and Rishit Hemant Dabhade[4]
School of Electronics Engineering[1,4]
School of Computer Science Engineering[2,3]
Vellore Institute of Technology, Chennai, India[1,2,3]
Thakur College of Engineering & Technology, Mumbai, India[4]

Abstract: *The three most basic amenities required for the survival of a human being are food, shelter and clothing. In today's tech-savvy generation, the latter two have witnessed a huge scientific boost. Unfortunately, even today, agriculture is considered as more of a man-power oriented field. Most of the farmers are untutored and have little to no scientific knowledge of farming. So, they have to rely on the hit and trial method to learn from experience which leads to wastage of time and resources. Our system focuses on building a predictive model to recommend the most suitable crops to grow in a particular farm based on various parameters. This can be helpful for the farmers to be more productive and competent without wasting any resources by farming the most competent crops.*

Keywords: Crop Recommendation, GUI, Machine Learning, Random Forest Classifier, Tkinter

I. Introduction

Agriculture crop recommendation is a new generation bubble that is engaging the masses. In most of the cases, the farmers are not well aware of the kind of crops they should be growing in their farms. This leads to a lot of confusion and affects the productivity. This is why we are focusing on figuring out the best crop to grow in order to get optimum yield.

We have gathered a dataset built by augmenting datasets of rainfall, climate and fertilizer data available for India. This will give us a better idea of the trends of crops considering different environmental and geographical factors. We can use this dataset to create a machine learning model for predicting the best suitable crop to grow at a particular place. Machine learning can prove to be the turning point of the agriculture industry. By predicting the right crop to be grown, we will help the farmers to decide the raw materials and other resources required much earlier than they would have figured it out otherwise. This will eradicate the problem of nutrients deficiency in fields occurring because of planting wrong crops which can scale down the production efficiency in a compound manner. India is still lacking behind in finding technological solutions for agriculture which is the primary source of income for about 50% people in the country. Promoting more scientific solutions is the need of the hour in order to take the agriculture industry of India to greater heights.

The main idea of the model is to provide the farmers with an ideal recommendation for growing crops taking into consideration the factors such as composition of soil, the environmental factors like temperature, humidity, rainfall and the geographical influence.

II. Design and Methodology

The main ideology of the research revolves around the concept of identifying the most suitable crop to be grown with the help of a machine learning model. Thus, the results can prove to be exceedingly beneficial for the agriculture farmers.

We have used a dataset from Kaggle containing 2200 values of 22 unique crops. We have applied machine learning algorithms on the dataset. The dataset is used to train the model according to the actual values and then test the model for its accuracy. According to Prof. Nischitha K., presently our farmers are not using technology and analysis, so there

DOI: 10.48175/IJARSCT-1509

may be a chance of wrong selection of crops for cultivation which will reduce their income. So, we have also included graphical user interface to make it seamless and attractive to operate for even the first-time users who have no prior experience of using any such app or facility. The GUI has been added using the Tkinter library of python. Our research aims to take into consideration the ground reality reflecting the actual requirements instead of assuming any of the factors. Since we are mainly catering to uneducated farmers, we have made it more and more visual and easier to operate.

This process is laying the foundation for further evaluation through addition of secondary factors having an impact on deciding the crop to be grown on a particular field in a particular place.

2.1 Research Approach

We have used scikit-learn to implement machine learning algorithms on the dataset. The model is trained in such a way that it learns from the data given to it and applies the same trends and knowledge to give optimum results for any given input. The model is then implemented using GUI after testing to give a visual look and clarity in accessing the machine learning results from the front-end.

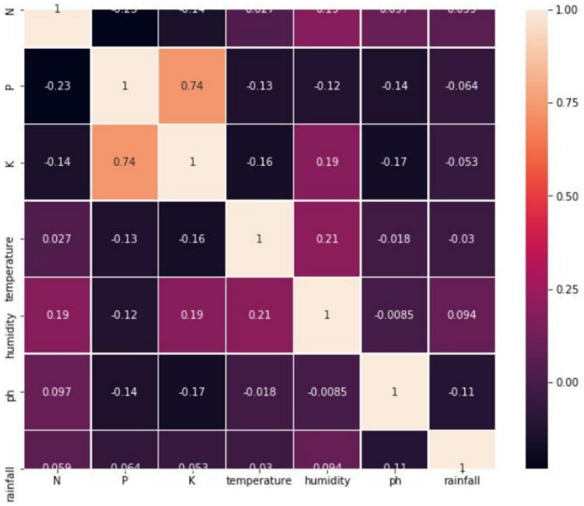


Figure 1: Explorative Data Analysis

2.2 Method Approach

The dataset is loaded in the ML model. Then after splitting of data into training and testing sets, it is trained and tested using three different algorithms. The best algorithm is chosen and then the model is finalized. This model now performs its action by taking in different factors as input and returning the optimal crop yield.

