

**CROP RECOMMENDATION SYSTEM USING WSN AND  
ML ALGORITHMS**

Project report submitted in partial fulfillment of the requirement for  
the degree of Bachelor of Technology

in

**Computer Science and Engineering**

By

Harshul Choudhary (191362)

Ujjwal Rajput (191366)

Under the supervision of

Dr. Alok Kumar

and

Dr. Pankaj Dhiman

to



Department of Computer Science & Engineering and Information  
Technology

**Jaypee University of Information Technology Waknaghat,  
Solan-173234, Himachal Pradesh**

## DECLARATION

We hereby declare that the work presented in this report entitled “ **Crop Recommendation System Using WSN and ML algorithms**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of our own work carried out over a period from January 2023 to May 2023 under the supervision of **Dr. Alok Kumar** (Assistant Professor (SG) department of ECE) and **Dr. Pankaj Dhiman** (Assistant Professor (SG) department of CSE). We also authenticate that we have carried out the above-mentioned project work under the proficiency stream Data Science.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Harshul Choudhary, 191362

Ujjwal Rajput, 191366

This is to certify that the above statement made by the candidates is true to the best of our knowledge.

Dr. Alok Kumar  
Assistant Professor (SG)  
ECE  
Dated:

Dr. Pankaj Dhiman  
Assistant Professor (SG)  
CSE  
Dated

# PLAGIARISM CERTIFICATE

## JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT PLAGIARISM VERIFICATION REPORT

Date: .....

Type of Document (Tick):  PhD Thesis  M.Tech Dissertation/ Report  B.Tech Project Report  Paper

Name: \_\_\_\_\_ Department: \_\_\_\_\_ Enrolment No \_\_\_\_\_

Contact No. \_\_\_\_\_ E-mail. \_\_\_\_\_

Name of the Supervisor: \_\_\_\_\_

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): \_\_\_\_\_

### UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/ revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

#### Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

### FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at .....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor)

Signature of HOD

### FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	<ul style="list-style-type: none"> <li>• All Preliminary Pages</li> <li>• Bibliography/Images/Quotes</li> <li>• 14 Words String</li> </ul>		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by  
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at [plagcheck.juit@gmail.com](mailto:plagcheck.juit@gmail.com)

## ACKNOWLEDGEMENT

Firstly, we express our heartiest thanks and gratefulness to almighty God for his divine blessing that made it possible to complete the project work successfully.

We are really grateful and wish our profound indebtedness to Supervisor **Dr. Alok Kumar Assistant Professor (SG)**, Department of ECE Jaypee University of Information Technology, Waknaghat and **Dr. Pankaj Dhiman Assistant Professor (SG)**, Department of CSE Jaypee University of Information Technology, Waknaghat. Deep Knowledge & keen interest of our supervisors in the field of “**Data Science**” to carry out this project. Their endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Alok Kumar**, Department of ECE, and **Dr. Pankaj Dhiman**, Department of CSE, for their kind help to finish our project.

We would also generously welcome each one of those individuals who have helped us straightforwardly or in a roundabout way in making this project a win. In this unique situation, we might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated our undertaking.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

**Harshul Choudhary**  
**(191362)**

**Ujjwal Rajput**  
**(191366)**

## TABLE OF CONTENT

List of Abbreviations.....	v
List of Figures.....	vi
List of Graphs.....	vii
Abstract.....	viii
Introduction.....	1-21
Literature Survey.....	22-29
System Design & Development.....	30-37
Experiments & Result Analysis.....	38-43
Conclusions.....	44-47
References.....	48-49

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Definition</b>
WSN	Wireless Sensor Network
ML	Machine Learning
GUI	Graphics User Interface
API	Application Programming Interface
AI	Artificial Intelligence
DT	Decision Tree
SVM	Support Vector Machine
RF	Random Forest
NB	Naive Bayes
GPU	Graphics Processing Unit
MLR	Multiple Linear Regression
SDLC	Software Development Life Cycle
SRS	Software Requirement Specifications
ASWT	As we know that
Approx	Approximately

## LIST OF FIGURES

Figure No.	Title	Page No.
1.1	Methodology	3
1.2	Design of ML algorithm	4
1.3	WSN nodes deployment in agricultural land	9
1.4	Building ML model	11
1.5	Design of Tkinter Window	13
1.6	Block diagram of overall methodology of proposed System	17
3.1	Waterfall Model	31
3.2	Sigmoid function for logistic regression	33
3.3	SVM hyperparameters	35
4.1	Dataset	38
4.2	Description of features	39
4.9	Crop Recommendation Assistant	44

## LIST OF GRAPHS

Figure No.	Title	Page No.
4.3	Confusion Matrix	39
4.4	Nitrogen value comparing different Crops	40
4.5	Potassium value with respect to different Crops	40
4.6	Accuracy Comparison of applied models	42
4.7	Accuracy Comparison of Ensemble Models (2 of 5)	43
4.8	Accuracy Comparison of Ensemble Models (3 of 5)	43



## **ABSTRACT**

In this project, we have developed a GUI that will assist farmers in selecting the most suitable crop for their land. Agriculture is the largest source of livelihood in India and approx. 70 percent of its rural households still depend primarily on agriculture for their livelihood.

However, India still has many growing concerns, as the Indian economy has diversified and grown. Looking at the current situation faced by the farmers in India, we have observed that there have been many suicides in India over many years, the main reason behind this is the change in weather conditions and frequent changes in the Indian Government system. Sometimes farmers are not aware of the crop which suits their soil quality, soil nutrients, and soil composition. This project aims to help farmers to check the soil quality to get good crop yield. Any farmer is interested in knowing how much yield he is about to expect. The prediction using various ML models will help the farmers predict the crop yield before cultivating it on the agricultural field. ML is an essential approach for achieving the practical and essential solution to this problem. This system considers various parameters like soil moisture, soil pH value, rainfall, and temperature all at once.

Based on all these parameters the system will predict the best crop for the farmer using the ML approach.

Keywords- Agriculture, ML, GU

# CHAPTER-1 INTRODUCTION

## 1.1 Introduction

Agriculture is considered as the most important cultural practice in India. Crop recommendations are new-generation bubbles that dominate the masses most of the time. Farmers are unaware of the types of crops they should grow on their farms. This causes a lot of confusion and it also affects productivity. Therefore, in this project, we propose an idea to predict the optimal harvest of the crops. Increased yields can be achieved by analysing all of these problems including weather, temperature, and other factors.

Compiled datasets containing precipitation, climate and fertilizer data available for India will allow us to better understand how crops evolved given different environmental and geographic factors. We have used this dataset to build a ML model to predict the best crop that should be grown on a particular land.

Machine learning could be a game changer in agri-business. What is needed now is to develop systems that provide farmers with predictive insights and help them make informed decisions about which crops to grow.

