Optimizing Crop Selection in Agriculture through Machine Learning-Driven Recommendation Systems

[1]G.Prasanthi ,[2]L. Surya Prakash, [3]D. Naga Durga, [4]Ch. Jahera,[5]T. Jagadeesh, [6]P. Sriram, [7]S. Dileep
[1,2,3,4,5,6,7]Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India.

---***---

Abstract - *In the present era, technology is playing a crucial role across various sectors to overcome challenges and achieve optimal results. In India, agriculture significantly influences the economy, employing half of the country's population. However, the agricultural industry faces challenges due to reliance on primitive practices and slow technological advancements. One major dilemma is deciding which crop is most suitable and profitable for their specific soil conditions, given the vast soil variations across regions. Effective technology has the potential to enhance yield and mitigate challenges. Often, farmers prioritize crops based on market value and financial gains, neglecting factors like soil conditions and sustainability. This project aims to leverage technologies such as machine learning to revolutionize crop recommendations, providing valuable assistance to farmers for improved outcomes. By, Simply inputting soil nutrient values into the system offers personalized recommendations on which crops or fruits to grow, coupled with suggested fertilizers for improved production. This innovative approach represents a stride towards sustainable and technology-driven agriculture, offering maximum assistance to farmers for informed decision-making..*

***Key Words*: Crop recommendation, Machine Learning, Decision Tree, SVM, Logistic Regression.**

1. INTRODUCTION

Other countries have begun to adopt and use contemporary techniques in order to profit financially. They already have a significant advantage over others in the field of farming and agriculture when it comes to using scientific and technological methods to improve the quality of their work. India, on the other hand, continues to use traditional farming methods and the associated technologies. It is often acknowledged that a substantial portion of the nation's income comes from agriculture alone. When determining the Gross Domestic Product's worth, income comes in rather handy. The need for food has increased dramatically as globalization becomes closer. In an effort to increase productivity, farmers apply a variety of chemical fertilizers, but this approach eventually damages the ecology. Effective crop production can be achieved by the farmer knowing precisely which crop to plant based on the soil composition and environmental factors. This will reduce crop loss. We have created a dataset comprising data on rainfall, weather patterns, and several soil nutrients. This will provide us a better grasp of agricultural production trends while taking environmental and regional aspects into account. Additionally, our technology forecasts any shortages of key ingredients needed to cultivate a given crop. The agricultural business stands to benefit greatly from our predictive method. With the assistance of our predictive method, the issue of nutrient shortage in areas that resulted from planting the wrong crop at the wrong time is eliminated. As a result, farmers production efficiency is decreased. The more science is used to agriculture, the more it will surely reach new heights.

We propose this method to teach farmers about the different types of minerals needed and the kinds of climate that are suitable for cultivating different kinds of crops. Furthermore, our research raises awareness of the lack of several minerals that are necessary for the growth of particular crops and offers remedies to close this gap. Our method considers climatic factors including temperature, precipitation, and humidity as well as soil characteristics.

Literary Survey

[1] Crop And Yield Prediction Model- Shreya S. Bhanose, Kalyani A. Bogawar. The crop recommendations in this report are based on productivity and season. The first step in the process is gathering data from many sources, such as the crop production area, crop production season, soil type, etc. Pre-processing of the gathered data, including data wrangling, has been completed in the following phase.

[2] Data Mining and wireless sensor network for agriculture Disease predictions. –Tripathy. This paper shows the relationships between most prevailed disease, pest and weather parameter of Groundnut crop. WSN was established in the test bed to obtain real-time weather parameters (Temperature, Humidity and Leaf wetness) at micro-climatic level and a few related weather parameters were taken from the nearby weather station.

[3] An Analysis of Agricultural Soils by using Data Mining Techniques– Ramesh Babu Palepu– This paper shows various problems faced in agriculture sector and how data mining solves the problem. Data mining applications are utilized by agricultural institutions for several purposes, including disease detection, pesticide optimization, and problem prediction, in order to make the best judgments possible.

[4] The Impact of Data Analytics in Crop Management based on Weather Conditions- A.Swarupa Rani. In this paper, the solution (Tillage) will help farmers increase agricultural output, decrease soil degradation in cultivated areas, have knowledgeable guidance on organic fertilizers/other fertilizers.As a result, both farmers and the environment would gain from this complete prediction.