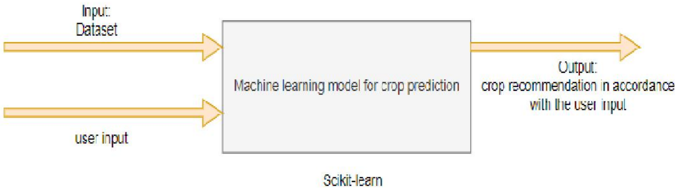
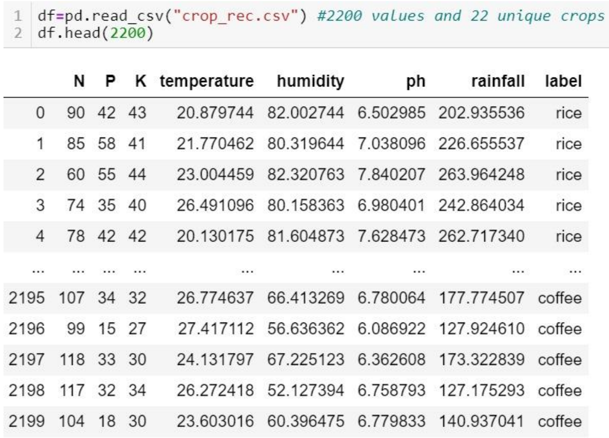

Input:
Dataset
user input
Machine learning model for crop prediction
Output:
crop recommendation in accordance
with the user input
Scikit-learn

Figure 2: Design of Machine Learning Algorithm

2.3 Kaggle

Kaggle is a subsidiary of Google which gives users a platform to get and publish data sets. Apart from this, it also allows the users to build models in an environment that is generally web-based and data-science oriented. Basically, it is a community for machine learning and data science enthusiasts to get data to work and a platform to display their work. It also hosts competitions where people can compete and hone their ML skills and also get some useful research ideas. We have taken our dataset from Kaggle which we have used to train our model.



```
df=pd.read_csv("crop_rec.csv") #2200 values and 22 unique crops
df.head(2200)
```

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
...	...	...	...	...	...	...	...	...
2195	107	34	32	26.774637	66.413269	6.780064	177.774507	coffee
2196	99	15	27	27.417112	56.636362	6.086922	127.924610	coffee
2197	118	33	30	24.131797	67.225123	6.362608	173.322839	coffee
2198	117	32	34	26.272418	52.127394	6.758793	127.175293	coffee
2199	104	18	30	23.603016	60.396475	6.779833	140.937041	coffee

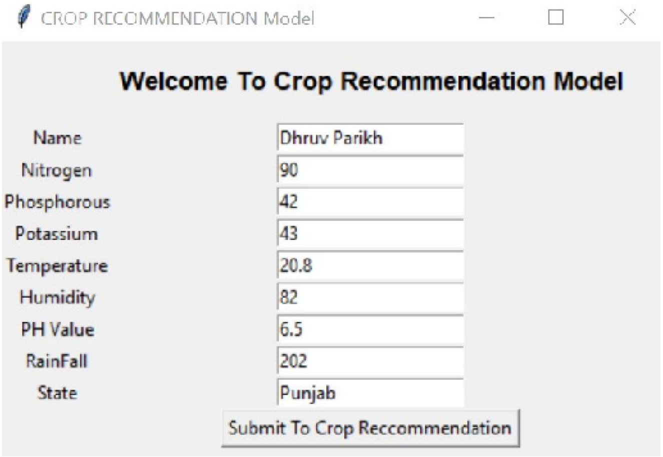
Figure 3: Crop Recommendation Dataset

2.4 Machine Learning

We have used machine learning to make our model capable of suggesting the optimum crop which can be sown by a farmer according to various input factors. Machine learning is a method of analyzing data to automate the building of an analytical model. It is in fact a branch of AI as it is based on the concept of systems learning from data and identifying some patterns to make decisions without much human intercession.