Thus, we had identified the farmer's dilemma as to which crop needs to be sown in a particular season. With this in mind, we have proposed an intelligent system that, before advising the user on the best crops, takes into account soil qualities (nutrient content, soil type, pH) as well as environmental factors (rainfall and temperature).

## 1.2 Problem Statement

Achieving high yields with comprehensive technical solutions requires soil quality analysis and advanced crop forecasting. This project's primary objective is to forecast the ideal crop using multiple ML techniques. This is very beneficial for farmers when planning their harvest and selling their crops. An ML model implementation provides the best crop forecast for a given region and season in a given region.

### **1.3 Objectives**

The objective of crop recommendation system is to give farmers precise advice on the best crop to grow based on a variety of factors, including weather patterns, soil composition, location, and crop diseases. Farmers with little technical experience should be able to use the system with ease. AI in agriculture has the potential to increase crop productivity and yield while decreasing the use of resources like water and fertilizers. A crop recommendation system's overall goal is to assist farmers in meeting the rising demand for food while promoting sustainable farming methods.

Using geographic and climatic boundaries, a technology-based crop recommendation system for agriculture helps the farmers increase crop production by recommending a suitable crop for their area. It is believed that the suggested crop recommendation model is successful and useful in recommending a reasonable crop.

The primary goal of this project is:

- To build a strong model to give correct and precise predictions of crop in a given state for the specific soil type and climatic circumstances.
- Give suggestions of the best appropriate crop yields in the area so the farmer does not face any losses.

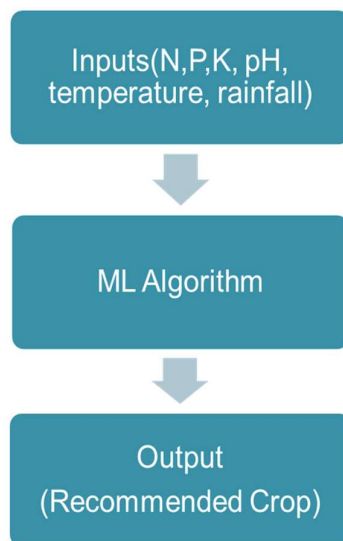
### **1.4 Methodology**

The main ideology of this research includes the concept of using various ML models to recommend the best crops to grow on a given land. The results are therefore very beneficial for farmers.

This project used a dataset containing 2200 values from 22 unique plants. We applied six different ML algorithms to this dataset. In addition, we trained the model according to the actual values, tested the model and calculated the accuracy. We also created a GUI [1] to make it easier for farmers to use the crop recommendation system. This is an attractive GUI [1] that can be used seamlessly even by first-time users. This GUI [1] was built using the Python Tkinter [2] library. The model is not only intended to appeal to uneducated farmers, but also to help anyone who wants to use a crop recommendation

system. So, in order to make this GUI [1] easier to use, we also added a language system.

This interaction creates the basis for additional assessments through the potential for auxiliary variables to influence the selection of yields developed in specific fields at specific locations.



**Fig.1.1 Methodology**

The functionality of the architecture (Fig.1.1) is as follows:

Our system accepts various inputs from users. Soil pH based on nutrient content and varying weather conditions. Soil pH, ambient temperature, precipitation, moisture, and various soil contaminants (N, P, K) are recorded using a third-party application. Additionally, these results are displayed in the plant dataset. We apply different ML models, calculate the accuracy of different models, and based on that accuracy; the system recommends a particular crop as a result.

#### 1.4.1 Research Approach

We have used Scikit-learn [5] to perform ML computations on the dataset. A model is prepared to take the information that a user gives it and apply similar patterns & information to produce the ideal result for the given input. Then run the model after testing, and finally using the GUI [1] to provide a visual look and clarity to arrive at the prediction results.

### 1.4.2 Method Approach

ML models are built up using the datasets. This is followed by the division of the data into preparation and test sets, which are then prepared and tested using three different calculations. The best calculation is chosen and the model is determined. The model does its job at this point by using various factors as information and returning the ideal yield.



**Fig.1.2 Design of ML algorithm**

### 1.4.3 Python Libraries

#### 1) Pandas

Pandas [3] is an open source library that is BSD (Berkeley Software Distribution) licensed. The area of data science makes extensive use of this well-known library. They are primarily employed for data editing, cleaning, and analysis. Data modelling and data analysis are simplified by Pandas without the need to convert to another language like R.

Pandas [3] can perform a variety of tasks, including: Example:

1. Dataframes can be sliced with pandas.
2. Combining and merging data frames can be done with pandas.
3. Concatenate the columns of two data frames.
4. To change the index value of the data frame.
5. You can change column headings in pandas NumPy.

## 2) NumPy

One of the most widely used open source Python libraries focused on logical processing is NumPy [4]. It has implicit number related capabilities for quick calculation and supports enormous networks and multi-faceted information. "Mathematical Python" is characterized by the expression "NumPy". It frequently serves as a sophisticated information storage unit for general knowledge, an irregular number generator, a straight polynomial math tool, etc. A NumPy Cluster is a python object that characterizes a N-layered exhibit with lines and segments. Python inclines toward NumPy [4] exhibits over records. Since it utilizes less memory, is quicker, and is more straightforward to utilize.

Utilizing the NumPy interface, pictures, sound waves, and other crude twofold information streams can be addressed as N-layered varieties of genuine qualities for representation. Full-stack designers require an immense measure of information to execute this ML library.

## 3) Scikit Learn

An open-source AI computation library that utilizes Python is called Scikit Learn [5]. It might very well be applied to computations for both directed and independent learning. This library includes the NumPy, Matplotlib, and SciPy bundles in addition to standard computations. Scikit discovers the most popular ways to use musical concepts on Spotify.

Scikit Learn [5] contains an enormous number of calculations that can be utilized to perform normal AI and information mining undertakings like dimensionality decrease, order, relapse, grouping and model choice.

## 4) SciPy

Scipy [6] is a Python library that is open-source, free, and used for logical reasoning, data management, and superior performance registers. Simple to-involve schedules are available in the library for quick estimations. This bundle supports information management and perception as well as significant level orders and depends on the NumPy

expansion. Scipy [6] and NumPy [4] are used together to do numerical calculations. While SciPy maintains numerical codes, NumPy allows you to sort and list exhibit information. Only a few of the various subpackages available in SciPy include Bunch, Constants, fftpack, Coordinate, Add, io, Linalg, Ndimimage, Odr, Upgrade, Signal, Scanty, Spatial, Exceptional, and Details. Using "from scipy import subpackage-name," you can import data from SciPy. Nevertheless, SciPy requires the bundles SciPy Library, Matplotlib, NumPy, IPython, Sympy, and Pandas.

#### 5) PyTorch

Facebook just debuted PyTorch [7], a Python library, with two noteworthy level features:

1. Tensor calculation with huge GPU speed increase (like NumPy)
2. Stage in light of profound brain networks that gives adaptability and speed.