2. System Design

To design software, we must grasp the concept of SDLC life cycles. Users can effectively manage and implement their systems with the use of a framework called the software development life cycle (SDLC). We decided to use the waterfall paradigm for our project. The life cycle notion is quite straightforward and easy to understand. The software is created using a sequential linear flow. This means that a process cannot start another until the previous one is completed. As a result, no processes overlap exists. The output of one segment is the input of the next component. The waterfall model includes the phases seen in "Fig.1". The phases that follow are all essential to the development of a system, and the output of each step depends on the step that came before it .

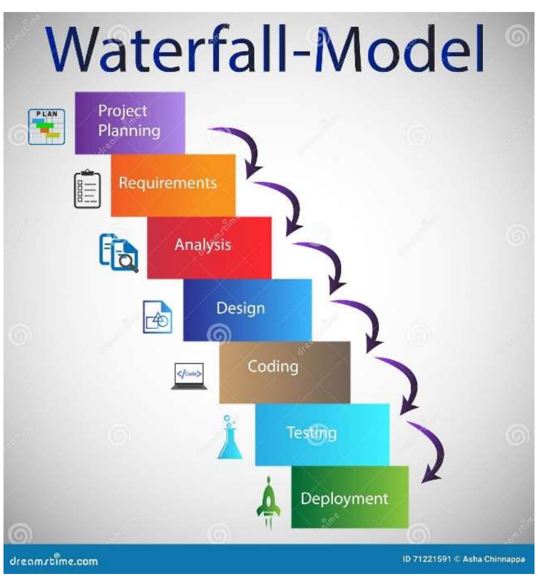


Fig. -1: Waterfall Model

2.1 Model Framework Overview

The following "Fig. 2" illustrates the key phases in creating a machine learning-based predictive model.

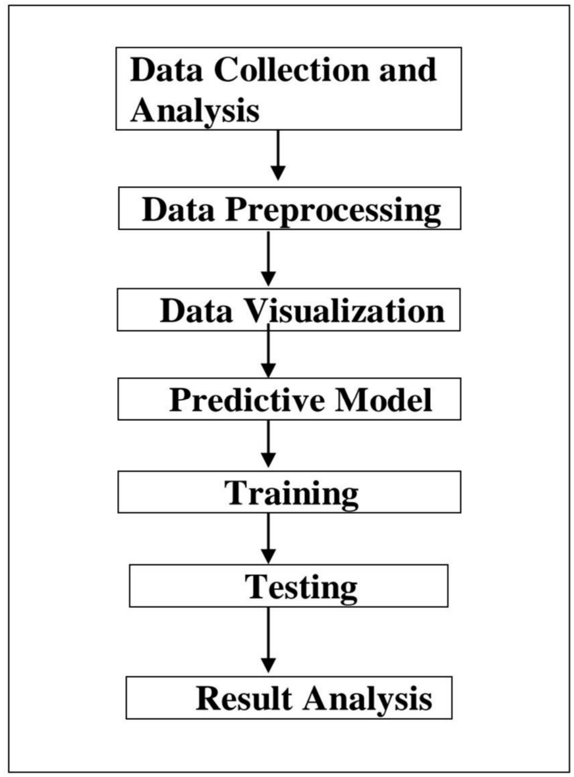


Fig. - 2: Project Framework

2.2 System Architecture

The structure and behavior of our project are defined by a conceptualization-oriented model called a system architecture. An architecture is a system's formally descriptive and representative structure that is arranged to support the logic behind all of the system's behaviors and structures. It also illustrates how dependent processes are on one another. A well-thought-out system architecture makes it easier for us to learn and comprehend the system graphically. It facilitates our understanding of our product in a way that makes it simpler for us to comprehend and connect to the key features and phases of a system. "Fig. 3 shows the system architecture of our recommendation model.

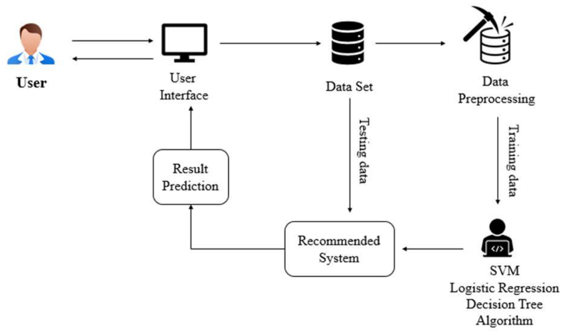


Fig. -3: System Architecture

3. Methodology and Implementation

The steps and procedures that need to be followed in order to put our prediction system into practice are listed below in chronological order.