2.5 Tkinter

We have used tkinter for developing a graphical user interface to integrate with the machine learning model to make its UI easier to use the system to obtain required results. Tkinter is a standard library of python used for creating GUI by combining with python in a fast and easy way to provide an object-oriented interface.



Welcome To Crop Recommendation Model

Name: Dhruv Parikh
Nitrogen: 90
Phosphorous: 42
Potassium: 43
Temperature: 20.8
Humidity: 82
PH Value: 6.5
RainFall: 202
State: Punjab

Figure 4: Design of Tkinter Window

III. Algorithm

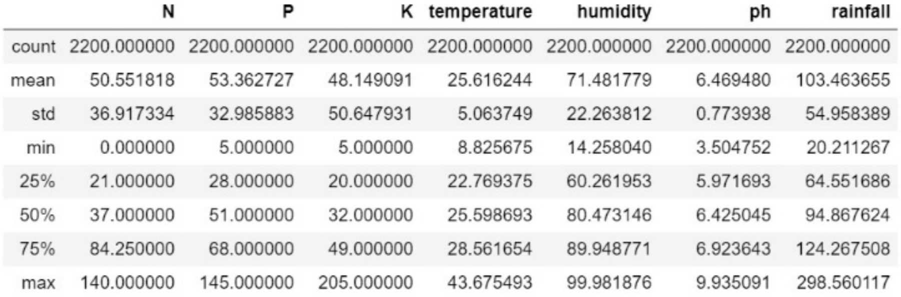
This section discusses the platform and major modules/libraries used to implement the programming of the algorithm for crop prediction through machine learning for prominent accuracy in utilization of land and crop cultivation.

3.1 Scikit-Learn

Scikit-learn is an ML library for python. It has various algorithms for classification and regression like logistic regression, random forest classifiers and support vector machines. We can operate it along with other python libraries like NumPy and pandas. Scikit-learn is one of the best libraries especially for supervised learning which involves training the model by loading a sample dataset which it can observe and structure its learning accordingly. It also gives us the provision to use train_test_split for making training and testing datasets.

3.2 train_test_split Library

We import train_test_split from scikit-learn to split our dataset into two different sets according to our needs- one for training the model and the other for testing the working and accuracy of the trained model so that we can choose the best possible algorithm. We apply different supervised learning algorithms on the training data and obtain results on testing data for all the applied algorithms. Then the best performing algorithm is selected for the model.



	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117

Figure 5: Explorative Analysis on Dataset

3.3 Logistic Regression

Logistic regression is a basic linear model that uses a logistic function for model creation. It categorizes the data into discrete classes by figuring out the relationship trends from the given dataset. It is easy to implement and very efficient to train and can classify unknown data records considerably quickly. But it by default assumes a linear relation between dependent and independent variables which can turn out to be a limitation in the performance of the model in some cases.

3.4 Support Vector Machines (SVM)

Support Vector Machines or SVM consists of a group of algorithms that analyze data for regression and classification. It represents different classes in a single plane iteratively to minimize the error and is also memory efficient. This makes it one of the best algorithms to use if the error persists in basic linear regression. But in case of noisy datasets and large datasets, its performance dips because of chances of overlapping of classes.

3.5 Random Forest Classifier

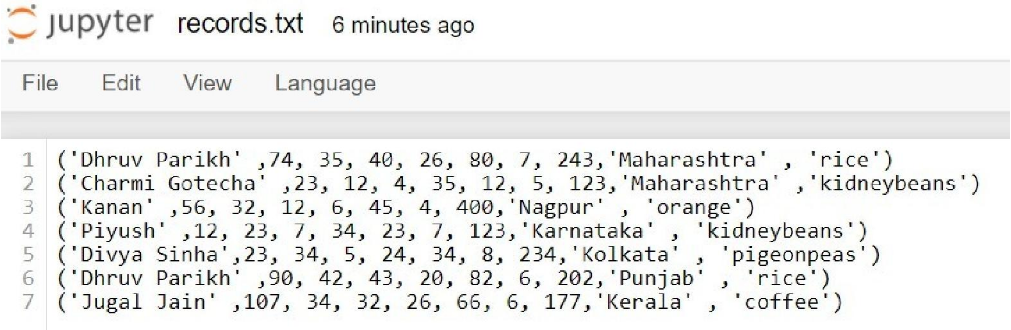
Random forest algorithm is one of the most famous and a widely used supervised learning technique. It contains a number of decision trees for different subsets of the data instead of working on the whole data as a single subset. This improves the accuracy of prediction of the model by several folds as it takes the average of predictions of all the trees and decides the final output on the basis of majority votes of the predictions. This makes it suitable even for the large and varied datasets as it can deliver results with high accuracy in very less amount of time.