Overall, PyTorch [7] is a powerful and flexible library for building and training deep learning models in Python. Its dynamic computation graph and strong debugging tools make it particularly well-suited for research and experimentation in deep learning.

#### 6) Matplotlib

A plotting library used in the Python programming language is called matplotlib.pyplot. It may very easily be used in shells, web application servers, python scripts, and other graphical user interface toolkits.

It is a charting Python package that offers protested situated APIs so that plots can be integrated into programmes.

There are a few toolboxes that are available that increase the usability of Python Matplotlib [8] ; it is not a component of the Standard Libraries that is introduced by default when Python. Some of them can be downloaded separately, while others can be sent along with the matplotlib source code but have additional requirements.

Matplotlib [8] provides a range of functions for creating different types of plots, including line plots, scatter plots, bar plots, and histograms. It also includes tools for adding titles, labels, and annotations to plots, as well as adjusting the color, style, and other properties of the plot elements.

Overall, Matplotlib [8] is a powerful and flexible library for creating visualizations of data in Python, and is widely regarded as one of the most important tools in the data science toolkit.

#### 7) Seaborn

Python's Seaborn [9] package allows users to create realistic designs. It builds upon Matplotlib [8] and closely integrates with Pandas [3] data structures.

It aids in the research and knowledge retention. In order to produce illuminating charts, its plotting skills operate on dataframes and exhibitions comprising whole datasets. Inside, crucial semantic planning and quantifiable collecting are performed. With its dataset-organized, informative programming interface, you can concentrate on the meaning of each element of your plots rather than the specifics of how to design them.

#### 1.4.4 Kaggle

Kaggle [10] is a popular platform for data science competitions and projects, where users can compete with each other to solve real-world problems using data. Kaggle [10] provides a range of datasets and challenges for users to work on, as well as a community of data scientists and machine learning practitioners who can collaborate and share knowledge. Users can also create and share their own datasets and projects on the platform.

Kaggle [10] is widely used by data scientists and machine learning practitioners for developing their skills, building their portfolios, and collaborating with other professionals in the field. It is also used by companies and organizations to source solutions to real-world problems and to identify top talent in the data science community.



#### 1.4.5 Wireless Sensor Network

In order to monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion, or pollutants, Wireless Sensor Networks (WSNs) are self-configured and infrastructure-free wireless networks. WSNs [11] cooperatively pass their data through the network to a central location or sink where the data can be viewed and analyzed. An interface between users and the network is provided by a sink or base station. By injecting queries and gathering results from the sink, one can receive necessary data from the network. A wireless sensor network [11] typically consists of millions of sensor nodes. Radio signals can be used by the sensor nodes to communicate among themselves. A wireless sensor node is outfitted with power supplies, radio transceivers, and sensing and processing equipment. A wireless sensor network (WSN individual)'s nodes are fundamentally resource constrained because of their confined processing power, storage space, and communication bandwidth. After being deployed, the sensor nodes are in charge of self-organizing a suitable network infrastructure and frequently communicating with them across several hops.

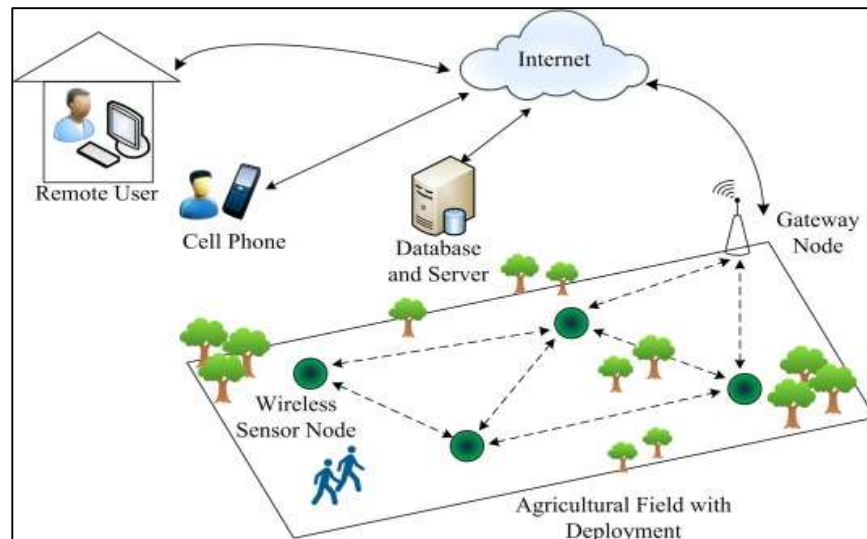
##### 1) WSN in Machine Learning:

One of the most well-known applications of artificial intelligence is machine learning (ML), in which computer algorithms create a mathematical model from a sample of data called "training data" in order to make predictions or judgments without being explicitly taught to do so. The listed justifications make the ML nature of WSNs [11] appropriate: Due to the complexity of WSN ecosystems, mathematical frameworks cannot be created. Additionally, some programmes employ data sets that need to be integrated in order for them to work effectively. Finally, in keeping with the nature of WSNs, ML algorithms do not require human interaction. WSNs also have unanticipated dynamics and behaviors. The resources and computing capabilities of nodes, as well as the requirement for sizable data sets for learning, provide the two key obstacles to ML in WSNs. The difficulty of applying ML algorithms to the integrity and confidentiality of security requirements is one of the most significant issues faced by ML algorithms with regard to the security of WSN networks. Therefore, machine

learning methods can aid in enhancing wireless network security, decreasing all types of congestion issues , assisting physical layer authentication processes , and assisting mistake detection . Additionally, ML algorithms have a significant advantage in packet analysis between WSN nodes and spotting suspect nodes.

## 2) WSN in Agriculture:

WSNs, or wireless sensor networks, are frequently used in agriculture to track and enhance the efficiency and effectiveness of farming. WSN [11] can be used to track real-time data on temperature, water supply, irrigation systems, and other agri-parameters. Farmers can produce crops in greater quantities and at lower cost by using the WSN [11] data. WSN may also be used for precision agriculture management, a technique for management that maximizes crop yields while minimizing waste. Most works on irrigation management are included in the most recent state-of-the-art. Crop field monitoring, water quality monitoring, and soil quality parameter monitoring are all examples of monitoring. Crop field checking is essential in horticulture to decrease asset squandering and increment yield in exercises like water system preparation. The data produced by the checking can be utilized to foresee crop well-being and creation quality. An ordinary remote sensor network for estimating ecological variables is displayed in Figure below.



**Fig.1.3 WSN nodes deployment in agricultural land [11]**

#### 1.4.6 Machine Learning

ML is a cutting-edge innovation that has been tested on a variety of master cycles, present-day cycles, and our regular schedules. It is a subset of man-made intellectual prowess (recreated insight), which is based on developing cunning PC systems employing verifiable ways to obtain information from open informational indexes. PC structures can use all of the client data with ML. It takes care of what has been modified while also adjusting to new circumstances or adjustments. Data-driven estimations cause approaches to action that weren't adjusted earlier to shift.