3.1 Dataset Collection

Gathering datasets is the first action we do while developing a machine learning project. The raw data in the dataset that we get from several platforms is rife with inconsistencies and ambiguities.

In this project, data has been acquired from the open-source platform Kaggle, comprising two fundamental datasets: the Soil Content Dataset, detailing the ratios of Nitrogen (N), Phosphorous (P), and Potassium (K), as well as soil pH; and the Climatic Condition Dataset, encompassing information on rainfall, humidity, and temperature. The amalgamation of these datasets has resulted in a final dataset with approximately 2200 rows and 8 columns, as depicted in "Fig. 4." This integrated dataset plays a pivotal role in agricultural analysis, providing a comprehensive understanding of the interplay between soil composition and climatic factors. Leveraging Kaggle's resources, this project aims to harness the collective insights from both datasets to inform decision-making in the agricultural domain, with the visual representation in Figure 4 serving as a key reference for the dataset's structure and attributes..

N	P	K	temperat	humidity	ph	rainfall	label
79	34	37	20.87974	82.00274	6.502985	202.9355	rice
78	21	20	21.77046	80.31964	7.038096	226.6555	rice
92	31	34	23.00446	82.32076	7.840207	263.9642	rice
98	38	45	26.4911	80.15836	6.980401	242.864	rice
79	53	43	20.13017	81.60487	7.628473	262.7173	rice
65	50	36	23.05805	83.37012	7.073454	251.055	rice
75	23	28	22.70884	82.63941	5.700806	271.3249	rice
95	35	21	20.27774	82.89409	5.718627	241.9742	rice
92	30	25	24.51588	83.53522	6.685346	230.4462	rice
69	35	60	23.22397	83.03323	6.336254	221.2092	rice
75	54	39	26.52724	81.41754	5.386168	264.6149	rice
98	36	44	23.97898	81.45062	7.502834	250.0832	rice
86	52	48	26.8008	80.88685	5.108682	284.4365	rice
86	20	31	24.01498	82.05687	6.984354	185.2773	rice
75	22	56	25.66585	80.66385	6.94802	209.587	rice
66	35	39	24.28209	80.30026	7.042299	231.0863	rice
61	46	51	21.58712	82.78837	6.249051	276.6552	rice
69	45	28	23.79392	80.41818	6.97086	206.2612	rice
80	26	44	21.86525	80.1923	5.953933	224.555	rice
100	46	34	23.57944	83.5876	5.853932	291.2987	rice
62	49	37	21.32504	80.47476	6.442475	185.4975	rice

Fig. -4: Final Crop Dataset

3.2 Data Analysis

A vital first step is a thorough investigation and comprehension of the data before moving further with the preprocessing stage. With great care, this initial examination is carried out in order to identify relevant characteristics from the dataset that will be essential factors in determining crop outcomes in later phases. This methodical procedure is essential in forming the basis of the prediction model. The immediate impact of this step on the precision and dependability of crop forecasts explains its importance. The complexities of our recommender system's data analysis procedure are depicted in a heatmap in "Fig. 5". This visual tool makes it easier to understand the links and patterns in the information, which enables decision-making about feature selection. It also highlights how important it is to handle this crucial stage carefully and methodically. The recommender system development process is guided by the heatmap, which is an effective tool for visualizing the data's underlying structure.

