CHAT BOT FOR CROP YIELD PREDICTION AND RECOMMENDATION THROUGH K-NEAREST NEIGHBOUR ALGORITHM

1J.Surya, 2B. Prabha   
*Departement of Computer Sciernce and Engineering,*   
*SRM Institute of Science and Technology,* Chennai, India.

*1*[*suryajeyasubramanian@gmail.com*](mailto:suryajeyasubramanian@gmail.com) *, 2jemi.prabha@gmail.com*

***Abstract***—**The majority of the population relies heavily on agriculture for their food, jobs, income, foreign exchange, and raw materials for the manufacturing industries, making it the backbone of the nation's economy. Farmers need to maintain quality and quantity controls over crop cultivation in order to meet market needs. Therefore, in order to cultivate and preserve crops, farmers need to be knowledgeable. Agriculture plays a major role in India. India is the world's leading producer of a wide range of crops. But the agriculture sector's yield is still insufficient if improved techniques aren't used. There's no trustworthy referral system for Indian farmers at the moment.  maximise output and profit, they confront a great challenge in figuring out which crop is best for their agricultural region. Indian farmers frequently have two difficulties: (i) choosing appropriate crops based on soil characteristics; and (ii) determining the soil's health. Because of this, their production has taken a significant hit. By addressing this problem, farmers are able to optimise resource allocation and prepare for future dangers. Crop yield prediction and crop recommendation play a vital part in agricultural decision-making processes. Accurate crop yield predictions may now be made with greater ease thanks to machine learning algorithms. The use of the Decision Tree method to crop yield prediction training is the main topic of this abstract. Crop yields for unknown data can be predicted using the Decision Tree model once it has been built. To get yield predictions and crop recommendations, the model can be updated with new input data, such as soil measurements or weather forecasts. Farmers can make educated judgements by knowing which factors contribute most significantly to differences in crop yield thanks to Decision Trees' interpretability. The chatbot allows users to communicate by providing details, after which the model predicts the crop and responds. By providing farmers with a prioritised list of crops and an improved user interface, the trained model proved beneficial in meeting their needs. Random Forest, K-Nearest Neighbour, and Decision Tree models are applied for the proposed chatbot development. The comparison of the output from the three algorithms determines the accuracy of crop forecast and advice. The input data is processed through three different machine learning algorithms, and the outcome is selected by comparing the accuracy of the results obtained from each approach. Similarly, comparing the output of three machine-learning algorithms is used to calculate the standard deviation of crop forecast and recommendation.**

**Index terms-Machine learning, Decision tree, K-Nearest Neighbour, Random Forest, Chatbot.**

1. INTRODUCTION

The principal aim of the project is to develop a chatbot that utilises a decision tree algorithm to estimate crop yield and suggest crops. The majority of the population relies heavily on agriculture for their food, jobs, income, foreign exchange, and raw materials for the manufacturing industries, making it the backbone of the nation's economy. Farmers need to maintain quality and quantity controls over crop cultivation in order to meet market needs. Therefore, in order to cultivate and preserve crops, farmers need to be knowledgeable. India is largely dependent on agriculture. India is the world's leading producer of a wide range of crops. Yet, the agricultural sector's yield is still insufficient if improved techniques aren't employed. There isn't a trustworthy referral system in place for Indian farmers. Choosing the suitable crop to grow in their area to increase profit and production is a major challenge for them. Indian farmers frequently have two difficulties: (i) choosing appropriate crops based on soil characteristics; and (ii) determining the soil's health. They are consequently facing a significant decline in productivity. By addressing this problem, farmers are able to optimise resource allocation and prepare for future dangers. Crop yield prediction and crop recommendation play a vital part in agricultural decision-making processes. Machine learning algorithms have become highly effective tools for precisely forecasting crop yields in recent times. The decision tree algorithm's usage in agricultural yield prediction training is the main topic of this abstract. Crop yields for unknown data can be predicted using the Decision Tree model once it has been built. To get production estimates and crop recommendations, the model can be updated with new input variables, such as soil measurements or weather forecasts. Farmers can make educated judgements by knowing which factors contribute most significantly to differences in crop yield thanks to Decision Trees' interpretability. The user can communicate with the chatbot by providing details, after which the chatbot will forecast the model's harvest and respond. By providing farmers with a prioritised list of crops and an improved user interface, the trained model proved beneficial in meeting their needs. Hardware specifications include a minimum RAM size of 8 GB and an Intel Core i5 or i7 CPU with a 3 GHz processing speed. It is a 500 GB hard disc. It makes use of a Standard Window keyboard. We're using an SVGA monitor. Windows 7/10 is the minimum operating system needed for software. Jupiter, vscode, and Anaconda are the necessary tools. Python and AIML are needed for the server-side script, and HTML and SS are needed for the front end. Helping farmers and other agricultural stakeholders make informed decisions about crop cultivation is usually the goal of this initiative.

**Contribution:**

1. The main goal of implementing a chatbot that uses decision trees to predict crop output and provide crop recommendations is to provide farmers with useful information and practical advice that will maximise crop yield, reduce risks, and improve agricultural sustainability.
2. Using a variety of data sources, including past yield records, weather trends, soil conditions, and crop management techniques, the chatbot seeks to forecast agricultural yields. Using decision trees as a predictive model, one may forecast agricultural yields by analysing past data. Planting schedules, resource allocation, and marketing tactics can all be more effectively planned by farmers when crop yields are predicted with accuracy. After analysing a number of variables, including soil type, climate, market demand, and resource availability, the chatbot makes recommendations for what kinds of crops to cultivate.
3. The most likely crops to flourish in a certain area or under particular environmental conditions can be identified with the use of decision trees, which aid in the analysis of large, complicated datasets. Recommendations may also take into account elements like profitability, insect resistance, and crop rotation techniques. Decision trees are a popular choice for agricultural dataset analysis because of their versatility in handling both numerical and categorical data. Their transparency and interpretability enable consumers to comprehend the logic underlying the chatbot's recommendations.
4. The chatbot can assess various aspects and make conclusions based on a set of metrices, emulating decision-making process of human experts, thanks to the decision trees' hierarchical structure. Farmers and other agricultural stakeholders can communicate with the chatbot to ask for advice and knowledge about crop cultivation. The chatbot can be used to improve user engagement and build user trust and reliability by offering precise forecasts and personalised recommendations. By interacting with the chatbot, users may get up-to-date market trends, real-time information, and advice on crop management best practices, all of which increase overall agricultural profitability and productivity.
5. RELATED WORKS