3.6Working Algorithm Applied

Jupyter notebook was used for creating the model. We started off with importing the required libraries that were pandas, matplotlib, seaborn and sklearn, which we felt were essential for carrying out proper analysis of the given dataset. A data frame was created to read the csv and operate upon it. We then plotted a heatmap to check the correlation between all the factors. Data splitting was done by splitting the dataset into test and train data. We then trained the model with linear regression and verified the accuracy with the testing data. The same was done for SVM and Random Forest Regressor also. The model with the best results, i.e., the Random Forest Classifier with 99.725% accuracy was selected to make the prediction model. Then we used tkinter to create the graphical user interface and load the model created in the application. The application created successfully predicts 22 unique crops accurately

3.7 File Management System

Creating separate database to store data of different modules is very inefficient. So, for storing the data at one place, we decided to use local file management system. Thelocal python database integrates seamlessly with pandas data frame and saves the data securely by calling a single function. Contrary to the most commonly used databases like MySQL and firebase, the local database used is more secure as it is end to end encrypted and is considerably faster because of local access.

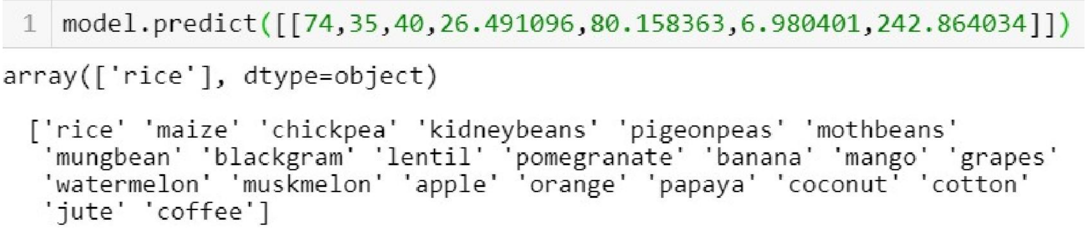


```
('Dhruv Parikh' ,74, 35, 40, 26, 80, 7, 243,'Maharashtra' , 'rice')
('Charmi Gotecha' ,23, 12, 4, 35, 12, 5, 123,'Maharashtra' ,'kidneybeans')
('Kanan' ,56, 32, 12, 6, 45, 4, 400,'Nagpur' , 'orange')
('Piyush' ,12, 23, 7, 34, 23, 7, 123,'Karnataka' , 'kidneybeans')
('Divya Sinha',23, 34, 5, 24, 34, 8, 234,'Kolkata' , 'pigeonpeas')
('Dhruv Parikh' ,90, 42, 43, 20, 82, 6, 202,'Punjab' , 'rice')
('Jugal Jain' ,107, 34, 32, 26, 66, 6, 177,'Kerala' , 'coffee')
```

Figure 6: Records saved in the local database

IV. Results and Discussion

We have used a sample data set from Kaggle, which consisted of various factors needed for the proper growing of a crop. We have taken into consideration these factors which will be taken as an input from the farmer/authority that will be deciding which crop to sow. A detailed analysis has been carried out on these factors and various inferences were generated from the results that we then took into consideration while predicting the optimum crop.



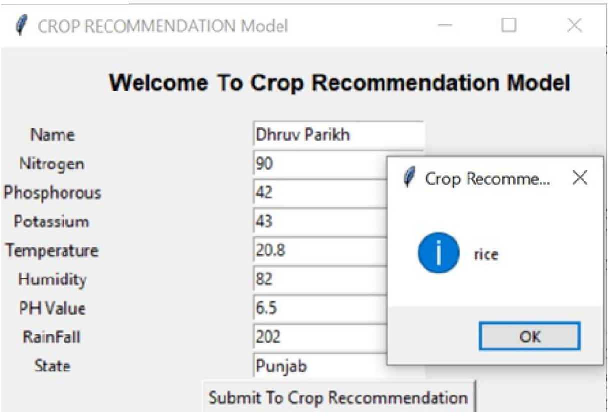
```
model.predict([[74,35,40,26.491096,80.158363,6.980401,242.864034]])

array(['rice'], dtype=object)

 ['rice' 'maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'mothbeans'
  'mungbean' 'blackgram' 'lentil' 'pomegranate' 'banana' 'mango' 'grapes'
  'watermelon' 'muskmelon' 'apple' 'orange' 'papaya' 'coconut' 'cotton'
  'jute' 'coffee']
```