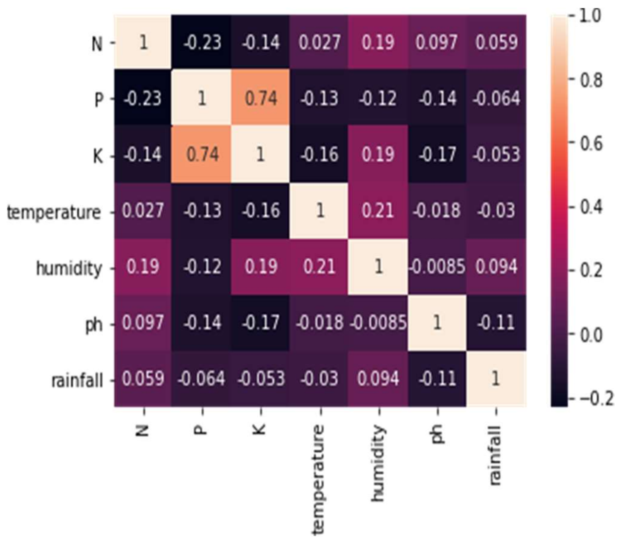


Fig. -5: Data Analysis as Heatmap

4. Data Visualization

The production quantity of various police officers has been plotted against the elements influencing their production. This will teach us about the ways in which different soil and climate elements influence the amount of produce. As seen in "Fig.6, the data that we plotted is displayed as bar graphs.".

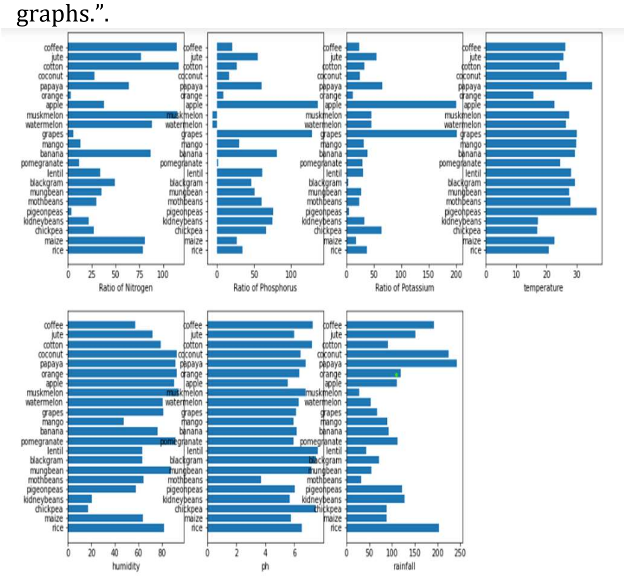


Fig. -6: Data Visualization

5. Machine Learning Algorithms

Several algorithms were trained and evaluated because our crop prediction system is a kind of classification problem. We make sure that precision is given top priority in order to obtain optimal accuracy and the best crop to be seeded. We evaluated each model's accuracy and determined which had the best overall.

5.1 Decision Tree

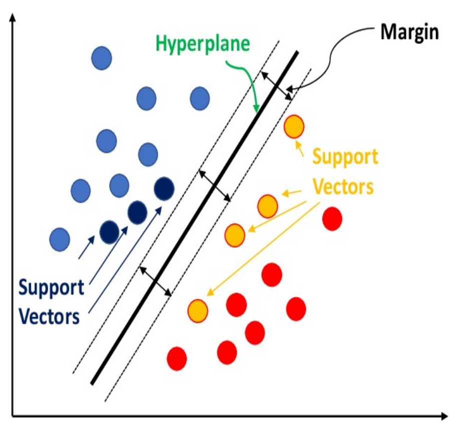
The first machine learning method we used to build our target crop system was the Decision Tree. Actually, the graphical representation and the parsing pattern it follows are what gave rise to its name. The leaf and decision nodes are the two fundamental types of nodes that comprise this system. Decision nodes are nodes that branch and make decisions. Conversely, leaf nodes are nodes that show the outcomes of decisions made under particular circumstances.

The accuracy of the tested model is found to be roughly 75.78%. Because of its poor precision, we haven't used it, but as a result, its accuracy is not too awful.

5.2 Support Vector Machine

In order to accurately categorize new data in the future, SVM is a machine learning technique that builds an optimal decision boundary or hyperplane to divide dimensional spaces into classes. With the help of support vectors, we can build a hyperplane. Consequently, each side of the generated hyperplane has two support vectors. The support vectors are the lines that are drawn between the two data points on either side that are closest to the hyperplane.

The accuracy of this model is roughly 96.08%. Consequently, this model outperforms the Random Tree approach in terms of accuracy. "Fig.7." displays the SVM classifier.



Fig, -7: SVM Classifier

Next, a pickle file is created using the SVM model. Then, this pickle file is used to implement and run as a website on our local server using flask, html, css, and a little bit of basic javascript..