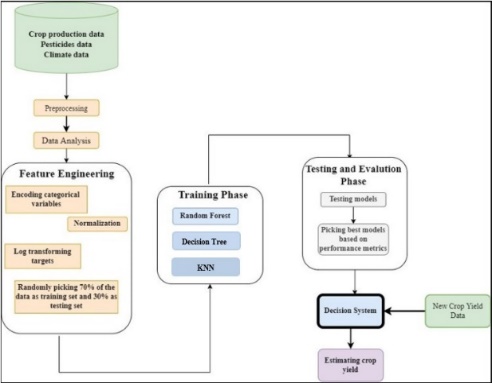
Experiments with support vector machines for high-resolution picture road recognition, Lai *et al.,* [1], In recent years, the pattern recognition community has given support vector machines a lot of attention. When used on several classical recognition tasks, they have shown results that are on par with or even better than those of classifiers like neural networks. We examine the use of edge-based features in Support Vector Machines (SVMs) to address the issue of road detection from remotely sensed photos. Comparing our experimental results to neural network classifiers and decision tree, we show highly encouraging results. Temperature and CO2 levels across four seasons' worth of winter wheat crops: an analysis, Batts *et al.*, [2], Four seasons of winter wheat were grown in the field at Reading, UK, from 1991/1992 to 1994/1995. Across the temperature gradient, the mean seasonal temperature varied by as much as 4°C. Reproductive development rate was temperature-sensitive, and neither crop length nor rate of development was affected by increased [CO2]. A 2°C increase in the ambient mean temperature during a 4-year period (10°C) resulted in a 42-day crop duration decrease (from 254) and a 16-day reproductive phase reduction (from 130). Generally speaking, crop biomass decreased as mean temperature rose and increased in response to elevated CO2 levels. However, the impact of elevated CO2 varied according to temperature and year, with a relative stimulation range of 6–34%. Temperature increases of 7–168% ranged in their impact on grain output, whereas warmer temperatures significantly decreased it. However, doubling [CO2] had the opposite effect. Increased [CO2] and temperature had both beneficial and detrimental effects on grain yield and biomass. A mere 1.0–2.0°C rise in the mean seasonal temperature offset the grain yield gain caused by the doubled [CO2] in each of the four years. Differences in climatic circumstances affected biomass partitioning and changed the roles of various yield components, resulting in changes in the year variance in the biomass responses and grain yield to temperature and [CO2]. Regression model comparison in data mining for agricultural yield data, Rub *et al.*, [3], These days, precision agriculture involves treating crops on a small scale using sensors and cutting edge GPS technology. This introduces a lot of data that is gathered and saved for use at a later time. Economic benefits and significant increases in efficiency are frequently the results of making effective use of these data. Data mining techniques should be employed to address the problem of the volume of data. Produce forecast based on available data is one of the unfinished challenges. One way to formulate and approach this from a data mining standpoint is as a multi-dimensional regression task. In order to assess suitable regression procedures, this research uses a sample of farm data to assess four distinct techniques. An advice regarding a particular method is given. Forecasting air pollution occurrences in the short term, Jorquera *et al.*, [4], Secure authorization for applications operating on many types of platforms is made possible by OAuth 2.0, a delegated authorization framework. When it comes to medical services, OAuth allows the patient seeking immediate clinical attention to approve automatic monthly payments from his bank account without having to give the clinic his credentials. Access tokens have a time limit and are temporary to maintain security. To give monthly payments for providing real-time medical services, the clinical app may use a new token to acquire a new access token. Refresh tokens must be stored safely to prevent leaks, as any nefarious actor could exploit them to gain fresh access credentials. OAuth 2.0 is susceptible to phishing attacks when attempting to access interoperable APIs because it no longer uses signatures and is entirely dependent on SSL/TLS. In this work, we develop a method where a client wishing to use a prior authentication and permission can request an OAuth access token from the authorization server by combining JSON web token (JWT) with OAuth 2.0. An IOT cloud platform for health care services has high security requirements, and experimental analysis demonstrates that the suggested system is realistically efficient, eliminates secure storage overhead by eliminating the need for a refresh token or to keep one, leverages signatures, and prevents various security assaults. An approach to assessing apples in actual time that takes characteristics out of flaws, Leemans *et al.,* [5], This study presents a multi-feature information fusion method based on D-S evidential theory and BP neural network to boost apple grading accuracy, as results that evaluate apples based on a single feature—like size, shape, or colour—are incorrect. Initially, the processed images of apples are used to extract attributes related to size, shape, and colour. The BP network classifier is used to classify apples based on several features. The results of the classifiers are then merged to create the basic probability assignment (BPA), which serves as independent evidence. In order to arrive at the final grading result, the choice is made utilising the D-S fusion rules of evidence. Based on size, shape, or colour features, The experimental findings show that, in terms of accuracy, the decision data combination method performs better in apple grading than the single feature-based method.. Speaker identification analysis, feature extraction, modelling, and testing methods, Jayanna *et al.*, [6], utilising the ANSI/AAMI SP10:2002 Standard examination to verify a new generation blood pressure monitor that is substantially less expensive. The method outlined in the SP10: 2002 American National Standard was used to evaluate the Spot Vital Signs gadget. Participants in the study who were obese or hypertensive were overrepresented. When the device readings were compared to auscultation by paired, trained, blinded observers, the mean+/-standard deviation for the systolic blood pressure was -1.0+/-4.1 mmHg, and for the diastolic blood pressure, it was -0.4+/-6.1 mmHg. They satisfied AAMI requirements. The very accurate device came at a much lower cost than similar professional-grade (as opposed to self-measurement) electronic blood pressure devices. For healthcare facilities where both cost and accuracy are top concerns, the Welch Allyn Spot Vital Signs system offers a more affordable option. In order to optimise the diagnosis and therapy of cardiovascular disease, the study population was oversampled with patients who were obese or hypertensive, the main target groups. In these individuals, reliable blood pressure recording is essential. Machine learning techniques' capacity to anticipate large agricultural yields, Alberto *et al.*, [7], Precise yield estimation for the many crops in agricultural planning is a critical topic. To find workable and efficient solutions to this issue, machine learning (ML) is a crucial strategy. The most accurate machine learning methodology has been the subject of numerous comparisons. For the purpose of agricultural planning, the number of assessed crops and methods is typically too small and offers insufficient data. The prediction accuracy of agricultural yield using ten crop datasets is compared between machine learning and linear regression techniques in this research. Methods such as support vector regression, k-nearest neighbour, perceptron multilayer neural networks, multiple linear regression, and M5-Prime regression trees were ranked. The models were validated using four different accuracy metrics: correlation factor (R), normalised mean absolute error (MAE), root relative square error (RRSE), and root mean square error (RMS). To develop the models, actual data from a Mexican irrigation zone were used. Samples from two years in a row were used to test the models. The findings demonstrate that the M5-Prime and k-nearest neighbour approaches yield the lowest average correlation factors (0.41 and 0.42) as well as the lowest average RMSE errors (5.14 and 4.91), RRSE errors (79.46% and 79.78%), and average MAE errors (18.12% and 19.42%). For large-scale crop yield prediction in agricultural planning, M5-Prime is an excellent tool since it produces the most crop yield models with the fewest errors. Learning techniques, topologies, and stability of recurrent neural networks for prediction, Mandic *et al.*, [8], For agricultural planning purposes, estimating yields accurately for the many crops in the plan is significant. To find workable and efficient solutions to this issue, machine learning (ML) is a crucial strategy. The most accurate machine learning methodology has been the subject of numerous comparisons. For the purpose of agricultural planning, the number of assessed crops and methods is typically too small and offers insufficient data. The prediction accuracy of agricultural yield using ten crop datasets is compared between machine learning and linear regression techniques in this research. Methods such as support vector regression, k-nearest neighbour, perceptron multilayer neural networks, multiple linear regression, and M5-Prime regression trees were ranked. Root mean square error (RMS), root relative square error (RRSE), normalised mean absolute error (MAE), and correlation factor (R) were the four accuracy measures that were utilised to validate the models. To develop the models, actual data from a Mexican irrigation zone were used. Two-year samples were used to evaluate the models. The findings demonstrate that the M5-Prime and k-nearest neighbour approaches yield the lowest average correlation factors (0.41 and 0.42) as well as the lowest average RMSE errors (5.14 and 4.91), RRSE errors (79.46% and 79.78%), and average MAE errors (18.12% and 19.42%). A highly useful tool for enormous crop yield prediction in agricultural planning, M5-Prime produces the greatest number of crop yield models with the lowest mistakes. Extended short-term memory. neural computing, Hochreiter *et al*., [9], Due to insufficient decreasing error backflow, recurrent backpropagation takes some time to learn how to store data over extended periods.. After a quick recap of Hochreiter's (1991) study, we introduce long short-term memory (LSTM), a novel, effective gradient-based approach, to solve this issue. LSTM can be trained to connect minimal time gaps more than 1000 discrete-time steps by truncating the gradient when necessary and imposing constant error flow using constant error carousels within special units. Units with multiplicative gates get the ability to open and close the constant error flow. With an O(1) computational cost per time step and weight, LSTM is local in both space and time. We use real-valued, distributed, local, noisy pattern representations in our simulated data studies. LSTM learns more more quickly and produces many better runs when compared to other methods as neural sequence chunking, recurrent cascade correlation, backpropagation in time, Elman nets, and real-time recurrent learning.. Recurrent network techniques prior to LSTM were unable to perform complicated, fake long-time-lag challenges. Massive acoustic modelling using long short-term memory recurrent neural network designs, Sak *et al.*, [10], Long Short-Term Memory (LSTM) recurrent neural network architecture was created to more accurately mimic time sequences and their long-range communications than usual RNNs.. For large-scale acoustic modelling in voice recognition, we investigate LSTM RNN architectures in this study. We have demonstrated that, when considering moderately-sized models trained on a single machine, LSTM RNNs outperform DNNs and traditional RNNs for acoustic modelling. In this work, we present the first distributed training of LSTM RNNs on a large cluster of machines via asynchronous stochastic gradient descent optimisation. We demonstrate that state-of-the-art voice recognition performance can be surpassed by a two-layer deep LSTM RNN with a linear recurrent projection layer in each LSTM layer. This design converges faster, uses model parameters more efficiently than the others, and performs better than a massive feed-forward neural network that is deep in additional options. Regression analysis and classification using random forests, Wiener *et al.*, [11], "ensemble learning" techniques, which produce numerous classifiers and combine their output, have attracted a lot of attention lately. Two popular techniques are bagging Breiman (1996) of classification trees and boosting (see, e.g., Shapire et al., 1998). When boosting, subsequent trees assign more weight to points those previous predictors got wrong. The prediction is finally decided by a weighted vote. Following trees in bagging are independently built using a bootstrap sample of the data set; they are not dependent on preceding trees. The forecast is finally decided by a simple majority vote. random forests, which give bagging an extra degree of unpredictability. Random forests modify the construction of the classification or regression trees in addition to use distinct bootstrap samples of the data for each tree. Using the best split among all variables, each node in a normal tree is split. To split each node in a random forest, the best subset of randomly selected predictors at each node will be used. Compared to many other classifiers, including discriminant analysis, support vector machines, and neural networks, this seemingly odd method shows out to be highly effective and stable against overfitting. It is also easy to use because it needs just two parameters: the total number of trees in the forest and the number of variables in the random subset at each node. Also, it is usually not very sensitive to the values of these parameters.