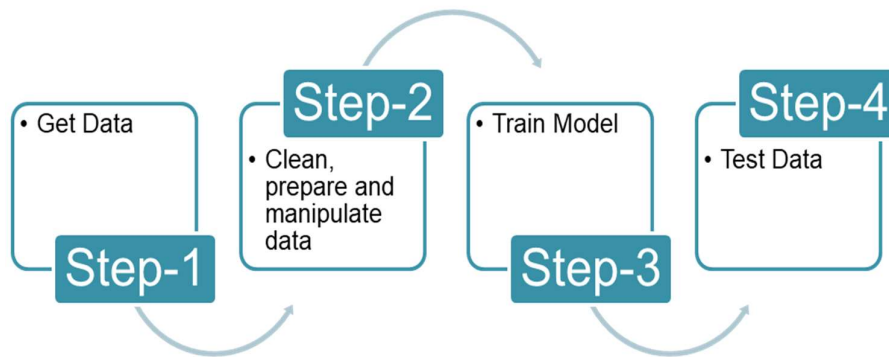
Finding a method to examine and observe the environment suggests that an electronic companion may actually look at communications and concentrate on the important information. Making calculations about the acting styles of prospective clients comes naturally with this learning. This aids in your ability to be responsive but also proactive and to truly understand your clients.

AI is a segment of simulated intelligence. For the most part, it's a fake cerebrum network with something like three layers. Mind networks with only a solitary layer can make surveyed assumptions. The development of extra layers can assist with growing improvement and accuracy.

ML is appropriate in many fields, and endeavors, and can foster long term. Coming up next are six authentic occurrences of how computer based intelligence is being used.

The ML system constructs assumption models from verifiable data and predicts the results for new data whenever it receives it. The amount of data is a factor in projected yield accuracy because a large amount of data makes it easier to develop a model that accurately predicts the outcome.

Assuming that we have a complex problem that requires the use of specific assumptions, rather than creating a code for it, we should simply deal with the data using conventional estimations. With the aid of these calculations, the machine develops the justification suggested by the data and predicts the outcome. ML has altered our perspective on the problem.



**Fig.1.4 Building ML Model**

Features of Machine Learning:

- ML makes use of data to identify various patterns in a given dataset.
- It can benefit from prior knowledge and subsequently improve.
- It is an invention that is information-driven.
- Due to the fact that it also deals with the enormous amount of data, ML is similar to data mining.

The condition for ML is incremental expansion. The ability of ML to perform tasks that would be unreasonably difficult for a person to carry out directly is the justification for the requirement for ML. We have some restrictions as humans because we can't actually access a large fraction of the data. To overcome these limitations, we need a few PC systems, and this is where machine learning (ML) comes in to simplify things for us.

We can schedule ML calculations by supplying them with a massive amount of data and allowing them to explore the data, develop the models, and anticipate the significant result normally. The amount of data determines how the ML calculation is introduced, and it tends not to be set up permanently by the cost capacity. We can save time and money with the aid of ML.

The motivations of ML make its significance clear to all. ML is being utilised in self-driving cars, sophisticated distortion, recognisable proof, facial affirmation, and Facebook's buddy concept, among other applications. Different leading companies, including Netflix and Amazon, have created machine

learning (ML) algorithms that use a significant amount of data to dissect customer interest and suggest further items.

Classification Of Machine Learning:

- Reinforcement learning,
- Supervised learning, and
- Unsupervised learning

#### 1. Supervised Learning:

Supervised learning is a type of machine learning technique where we prepare the ML framework with labelled data and then rely on it to predict the outcome. The framework creates a model using named information to comprehend the datasets and learn about every piece of information; after preparation and handling are complete, we test the model by providing an example piece of information to see if it is capable of foretelling a particular outcome.

Planning input information with result information is the goal of supervised learning. A pupil studying under an educator's supervision is equivalent to supervised learning, which depends on oversight. Spam filtering is an example of supervised learning in action.

Two categories can be made out of it:

- Regression
- Classification

#### 2. Unsupervised Learning:

Unsupervised learning, which uses unlabelled input features, is a learning approach in which a computer learns with very little supervision.

The computer is instructed to prepare the arrangement of unmarked, ungrouped, unsorted data, and the calculation is required to follow up on such data without supervision. Rebuilding the material into new highlights or a collection of items with related examples is the goal of unsupervised learning.

With unsupervised learning, there is no predetermined result. The algorithm searches through a vast amount of material looking for useful experiences.

It is classified into 2 categories:

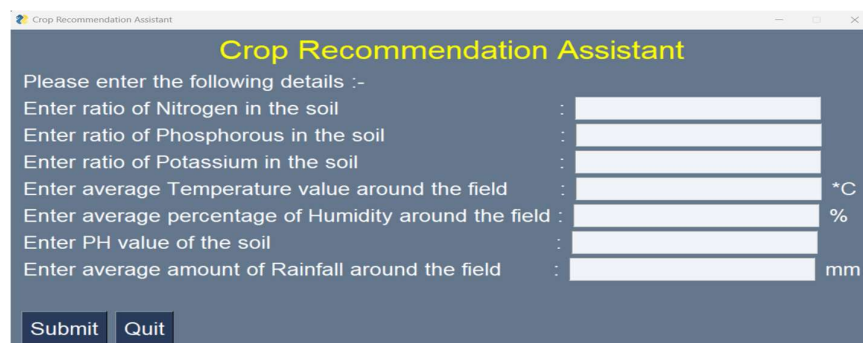
- Clustering,
- Association

### 3. Reinforcement Learning:

A learning specialist uses the feedback-based learning strategy known as reinforcement learning, where they are rewarded for doing something properly and penalized for doing something poorly. With these contributions, the specialist develops appropriately and improves performance. In reinforcement learning, the expert investigates and engages with the environment. A specialist's goal is to get as many awards as possible, so they improve their performance.

#### 1.4.5 Tkinter

The default GUI [1] library for Python is called Tkinter [2]. When Python and Tkinter [2] are used together, a quick and easy method for creating GUI [1] applications is provided. Tkinter [2] provides strong points for an appropriate connection point to the Tk GUI [1] toolbox. Tkinter [2] offers a variety of controls, such as buttons, marks, and text restraints that are used by a GUI [1] programme. Typically, these are referred to as gadgets. We have used Tkinter [2] for fostering a graphical UI to consolidate with the ML model to make its UI clearer to use the system to come by the necessary outcomes. Tkinter [2] is a standard library of python used for making GUI [1] by combining with python in a speedy and straightforward way to give an arranged interface.