5.3 Logical Regression

One ML algorithm that models the relationship between dependent and independent variables is logical regression. It is mostly used to solve problems with categorical data. It is a very basic, effective paradigm that is easy to implement.

The logical regression model has a 88.88% accuracy rate. While it is less accurate than SVM, it is still better than random tree algorithm. As a result, this model is dropped..

5.4 Random Forest

We have already applied, studied, and experimented with the decision tree approach. It won't be too hard to understand the random forest method, then. It is an extremely well-liked ensemble learning algorithm. It is a composite of many randomly selected trees from various dataset subnets. A random forest network with an abundance of random trees has a high model accuracy. The tree with the most votes is chosen to make the decision. Therefore, the accuracy of the model increases with the number of decision trees in the forest. This also solves the overfitting issue.

The random forest method has the best accuracy, at 93.73% percent. Finally, we use this approach to the creation of our model and the launch of our project, which is a website. The random forest model is depicted in "Fig. 8" below. Generally speaking, a forest is made up of several types of trees. In this case as well, we can state that random forests are composed of different decision tree subnets.

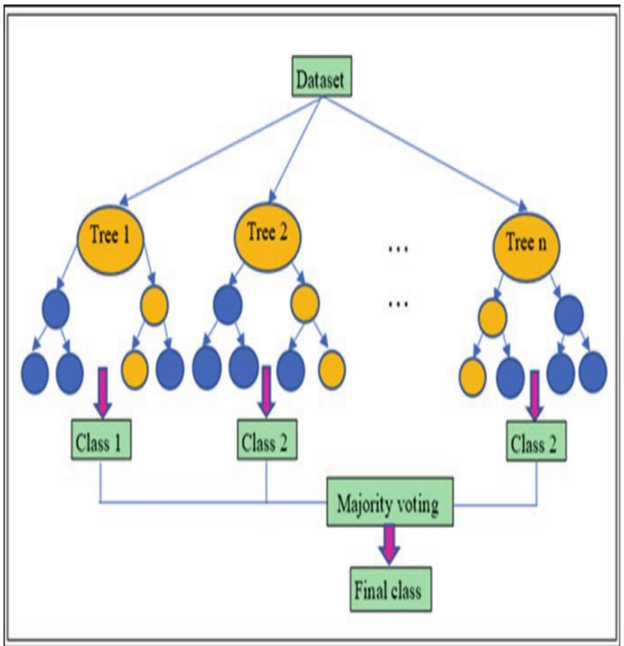


Fig. -8: Random Forest

6. Testing

Here, we have benefited from The cross validation training method is one way to evaluate how well the arithmetic model generalizes to different datasets. K-folding cross validation has been used in this instance. In this type of cross validation, we only need to train on about one set of data; the remaining sets are used to train our mold.

Furthermore, all of the separately constructed functional components of the program are integrated into our crop recommendation system project. Several steps were taken in compliance with the architecture of the system. In this study, we have experimented with and assessed several machine learning algorithms to provide farmers a dependable and accurate model for crop prediction. Each model is tested to make sure it satisfies a set of specifications, such as FP rate, accuracy, and precision. We have also looked at the cross-validation scores of all the methods we tried to utilize to build this system.

We evaluate multiple performance metrics for our clone of the recommendation algorithm. Below are the explanations for each performance attribute..

6.1 Accuracy

The most important consideration when assessing a machine learning model is accuracy. We determine whether or not our machine learning model is relevant in the actual world based on accuracy. When an algorithm performs well, it eventually indicates that the system is becoming more like the real world.

$$\textbf{Acc.} = \frac{\text{No. of Correct Predictions}}{\text{Total No. of Predictions}}$$

Also, it is evaluated as:-

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

6.2 Precision

Precision is calculated by simply plotting the confusion matrix. From that, we may determine the values of TP, FP, TN, and FN. We calculate the sum of True Positive and False Positive as well as the True Positive value in the numerator to assess precision.
Formula is :-

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

6.3 Recall

When we take the sum of True Positive and False Negative values as well as the True Positive value in the numerator. It displays the percentage that the real positive is equal to.