1. METHODOLOGY

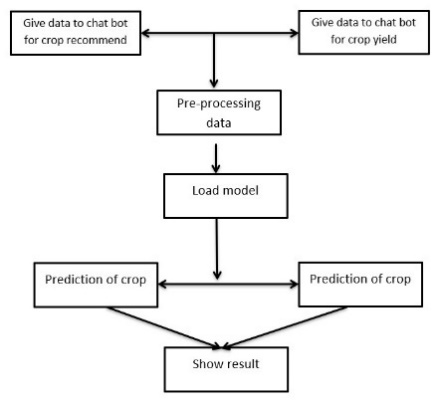
The suggested system's mechanism. Accuracy rates of several combinations were compared concurrently. We've included the KNN, AdaBoost, Random Forest, and Decision Tree regression algorithms. Count\_name, Year, Yield\_value, Average\_rainfall, Pesticide\_tons, and Average\_temperature are the parameters that we have taken into account. The best ensemble model out of all the models we looked at was determined by comparing the metric values of each ensemble. Once the optimal ensemble model has been identified, the system is prepared for the prediction model to be created. A few drawbacks include: 1) Lower accuracy than the suggested system; 2) Text box details and forecast accuracy; and 3) No crop recommendation.

***A. Proposed Architecture***

First, the system gathers historical agricultural data, which includes variables like weather patterns, soil properties, crop management techniques, and historical production records of eligible crops. Next, a prediction model is built using the Decision Tree algorithm. By using recursive partitioning, it creates a tree structure with internal nodes that reflect decisions based on certain input variables and leaf nodes that represent expected crop production outcomes. The crop yield forecast and crop recommendation system is now straightforward to use thanks to the incorporation of a chatbot. The Chatbot can communicate with researchers, farmers, and other stakeholders, doing away with the need for intricate data analysis or specialised technical expertise. Benefits include: 1) Higher prediction accuracy is provided by the decision tree method; 2) Users can interact with chatbots to forecast formation.