**Fig.1.5 Design of Tkinter window**

#### 1.4.6 Agriculture Season in India

The farming harvest season in India is from July to June. The Indian farming season is ordered into three fundamental times of development:

##### Rabi:

Crops for rabi are harvested in the summer from April to June after being planted in the winter from October to December. The important rabi crops include a percentage of wheat, grain, peas, gramme, and mustard. Although these yields cover a considerable portion of India, the growth of wheat and other rabi crops is particularly important in the north and north western states of Punjab, Haryana, Himachal Pradesh, Jammu & Kashmir, Uttarakhand, and Uttar Pradesh. The availability of precipitation over an extended period of cold weather due to the western quiet tornadoes aids in the success of these yields. However, the growth of the previously mentioned initiatives has also taken into substantial account the progress of the green revolution in Punjab, Haryana, western Uttar Pradesh, and areas of Rajasthan.

##### Kharif:

In different parts of the country, kharif crops start to grow as soon as the rain starts, and they are harvested between September and October. Paddy, maize, jowar, bajra, tur (arhar), moong, urad, cotton, jute, peanuts, and soybean all achieved significant yields during this season. Assam, West Bengal, Odisha, Andhra Pradesh, Telangana, Tamil Nadu, Kerala, and Maharashtra—particularly the (Konkan coast) alongside Uttar Pradesh and Bihar—are likely the key rice-developing areas. Paddy has recently become a key crop for Punjab and Haryana as well. Three yields of paddy are produced annually in states like Assam, West Bengal, and Odisha. Aus, Aman, and Boro are these

Zaid:

The Zaid season is a brief period of time that spans the middle of the year between the rabi and kharif seasons. Watermelon, muskmelon, and cucumber are among the produce that is presented during "zaid." In a short period of time, crops are cultivated essentially under a fake water system during the mid-year editing season. Crops are sown in the spring (Feb-Walk) and harvested in the summer (April-June).

#### 1.4.7 Crop Classification

Various food and non-food crops are grown in different regions of the country based on variations in the soil, environment, and development practices. These yields can be characterized based on various measures identified underneath:

Crops classified according to utility:

Utility crops are further classified into six groups based on their use:-

Seed:

All crop plants produce seeds as their main byproduct, and these seeds are used to make food. Due to their unpleasant qualities, such as those of apple or cherry seeds, not all plant seeds are used as food. Grain, vegetable, and nut ingredients are included in tasty seeds. These are abundant in common food fats.

Pulses: These plants produce foods such as peas, lentils, gramme (chana), and green and black gramme (masur). They provide the protein in common foods.

Vegetables:

The vegetables that come to mind for these yields are mostly spinach, bamboo shoots, carrots, verdant vegetables, microgreens, etc. A plentiful source of vitamins, minerals, and fibre is found in vegetables.

Oilseed Crops:

Oil is extracted from oil-producing plants by crushing the seeds. The oil is afterwards used for frying or other cooking purposes. For instance, sesame (till),



rapeseed, sunflower, flaxseed, soybean, peanut, mustard (sarson), and so forth. Oils are abundant in crops with seeds.

#### Fiber Crops:

These plants have been specifically bred to produce strands for fillers, ropes, and other materials. Stringy harvests include flax, cotton, kenaf, jute, sun hemp, modern hemp, and sun hemp.

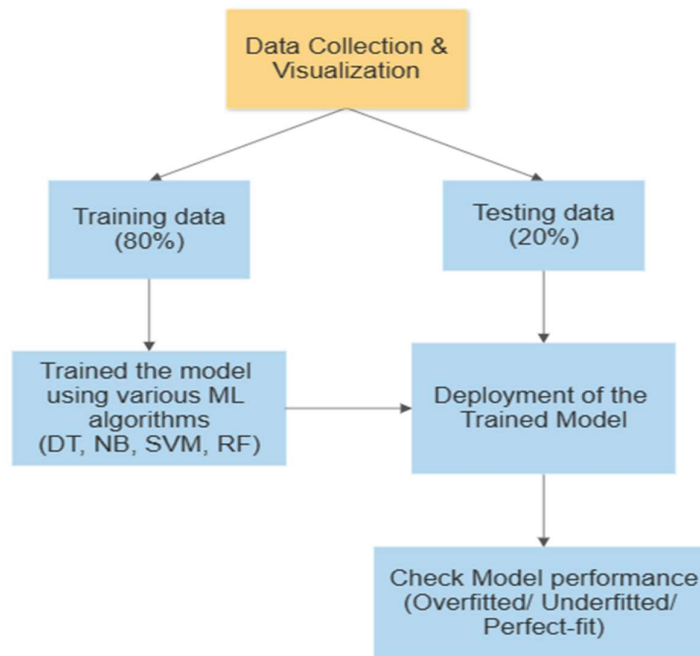
#### Cash Crops:

Additionally, these plants are referred to as cash crops. These yields are provided for further benefit and usage on a contemporary level. For instance, opium and tobacco (alkaloids/smoking), cotton and jute (fiber), sugarcane and sugarbeat (sugar), espresso and tea (drinks), and cotton and jute (fiber). Ranch crops include coconut, elastomer, coffee, and tea.

#### Feed Crops:

Feed crops are one type of harvest that are primarily grown for animal feed. No matter if they are developed, typical fields and meadows are included. The scavenges are dubbed annual or eternal harvests and are grown and harvested like any other yield. Rummage crops are specially created for animals that come in contact with them or are stored as silage or roughage. Feed crops aid in achieving production goals for growth or weight gain and help to alleviate sporadic supply and demand bottlenecks for feed. For animals, these plants provide green fodder. For instance, oats, sorghum, berseem, and so forth.

**Proposed system:**



**Fig. 1.6 Block diagram of Overall methodology of proposed system**

In this structure, we have suggested a process that is broken down into various stages:

- 1) Data Collection and,
- 2) Data Visualization
- 3) ML Algorithms
- 4) Crop Recommendation System
- 5) Recommended Crop

a) Data Collection

The most well-known method for social events and analyzing data from many sources is information assortment. For the dataset to provide an approximate informational index to the framework, it should contain the corresponding qualities. These criteria will be taken into account for crop suggestions:

- 1) Soil pH
- 2) Humidity
- 3) NPK Levels

4) Information on the crop

5) Temperature

#### b) Data Pre-processing

The next step is to pre-handle the information after collecting it from a variety of sources so that the model may be created. Information pre-handling can be conducted in a variety of methods, starting with browsing the obtained dataset and moving via information cleansing. During data cleansing, some dataset attributes are overly numerous. are not taken into consideration during cutting. Therefore, we should get rid of annoying attributes and datasets that have some missing data. To obtain them, we should eliminate or replace these lacking qualities with undesirable nan values.