Figure 7: Unique Crops Labeled by Algorithm

The dataset has been analyzed by plotting graphs and maps to check the effect and correlation of different attributes. The attributes taken into consideration are the ratio of nitrogen in soil, ratio of phosphorous in soil, ratio of potassium in soil, temperature, humidity, pH of soil and the amount of rainfall. This covers varied factors which affect the growth of crops and are the deciding factors in which crop to be chosen to sow. The model then shows the label as output which is the recommended crop. After applying three different algorithms- logistic regression, SVM and Random Forest Classifier, we found out that Random Forest Classifier gave the best results and so the model has been made using the same. The model has then been incorporated with GUI using Tkinter which makes it a complete application which can directly be put to use without further updates required.



CROP RECOMMENDATION Model

Welcome To Crop Recommendation Model

Name	Dhruv Parikh
Nitrogen	90
Phosphorous	42
Potassium	43
Temperature	20.8
Humidity	82
PH Value	6.5
RainFall	202
State	Punjab

Crop Recomme...

rice

Figure 8: Machine Learning Algorithm With GUI

This research will give a deep insight into the various conditions affecting the crop yield which can help the farmers to reduce losses and generate a greater revenue. This will also positively contribute towards reducing environmental depletion and maintaining a balance between continued agricultural growth and the ecological health of the land upon which humans depend.

V. Conclusion and Future Work

We have worked on a sample dataset from Kaggle which has taken into consideration records obtained from a broad agricultural demography. Farmers generally use hit and trial method which leads to wastage of land and resources or even disproportionate growth of crops. We are trying to break all such taxing walls by providing them with an accurate and justified model made by machine learning using random forest classifier to identify the correct crop to be grown in their farms. This will help them in improving their crop production both qualitatively and quantitatively. This will also help them to maintain the quality and nutrition contents of the soil.

As concerning future score, when the farmers sow a particular crop, there might face some issues or diseases in the crop before harvesting. In that case, they can upload the photographs of the crop and the soil report. Then the AI model can identify the problems and provide them with probable solutions. We can also provide IOT solutions through APIs or virtual agents which can help the farmers connect with raw material dealers, who can provide them with the materials required for instance seeds and fertilizers according to the crop recommended to them by the model.

VI. References

[1] 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

[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] 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] 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].

[5] 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.

[6] 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

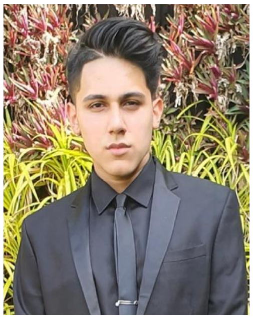
[7] Monali Paul, Santosh K. Vishwakarma, Ashok Verma (2015), 'Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach', International Conference on Computational Intelligence and Communication Networks

Biography

Mr. Dhruv Piyush Parikh is currently pursuing his B.Tech degree from VIT University at Chennai. His research interests are Machine Learning, Computer Vision, Voice applications, Internet of Things (IoT), Data Science, Website Development, Deep Learning, Product Management and Embedded Systems. He has successfully developed a personal virtual assistant collaborated with a dedicated secure server. He has published over 31 algorithms on various open-source platforms. He has worked on more than 8 research paper and patent system development.

Mr. Jugal Jain is currently pursuing his B.Tech degree from VIT University at Chennai. His research interests are Data Analytics and R lab analysis through diversified analysis on market trends.

Mr. Tanishq Gupta is currently pursuing his B.Tech degree from VIT University at Chennai. His research interests are Data Analysis and tapping new endeavors of technologies.



Mr. Rishit Hemant Dabhade is currently pursuing his B.Tech from Thakur college of engineering & Technology in Mumbai. He has completed his diploma in E&TC engineering from Thakur Polytechnic in Mumbai with first class distinction and ranked first in the college in his final A.Y. 2018/19. His research interests are Engineering Management, Product Management, Robotics, Mechatronics & Internet of Things (IoT).

DOI: 10.48175/IJARSCT-1509