**Fig1.** System architecture of the proposed crop prediction and recommendation



**Fig2.** Data model of the proposed crop prediction

In a software development project, it's crucial to organize the code into modular components to enhance maintainability, reusability, and collaboration among team members. Here are some key modules for the "Chat Bot for Crop Yield Prediction and Crop Recommendation" project:

1. ***Data Collection and Preprocessing Module:***

This module is responsible for collecting, cleaning, and preprocessing data from various sources, including crop yield data, soil data, and weather data. It includes submodules like:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | State\_Name | District\_Name | Crop\_Year | Season | Crop | Temperature | humidity | soil moisture | area | Production |
| 0 | Andaman and Nicobar Islands | NICOBARS | 2000 | Kharif | Arecanut | 36 | 35 | 45 | 1254.0 | 2000.0 |
| 1 | Andaman and Nicobar Islands | NICOBARS | 2000 | Kharif | Other Kharif pulses | 37 | 40 | 46 | 2.0 | 1.0 |
| 2 | Andaman and Nicobar Islands | NICOBARS | 2000 | Kharif | Rice | 36 | 41 | 50 | 102.0 | 321.0 |
| 3 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Banana | 37 | 42 | 55 | 176.0 | 641.0 |
| 4 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Cashewnut | 36 | 40 | 54 | 720.0 | 165.0 |

**Table1.** Different crop data from Andaman and Nicobar Islands

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Area  (sq. feet) | N | P | K | temperature | humidity | ph | rainfall | Crop |
| 1270 | 6 | 140 | 205 | 17.665584 | 82.929034 | 6.313086 | 69.867126 | grapes |
| 1481 | 98 | 22 | 47 | 29.072653 | 91.915332 | 6.341401 | 28.835684 | muskmelon |
| 1832 | 38 | 14 | 30 | 26.924495 | 91.201060 | 5.570745 | 194.902214 | coconut |
| 293 | 35 | 63 | 76 | 17.815645 | 17.607566 | 7.714153 | 90.820976 | chickpea |
| 1307 | 85 | 22 | 53 | 25.965342 | 89.770767 | 6.849472 | 59.463386 | watermelon |

**Table2. C**rop recommendation based on the land data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Crop\_Year | Temperature | humidity | soil moisture | area |
| 9931.000000 | 9931.000000 | 9931.000000 | 9931.000000 | 9931.000000 |
| 2006.076025 | 34.445675 | 44.773034 | 53.108146 | 13299.452361 |
| 5.153237 | 3.499294 | 6.662943 | 5.259584 | 46476.817881 |
| 1997.000000 | 25.000000 | 35.000000 | 45.000000 | 0.200000 |
| 2002.000000 | 34.000000 | 40.000000 | 50.000000 | 160.000000 |
| 2006.000000 | 36.000000 | 42.000000 | 54.000000 | 1071.000000 |
| 2011.000000 | 36.000000 | 50.000000 | 55.000000 | 6265.500000 |

**Table3.** Data Collection for crop prediction

***Data Scraper*:** Collects data from external sources.

***Data Cleaner*:** Handles missing values and data quality issues.

***Feature Engineer*:** Extracts relevant features from the collected data.

***Data Integrator*:** Combines data from different sources into a unified dataset.

1. ***Decision Tree Model Module:***

This module focuses on developing, training, and evaluating the decision tree models for crop yield prediction. It includes submodules like:

***Model Trainer*:** Trains decision tree models on the prepared dataset.

***Model Evaluator*:** Measures the model's performance using appropriate metrics.

***Model Tuner*:** Fine-tunes the model's hyperparameters for better accuracy.

***Model Serializer*:** Saves and loads trained models for future use.

1. ***Chat Bot Interface Module:***

This module deals with the user interface of the chat bot, enabling users to interact with the system. It includes submodules like:

***User Input Processor*:** Handles user queries and extracts relevant information.

***NLP (Natural Language Processing)*:** Understands and interprets user messages.

***Chat Bot Logic*:** Contains the core logic for responding to user queries and providing predictions and recommendations.

***Visualizer*:** Generates graphs, charts, or visual aids for better user understanding.

1. ***Crop Recommendation Module:***

This module is responsible for suggesting suitable crops to users based on the decision tree predictions and environmental factors. It includes submodules like:

***Recommendation Algorithm*:** Implements the logic for crop recommendations.

***Customization Options*:** Allows users to customize recommendations based on their preferences.

***Crop Database*:** Stores information about various crops and their characteristics.

1. EXPERIMENTAL RESULT & ANALYSIS
2. ***Decision Tree Algorithm:***

A well-liked and adaptable machine-learning technique for both classification and regression applications is the decision tree. They work in a variety of industries, such as marketing, banking, and healthcare. The workings, construction, pruning, and applications of the decision tree algorithm will all be covered in this paper.With a single root node branching into several decision nodes, which in turn lead to leaves representing class labels or numerical values, a decision tree is a hierarchical structure that resembles an upside-down tree. The decision made by each internal node in the tree is based on the value of a characteristic, and the anticipated output is stored in each leaf node. Based on the feature that best distinguishes the data at each node, the decision tree method recursively divides the dataset into subsets. At every split, the algorithm seeks to maximise information gain (entropy reduction) or minimise impurity. Gini impurity and entropy are two popular impurity metrics. The method keeps splitting the data until a predetermined threshold is reached, either a maximum depth or a minimum quantity of samples per leaf.

***B. Random Forest Algorithm:***

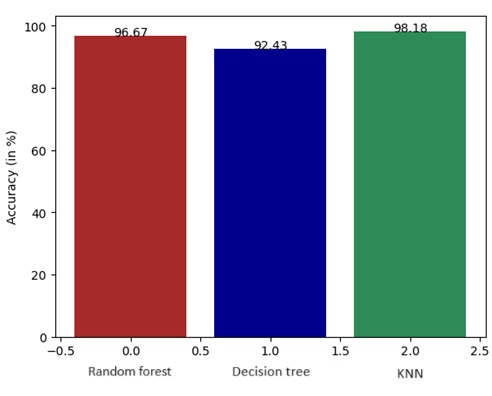
The decision tree is a popular and flexible machine-learning method that may be used for regression as well as classification. They are employed in a range of sectors, including finance, healthcare, and marketing. This paper will discuss the building, pruning, workings, and applications of the decision tree method.A decision tree is an upside-down tree-like hierarchical structure made up of multiple decision nodes branching from a single root node, which in turn leads to leaves representing class labels or numerical values. Every leaf node in the tree stores the expected output, and every interior node bases its choice on the value of a characteristic. Recursively, the decision tree technique partitions the dataset into subsets based on the feature that best separates the data at each node. The algorithm looks for impurity minimization or information gain maximisation (entropy reduction) at each split. Two widely used impurity measurements are entropy and Gini impurity. The process continues to split the data until a predefined limit is met, which could be a minimum number of samples per leaf or a maximum depth.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Area(sq.feet) | 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 |

**Table4.** Recommendation of crops based on data of the place

|  |  |  |  |
| --- | --- | --- | --- |
|  | Algorithms | Accuracy | Standard Deviation |
| 0 | Random Forest | 96.67 | 0.691015 |
| 1 | Decision-tree | 92.43 | 2.520343 |
| 2 | KNN Classifier | 98.18 | 0.668450 |