#### c) ML algorithms

We have trained and tested various ML algorithms because our crop recommendation system comes under classification algorithms. Therefore, to predict the most suitable crop and optimum accuracy, we compared the accuracy of all the ML models and selected the one which is having the highest accuracy. The algorithms which we have applied in our project are Logistic Regression, DT, KNN, Naive Bayes, and SVM.

##### i) Logistic Regression

It is among the easiest supervised learning classification problems that predicts the probability of the target variable, where the dependent variable is either coded as either 0 or 1.

By first importing the Logistic Regression library from sklearn. Linear Class, then creating the LogReg classifier object, and finally fitting our data, we used the Logistic Regression approach in our model.

##### ii) Decision Tree

It is a supervised ML technique that is mainly used for implementing classification problems and the prediction is done on the basis of entropy and

information gain. The accuracy of the tested model came out to be 90.9%. Since its accuracy is not bad we have not taken it because its accuracy was not the highest.

In our model, we used the Decision Tree technique as follows:

- i) To begin, we imported the DecisionTreeClassifier library from the tree class in sklearn.
- ii) Next, the Decision Tree classifier object was built.
- iii) Lastly, we fitted the data.

iii) SVM

Is also a supervised ML algorithm that is primarily used for solving classification problems. It creates the best line also known as the decision boundary that classifies different data points into correct categories as follows:

- i) First, we imported the SVC library from sklearn.svm Class;
- ii) Next, we generated the SVM classification object; and
- iii) Finally, we fitted our data.

iv) Random Forest

A large number of choice trees are built during the preparation phase to create random forests, also known as random decision forests, which are an ensemble learning technique for grouping, relapsing, and various tasks. The class selected by the majority of trees is the outcome of the random backwoods for classifying errands. The mean or normal expectation of the singular trees is returned for relapse projects. Choice trees have a tendency to overfit their preparation set in arbitrary choice timberlands. In general, irregular forests outperform choice trees, but their precision is inferior to that of angle-supported trees. However, the quality of the information can affect how well they function.

We used the RF technique in our model as follows:

- i) First, from the sklearn neighbors Class, we imported the Random Forest library.
- ii) Following that, we developed the RF classification object.

iii) Then, we fitted the data.

v) Naive Bayes

Based on the Bayes theorem, the Naive Bayes classifier is a probabilistic classifier. which performs best in multiclass predictions as compared to other algorithms.

We used the NB technique in our model by:

first importing the Gaussian Classifier library from the sklearn. Naive bayes Class, creating the NB classifier object, and then fitting the data.

## **1.5 Organization**

This project report's content is divided up into five chapters;

Chapter 2 of the project report describes the literature survey, where different research papers related to this project work have been included. Around 10 research papers are added in this report clearly based on the crop recommendation system.

Some of them just predicted the crops using different ML models and further selected the best one using accuracy. Others have also added GUI [1] or some website for better interaction and to ease the use of the models.

Chapter 3 summarizes the system development where various analytical and developmental analysis is explained along with the design and algorithms of the model development. Model development technique is explained further by adding various ML models including DT, NB, SVM, etc.

Their mathematical explanation is also shown using the formulas used.

Chapter 4 provides an account of the performance analysis where we have mentioned the most suitable model for this project after comparing the accuracies of these models. It also provides the outputs and step by step results at various stages.

Chapter 5 presents a brief summary of conclusions, and future work based on the research done during the implementation of the project.

In the end, References is added where all the research papers are mentioned that were needed for the better implementation of the model.

## CHAPTER-2 LITERATURE SURVEY

Farming assumes an essential part of India's economy. 54.6% of the complete labour force takes part in farming and unified area exercises (Enumeration 2011) and represents 17.1% percent of the Gross Value Added (GVA) for the country in 2017–18 (at current costs). But there are gains notwithstanding a reduction in farming's contribution to Net Worth Added. Ranchers play an important role because food is a basic necessity that depends on the success of agribusiness. The first layer neurons in the multi-facet discernment feed-forward brain organization provide the outcome to the succeeding layer in a unidirectional manner so that the neurons are not gotten from the opposite request. The feed-forward brain network assigns distinct responsibilities for each layer by integrating the three layers of information, yield, and concealment. The resulting layer consists of a variety of neurons that are equal to the previously recorded number of amounts taken from the information and made available for perceptron reactions. The At the point when a SOM is given too little data or an excess of superfluous data in the loads, the groupings found in the guide may not be completely exact or enlightening.

The study in [14] intended to suggest the most appropriate harvest in view of info boundaries like Nitrogen (N), Phosphorous (P), Potassium (K), PH worth of soil, Stickiness, Temperature, and Precipitation. The Paper anticipated the precision representing things to come creation of eleven distinct harvests, for example, Utilizing various regulated AI techniques in India, this study examines the yield potential of rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mungbean, dark gramme, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and espresso crops.. The proposed framework applied various types of AI calculations like Choice Trees, Gullible Bayes (NB), Backing Vector Machine (SVM), Strategic Relapse, Arbitrary Woodland (RF), and XGBoost. Various exactnesses were determined of this multitude of applied models and the outcomes were as per the following:

The most dependable model was XGBoost with a precision of 99.31 percent followed by Arbitrary Backwoods model and Credulous Bayes with an exactness of almost 100%. The following model was Calculated Relapse which estimated the precision of 95.22 percent followed by Choice Tree, whose exactness was 90%. Among this multitude of applied models, the model with least exactness was SVM, i.e, just 10.68 percent.

This paper successfully proposed and carried out a savvy crop suggestion framework, which can be handily utilized by ranchers all over India. This framework planned to assist the ranchers in coming to an educated conclusion about which trimming to develop contingent upon certain boundaries like Nitrogen, Phosphorus, Potassium, PH Worth, Temperature, Dampness, and Precipitation.

The suggestion of different yields of India was made utilizing different AI calculations like Innocent Bayes, Choice Tree, Strategic Relapse, Arbitrary Woods, Backing Vector Machine, and XGBoost. These six different categories of AI computations were the focus of the experiment, and XGBoost achieved the highest level of accuracy.

The study in [15] used Gradient Boosting calculation to anticipate the best reasonable harvest for a specific land in view of a portion of the dirt variables which incorporates N, P, K and pH of the dirt and ecological elements which incorporates temperature, moistness, and precipitation. Different models were additionally carried out alongside the Slope Helping which are KNN, Gullible Bayes, Choice Tree, and Irregular Woods, yet the most elevated precision was by given by the Inclination Supporting calculation which was 98.18 percent, trailed by the exactness of Arbitrary Backwoods which is 98%, aside from it the exactness of KNN estimated was 97.45 percent. The following model was Innocent Bayes which estimated the precision of 96.72 percent followed by the Choice Tree which estimated the exactness of 97.09 percent.