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

7. Result and Performance Analysis

This dataset is subjected to a thorough examination, and during the testing phase, a variety of outcomes are produced to verify the model's accuracy and help us develop a recommendation system for the best harvest. Additionally, our system forecasts a lack of nutrients. The Jupyter notebook results are displayed in "Fig.9" below.

```
# 21st label is jute!

data = np.array([[83, 45, 60, 28, 70.3, 7.0, 150.9]])
prediction = svc.predict(data)
pred = prediction[0]
print(labels_map_new[pred])


data = np.array([[104,18, 30, 23.603016, 60.3, 6.7, 140.91]])
prediction = svc.predict(data)
pred = prediction[0]
print(labels_map_new[pred])

jute
coffee
```

Fig. -9: Results of Prediction in Jupyter Notebook

Appropriate analysis and visualization of the data are done to fully comprehend the various attributes that impact crops. Various machine learning algorithms are techniques that have undergone accuracy checks and tests. In "Fig. 10," the correctness proportion of several algorithms is displayed and shown.The accuracy percentages are presented as a bar graph.

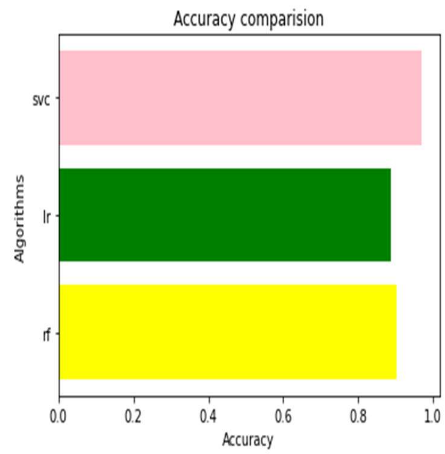


Fig. -10: Accuracy comparision of Algorithms.

To fully comprehend the various attributes that effect crops, the data is appropriately processed and visualized. It is possible to find many machine learning algorithms that have undergone accuracy checks and tests. "Fig. 10" plots and describes the correctness proportion of several algorithms.A bar graph is used to illustrate these accuracy percentages.

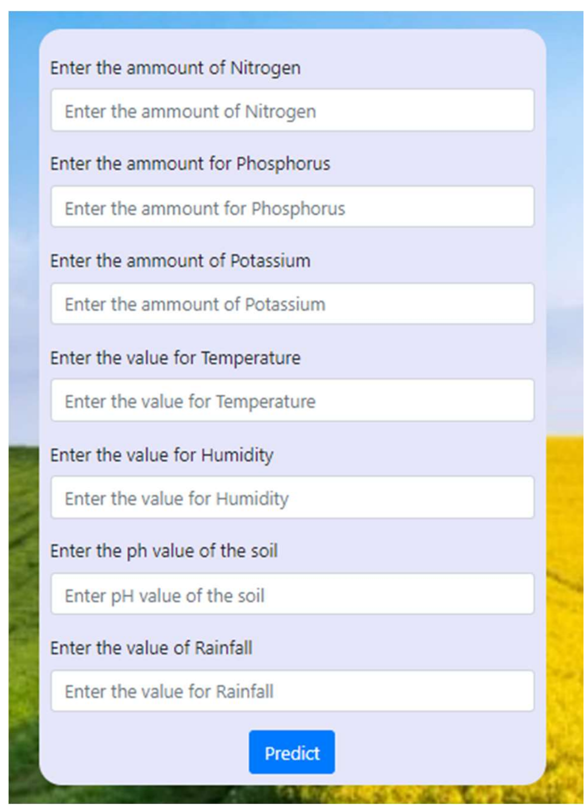
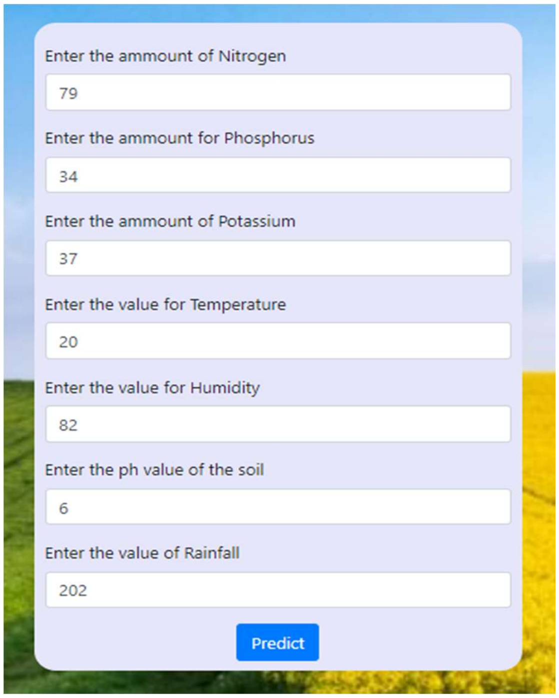


Fig. -11: Interface of Crop Prediction System

The output of our crop recommendation system is shown in "Fig. 12". Our algorithm makes predictions about the best crop to grow within the user interface.