**Table5.** Comparison of results from different algorithms

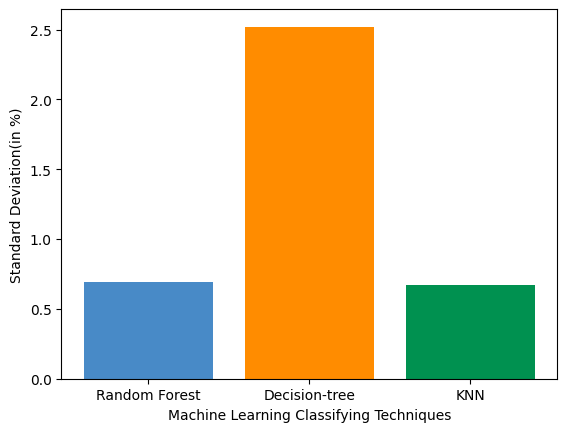
****

**Fig3.** Accuracy **c**omparison of Random forest, Decision tree and KNN algorithms

The two phases of Random Forest's operation are the creation of the random forest through the combination of N decision trees and the prediction of each tree generated in the first phase. Since the outcome is determined by a majority vote or average, the algorithm seeks to remove overfitting. This demonstrates the parallelization property, as each decision tree that is produced is independent of the others. Hence, Majority Voting or Averaging is the basis for the algorithm's final output for both classification and regression.

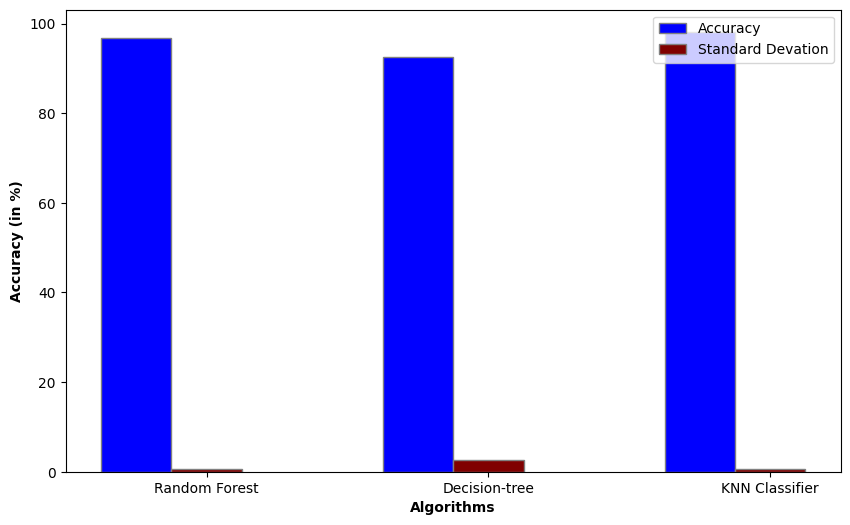
1. ***KNN (K- Nearest Neighbour):***

Classification and regression difficulties are addressed by the robust and user-friendly K-Nearest Neighbours (KNN) technique in machine learning. Using the idea of similarity, KNN uses the K nearest neighbours of a new data point in the training dataset to assume the label or value of that data point. Being non-parametric—that is, not assuming anything about the data distribution—it is highly disposable in real-world situations (in contrast to other algorithms like GMM, which assume a Gaussian distribution of the given data). The main reason for the popularity and versatility of the (K-NN) method in machine learning is its simplicity and ease of implementation.

****

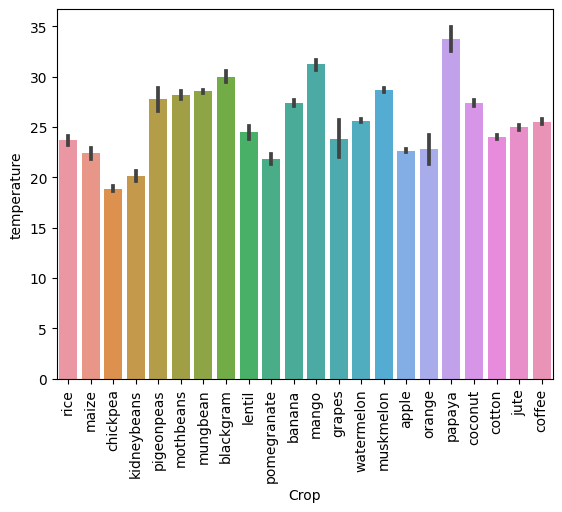
**Fig4.** Standard deviation comparison of algorithms

Pattern recognition, data mining, and intrusion detection are three major applications for this supervised learning domain member. Regarding the underlying data distribution, no assumptions are needed. Its versatility in handling different kinds of datasets for classification and regression tasks stems from its ability to handle both numerical and categorical data. Predictions are based on the degree of similarity between data points in a particular dataset using this non-parametric method. Different algorithms have different sensitivity to outliers than K-NN. Using a distance metric, like Euclidean distance, the K-NN algorithm locates the K closest neighbours to a given data point. Next, the average of the K neighbours or the majority vote determines the class or value of the data item. To make predictions, the calculated distance of the algorithm is divided by each new data point in the test dataset and every data point in the training dataset.. With this method, the algorithm may adjust to various patterns and forecast outcomes by using the data's local structure. Following the calculation of the distances between each new data point and every other data point in the training dataset, the algorithm finds the K nearest neighbours by utilising these distances. The method predicts things based on the labels or values connected to the K nearest neighbours after determining who they are. The anticipated label for a new data point in classification tasks is allocated to the majority class among K neighbours. The predicted value in regression tasks is determined by taking the mean or weighted mean of the values of the K neighbours.