The study in [16] collected a dataset, which contained data about precipitation, climatic circumstances and different soil supplements. That gave a superior comprehension of patterns of yield creation with regards to geological and climate factors. The framework likewise anticipated the absence of parts for growing a particular harvest. The prescient framework has likewise ended up being an incredible aid for the farming business. The issue of supplement deficiency in regions, which occurs because of the explanation of establishing mistaken crops at some unacceptable timeframe, is diminished with the assistance of this prescient framework.

This prescient framework proposes by giving farmers information about prerequisites of various minerals and climate or encompassing circumstances, which are reasonable for developing specific yields.

Likewise, this venture redirected the centre towards the lack of various minerals important to develop a few harvests, and proposed solutions for taking out their deficiency. The framework considered factors which incorporate soil arrangement like N, P, K and pH of the dirt, and ecological variables like temperature, precipitation and dampness. This framework produced different outcomes in the testing stage for checking the exactness of the models executed, for suggesting the best appropriate harvest.

Various models were executed which incorporates Choice Tree, Innocent Bayes, SVM, Strategic Relapse, Irregular Backwoods and XGBoost and in the wake of applying every one of the above expressed models this framework at long last reached resolution to utilize the Arbitrary Woods calculation to fabricate the prescient framework as the exactness of the Arbitrary Woodland is 99.09 percent which is the most extreme among every one of the models carried out. Likewise, this model was executed as a site, which utilizes html, css, flagon and JavaScript.

Based on the available data, the study in [17] assesses crop yield production. To increase harvest efficiency, the harvest yield was predicted using data mining

techniques. Putting all the information together at the outset of an information mining plan is essential.

The data used in this proposed work were gathered from the 31 locations in Tamil Nadu, India, for the years 2000 to 2012, and they total about 10,000 records. The initial data collection is carried out for the Indian state of Tamil Nadu. Each region in this collection is identified by its unique longitude and geographic breadth.

Nine information highlights—Year, Region, Harvest, Region, Tanks, Bore Wells, Open Wells, Creation—are selected from the information.

This framework goes through two stages: Training and the Testing stage.

In the preparation stage the information was assembled and pre-processing was finished. The pre-handled information was bunched utilizing the k-implies grouping calculation. The affiliation rule mining process was applied on bunched information to track down the standards. The preparation stage closes with various produced rules.

In the testing stage, the yield esteem is anticipated in view of the produced rules. The work begins with a preprocessing step. In this step the gathered information was pre-handled. In the preprocessing, a few information was eliminated from the informational index which were not reasonable for the yield creation.

After the preprocessing step, the informational collection was bunched utilizing the k-implies grouping calculation.

The study in [18] stated the regression models for predicting the yield of crops like cotton, wheat and maize depending on soil weather and crop parameters. Correlation was made between the different relapse models in view of  $R^2$ ,  $R^2$  and  $1.1PPE$  measurements.

The review made sense of the tests done on wheat, cotton and maize informational collections utilizing quadratic, unadulterated quadratic, direct, polynomial, summed up straight relapse and stepwise straight relapse models. It likewise looked at the outcomes acquired from them. The execution was finished on the Windows 7 working framework involving MATLAB R2013a as

the programming instrument. Precision of these expectation models were estimated utilizing ( $R^2$ ), Root Mean Square Blunder (RMSE) and Mean Rate Expectation Mistake.

The prediction results for wheat yield forecast showed the outcomes in light of  $R^2$  measurement and MPPE values for wheat yield expectation utilizing straight, unadulterated quadratic, communications, quadratic, polynomial and GLM models. As noticed, the  $R^2$ -an incentive for the GLM relapse model was higher than the other models while MPPE is lower. From this it tends to be induced that the GLM model precisely predicts the wheat yield than different models. In this, the GLM model had a lower RMSE esteem than different models.

The prediction results for maize yield expectation showed the outcomes in light of  $R^2$  measurement and MPPE, for maize yield expectation utilizing direct, unadulterated quadratic, connections and quadratic and polynomial models. As can be noticed, the  $R^2$  for the unadulterated quadratic relapse model was seen as higher while MPPE was lower than the other models. From this, it was derived that the unadulterated quadratic model precisely anticipated the wheat yield better compared to different models. Here, the unadulterated quadratic model had a lower RMSE esteem than different models.

The prediction results for cotton yield forecast showed the outcomes in light of  $R^2$  Measurement and MPPE, for cotton yield forecast utilizing direct, unadulterated quadratic, cooperations, quadratic, polynomial and proposed SLR relapse models. As can be noticed, the  $R^2$ -an incentive for the SLR model was higher while 11: FPE was lower than the other models. From this it was derived that the SLR model precisely anticipated the cotton yield better compared to different models. It obviously showed that the proposed SLR relapse model had a lower RMSE esteem than different models.

The results showed that the proposed relapse model is a reasonable strategy for predicting yield creation. The aftereffects of various models were analysed in light of the root mean square,  $R^2$  measurements and rate expectation blunder. The model which gave the lower Root mean square, rate expectation mistake

and Higher R2 measurements values was viewed as the best model for crop yield forecast.

The study in [19] gave a quick analysis of data mining techniques, horticultural practices, ranch types, soil types, and expectation using the Different Straight Relapse (MLR) method for the selected district. This work primarily focused on deconstructing rural examinations of organic and inorganic farming, time plant development, benefit and loss of knowledge, and dissecting land business land in a specific location and correlation between watered and unirrigated land. It focused informational collections on the natural, inorganic, and land from which the expectation in agribusiness would be achieved. The goal was to evaluate the differences in cultivation practices and predictions between organic and inorganic farming. Finding appropriate information models that achieve high precision and high oversimplification was the goal of the research. concerning yield forecast capacities. The model used various techniques for agriculture that included:

1. Genetic Algorithm
2. ANN
3. Nearest Neighbour
4. Rule Induction, and
5. Memory based reasoning.

It looked at information mining's role in the horticulture industry and how some creators tied it to the farming industry. It also looked at other information mining applications to address various rural concerns. This document collected the work of multiple authors in one place to make it easier for specialists to learn about the state of information mining techniques and applications in the horticulture industry today.

The study in [20] proposes to help farmers to check the soil quality to get a good crop yield. The prediction assisted the farmers with foreseeing the yield of the harvest prior to developing onto the agribusiness field. AI is a fundamental methodology for accomplishing pragmatic and successful answers for this issue.

According to the analysis carried out using an information mining approach, the work aims to help ranchers check the dirt quality. The process of dissecting information into smaller pieces based on opposing viewpoints and compiling those pieces into useful data is known as information mining. As a result, the framework moves toward reasonable composts to estimate the dirt quality and increase the harvest output for a better development based on their dirt types.

The assignment suggested a strategy to predict the harvest's yield. Before making any improvements to the field, the rancher will carefully assess the harvest's yield in relation to the area of land. The project's goal is to help ranchers check the quality of the soil using information mining techniques.

The system had five units.