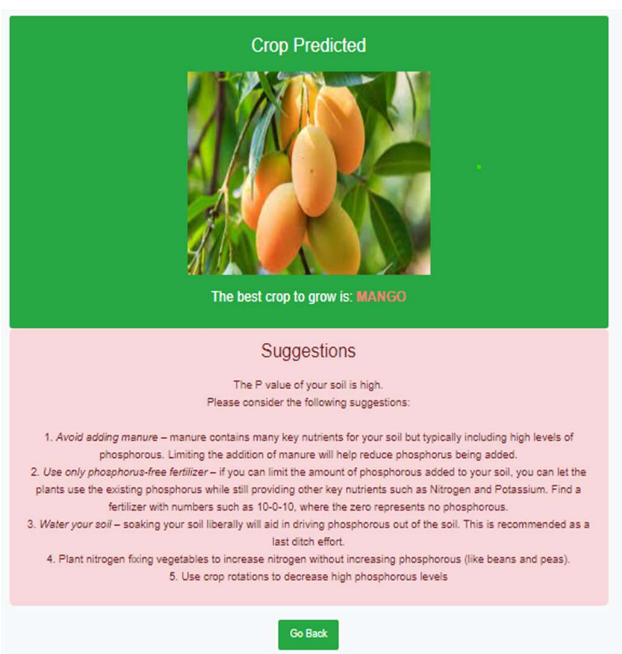


Fig. -12: Result of Predicted Crop

8. CONCLUSIONS

Since agriculture is essential to the economy of our nation, even the smallest investment made in this area has a significant impact on the nation as a whole. As a result, we must take it more seriously. The farmers in our nation often have great difficulties in choosing which crops to plant because of the paucity of scientific information regarding the various aspects influencing crops. so, as a result of lower productivity, experience a decline in profit. However, our approach will provide them a chance to cultivate crops that will yield the highest profit. Both the volume and quality of their output will rise dramatically. Additionally, it will support them in preserving the soil's nutrient content. Both the quantity and quality will be increased.

9. Future Work

The objective of our system is making a robust model. Also, trying to incorporate bigger dataset. I'm attempting to make my website's UI and CSS better in the future. Additionally, I'm working to add new features to my product, such the ability to predict plant diseases. attempting to understand and use web scraping as a data collection method.

REFERENCES

[1] https://en.wikipedia.org/wiki/Agriculture_in_India.

[2] https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[3] Shreya S. Bhanose, Kalyani A. Bogawar (2016) “Crop And Yield Prediction Model”, International Journal of Advance Scientific Research and Engineering Trends, Volume 1,Issue 1, April 2016.

[4] Tripathy, A. K., et al.(2011) "Data mining and wireless sensor network for agriculture pest/disease predictions." Information and Communication Technologies (WICT), 2011 World Congress on IEEE..

[5] Ramesh Babu Palepu (2017) ” An Analysis of Agricultural Soils by using Data Mining Techniques”, International Journal of Engineering Science and Computing, Volume 7 Issue No. 10 October.

[6] A.Swarupa Rani (2017), “The Impact of Data Analytics in Crop Management based on Weather Conditions”, International Journal of Engineering Technology Science and Research, Volume 4,Issue 5,May.

[7] S. Pudumalar, E. Ramanujam, R. Harine Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop Recommendation System for Precision Agriculture", IEEE International Conference on Advanced Computing, pp. 32-36, 2017.

[8] R. Kumar, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to Maximize Crop Yield rate using Machine Learning Technique", IEEE International Conference on Smart Technologies and Management for Computing Communication Controls Energy and Materials, pp. 138-145, 2015.

[9] N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy, "Rice Crop Yield Prediction in India using Support Vector Machines", IEEE International Joint Conference on Computer Science and Software Engineering, pp.