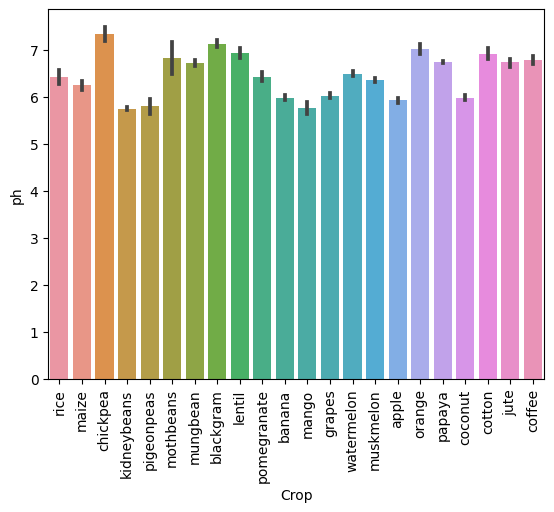
****

**Fig5.** Comparision of algorithms based on accuracy and

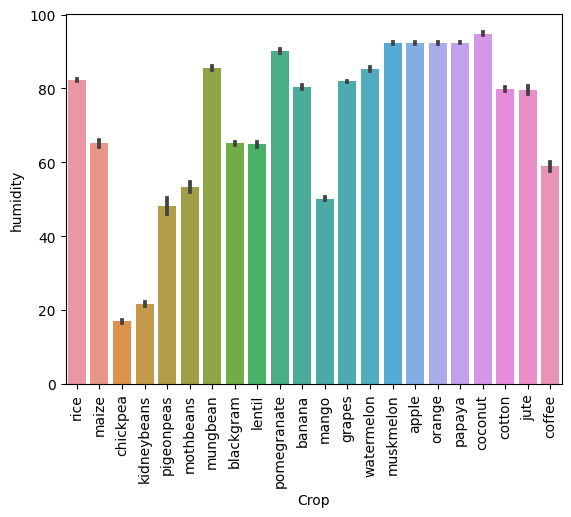
standard deviation

****

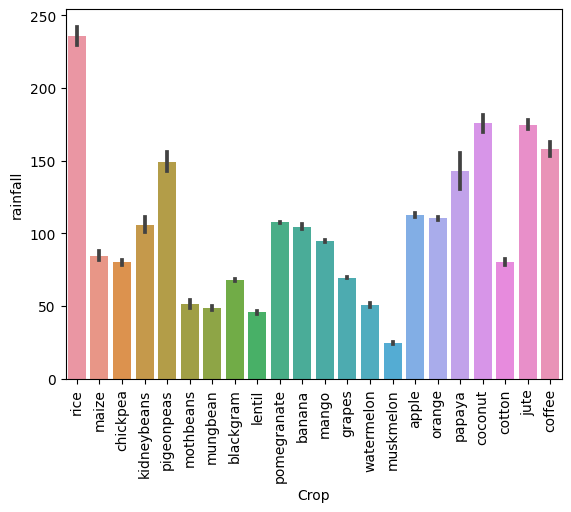
**Fig6.**  Crop dataset visualisation on temperature

****

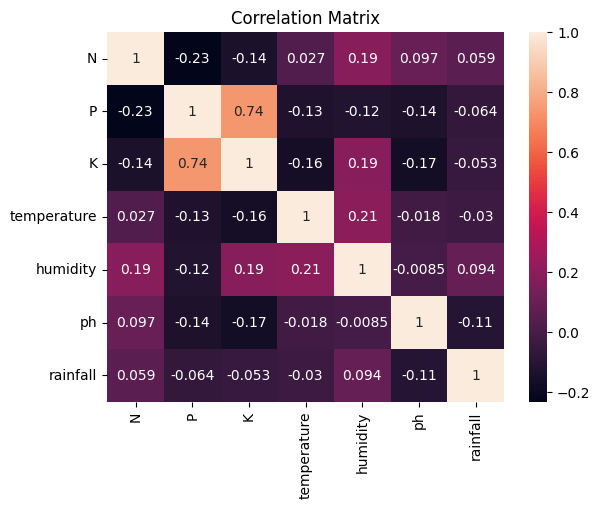
**Fig7.** Crop dataset visualisation on ph

****

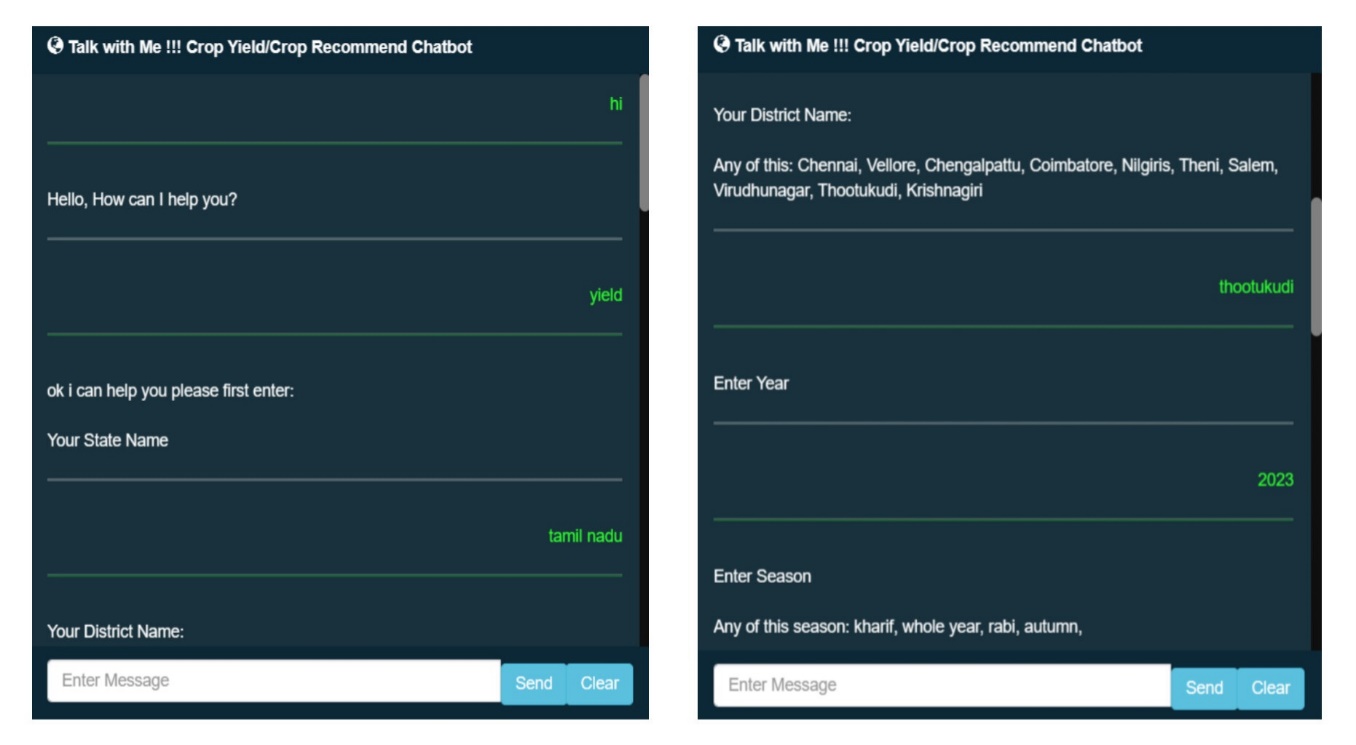
**Fig8.** Crop dataset visualisation on humidity

****

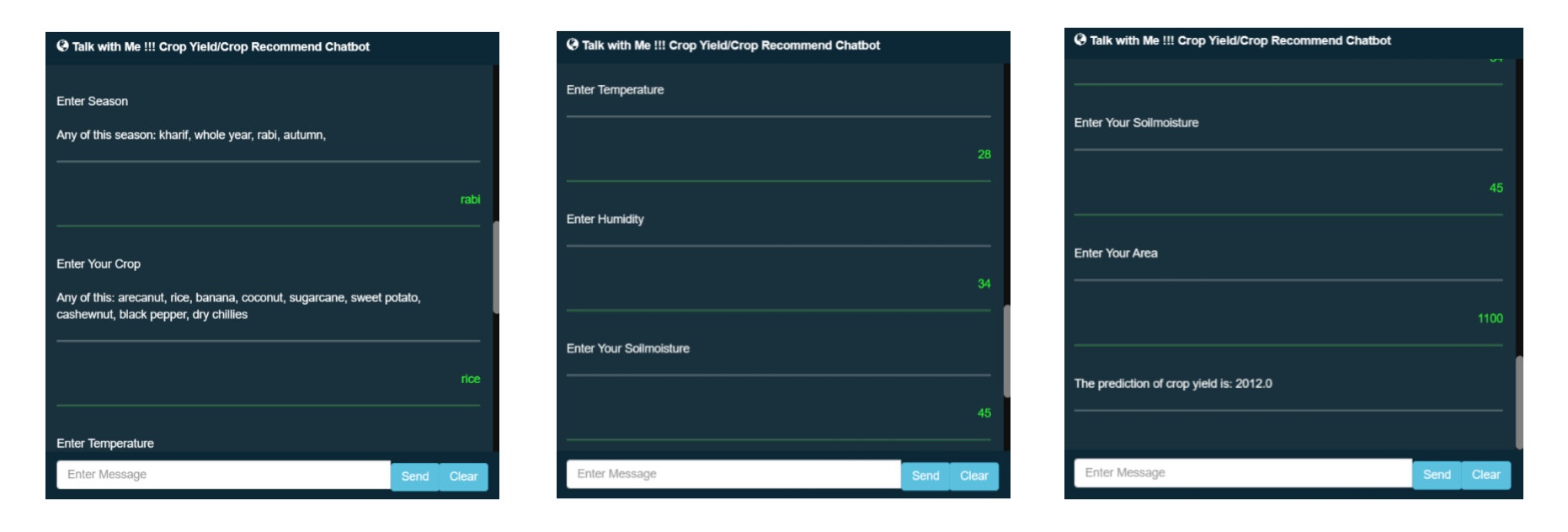
**Fig9.** Crop dataset visualisation on rainfall

**** **Fig10.** Correlation matrix based different crop’s dataset

1. ***Results Discussion:***
2. ***Crop yield predict:***

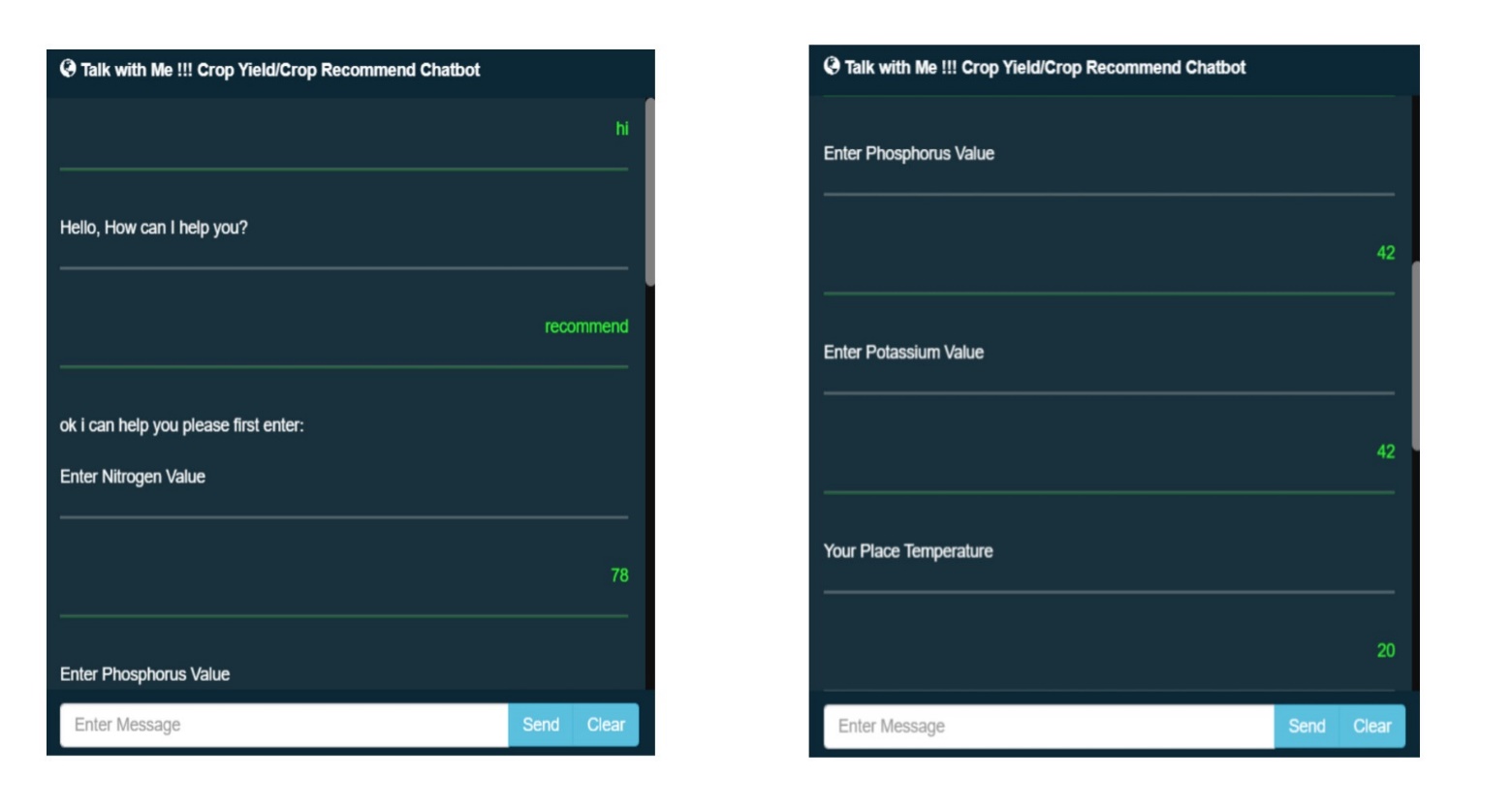
****

**Fig10.** Collection of location details

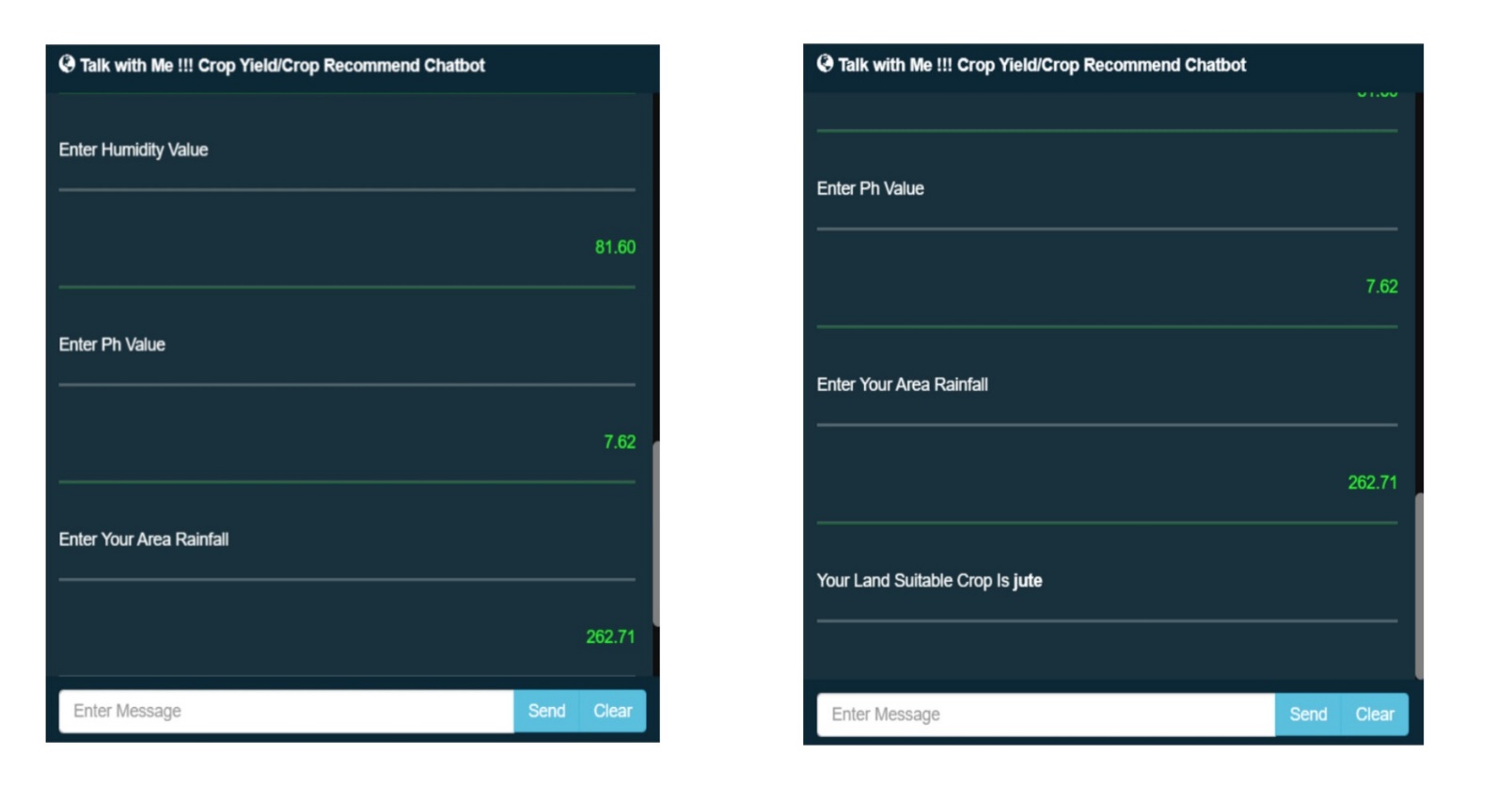


**Fig11.** Collection of Climate and crop details

1. ***Crop Recommend:***



**Fig12.** Collection of Chemical details present in the soil



**Fig13.** Collection of Climate details

1. CONCLUSION

The creation of the "Chat Bot for Crop Yield Prediction and Crop Recommendation" project is a noteworthy step forward in the use of technology to help farmers choose crops and manage their crops. This chatbot can provide farmers the information they need to grow and preserve crops. Farmers can identify the factors that most significantly affect crop production changes and make educated decisions by using Decision Trees, which are easy to read. The initiative seeks to offer significant assistance to the farming community by means of a methodical approach and a well-organized modular model. This technology lowers the likelihood of crop failure and boosts output by giving the farmer information that other farmers overlook, assisting them in selecting the best crop. It also keeps them from suffering losses. Millions of farmers nationwide will be able to utilise the system if it is made available online.

1. REFERENCE

[1] Lai J.Y., Sowmya A., Trinder J., (2005). Support vector machine experiments for road recognition in high resolution images, Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005.

[2] Batts G.R., (1997). Effects of co2 and temperature on growth and yield of crops of winter wheat over four seasons, journal ISSN:1161-0301.

[3] Rub G., (2009). Data mining of agricultural yield data: a comparison of regression models, Advances in Data Mining. Applications and Theoretical Aspects, 9th Industrial Conference, ICDM 2009, Leipzig, Germany, July 20-22, 2009.

[4] Jorquera, H., JR Pérez-Correa, A. Cipriano, and G. Acuña, (2001). Short Term Forecasting of Air Pollution Episodes, in Environmental Modeling (vol. 4), P. Zannetti (ed.), WIT Press.

[5] Leemans V., (2004). a real time grading method of apples based on features extracted from defects, Journal of Food Engineering 61(1).

[6] H.S. Jayanna, S.R.M Prasanna, (2009). analysis, feature extraction, modeling and testing techniques for speaker recognition, IETE Technical Review 26(3).

[7] Alberto González-Sanchez, (2014). predictive ability of machine learning methods for massive crop yield prediction, Spanish Journal of Agricultural Research 2014 12(2): 313-328.

[8] Mandic D. P., (2001). recurrent neural networks for prediction: learning algorithms, architectures and stability, ISBN: 978-0-471-49517-8.

[9] Hochreiter S., (1997). long short-term memory. neural computation, Neural Computation 9(8):1735-80.

[10] H. Sak, F. Beaufays, (2014). long short-term memory recurrent neural network architectures for large scale acoustic modelling, arXiv:1402.1128 [cs.NE].

[11] Wiener M., Andy Liaw, (2002). classification and regression by randomforest, R News, 2 (3): 18--22 (2002).

[12] Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020). Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 736-741). IEEE.

[13] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

[14] Chandgude, A., Harpale, N., Jadhav, D., Pawar, P., & Patil, S. M. (2018). A Review on Machine Learning Algorithm Used For Crop Monitoring System in Agriculture. International Research Journal of Engineering and Technology (IRJET), 5(04), 1470.

[15] Patil, A., Kokate, S., Patil, P., Panpatil, V., & Sapkal, R. (2020). Crop Prediction using Machine Learning Algorithms. International Journal of Advancements in Engineering & Technology, 1(1), 1-8.

[16] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in plant science, 10, 621.

[17] Batts G. R. (2021). Support vector machine experiments for road recognition in high resolution images.