In soil test examination, the client enters pH and site. Result of this unit is investigation aftereffects of the portion of supplements in that dirt. Soil crop matching unit finds the matching yield that could be developed in that dirt by correlation with the harvest data set. In the harvest data unit, clients can choose a yield and view data about it. In the compost data unit, the client can choose a manure and show data about it.

The study in [21] stated that it concentrated on crop prediction using artificial neural networks and existing data. Since the majority of the computationally intensive work is done during the training phase alone and there is no need for a testing phase, an artificial neural network that is utilized for classification and prediction has been created. The proposed approach was designed to aid farmers in making crop output predictions based on a variety of variables, including temperature, rainfall, humidity, soil nutrients, etc. The Artificial Neural Network algorithm, which is based on the biological neural network of the human brain, was used to achieve it. The ANN, which is employed to resolve complicated issues, is based on the neural mechanisms of the human brain. Three layers, including the input layer, the hidden layer, and the output layer, make up an artificial neural network. The input layer is the first layer of a neural network and is responsible for serving the input to the layers above it.

The proposed framework considers natural variables, soil factors, climate, soil ripeness and creation information throughout the last year and recommends the most noteworthy yielding harvest that can be developed under given ecological circumstances. The ANN innovation records generally potential yields and offers ranchers the chance to choose the most beneficial ones.

Mix of man-made brainpower and agribusiness will help the greater part of 's ranchers soon. Contrasted with others, Brain Organization is the best answer for horticultural issues (e.g., Crop yield expectation). The utilization of ANNs assumes a significant part in foreseeing agribusiness in view of specific key variables and furthermore in evaluating future yields in light of current/verifiable information.

## **CHAPTER-3 SYSTEM DESIGN & DEVELOPMENT**

### **3.1 System Design**

To develop a software program, we should be familiar with the concept of SDLC life cycles. A SDLC cycle is a software development framework which allows a user to manage and implement its system in a proficient manner. In our project the model that we have used is the waterfall model. It is a very basic and easy to use life cycle model. It uses a linear sequential flow to make the software. This means that the next process cannot be started without completing the previous process. Thus, there is no overlapping of processes. The result of one fragment is the input for the next part. The waterfall model includes the stages as shown in the “Fig.3.1”. All of the below mentioned stages are very important for system development and input of a process is the outcome of the previous one itself.

### **3.2 System Development**

System Development is the development of a system for a unique situation. Having a proper methodology helps us in bridging the gap between the problem statement and turning it into a feasible solution. It is usually marked by converting the System Requirements Specifications (SRS) into a real-world solution.

The following inputs are required for system design:

- Statement of work.
- A plan for determining requirements.
- Analysis of the current scenario.
- A conceptual data model and metadata are proposed system needs (data about data).

Through the phases of requirement origination, analysis, design, implementation, testing, and maintenance, progress is shown as owing slowly downwards (like a waterfall) in the waterfall model, a sequential software development process Intelligent Crop Recommendation System using ML.

The system specifications are converted into a software representation while keeping the requirements in mind. The designer places a focus on things like algorithms, data structure, software architecture, etc. during this phase.

### 3.3 Model Development

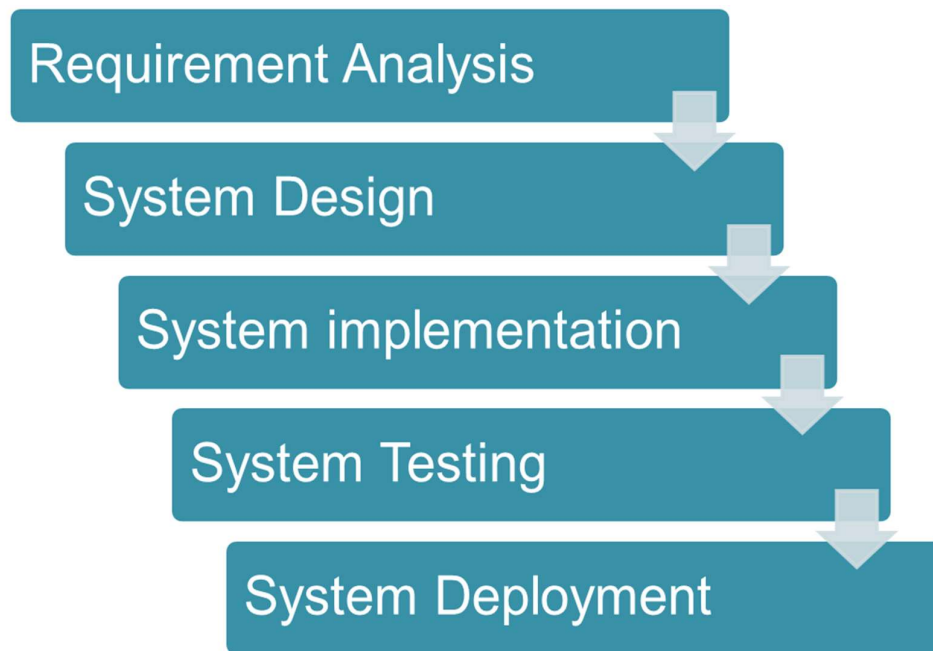


Fig. 3.1. Waterfall Model

#### 3.1 System Requirement Analysis

A crop recommendation system using wireless sensor network and machine learning algorithms can be analysed using a system analysis approach. System analysis involves studying the system's components, functions, and interactions to identify ways to improve its performance and efficiency.

The system analysis of a crop recommendation system using wireless sensor network and machine learning algorithms involves the following steps:

1. Identify the system's goals and objectives
2. Identify the system's inputs and outputs
3. Identify the system's processes
4. Identify the system's constraints and limitations
5. Evaluate the system's performance



By analysing the crop recommendation system using wireless sensor network and machine learning algorithms, it is possible to identify ways to improve the system's performance and efficiency. For example, improving the accuracy and reliability of the data collected by the WSN [11], and using advanced machine learning algorithms can help generate more accurate recommendations. Additionally, using a user-friendly interface can help increase farmer adoption of the system.

### 3.2 Performance Analysis

The performance analysis of a crop recommendation system using wireless sensor network and machine learning algorithms involves evaluating the system's accuracy, response time, and scalability. Here are some key factors to consider when analysing the performance of such a system:

1. Accuracy
2. Response Time
3. Scalability
4. User Adoption

To improve the performance of the crop recommendation system, the following measures can be taken:

1. Improve Data Collection
2. Advanced Machine Learning Algorithms
3. Real-Time Analysis
4. Cloud-based Deployment
5. User Feedback

### 3.3 Economical Analysis

An economic analysis of a crop recommendation system using wireless sensor network and machine learning algorithms involves evaluating the system's costs and benefits. Here are some key factors to consider when analyzing the system's economic viability:

1. Deployment Costs
2. Maintenance Costs

3. Benefits
4. Return on Investment (ROI)
5. Adoption Rate

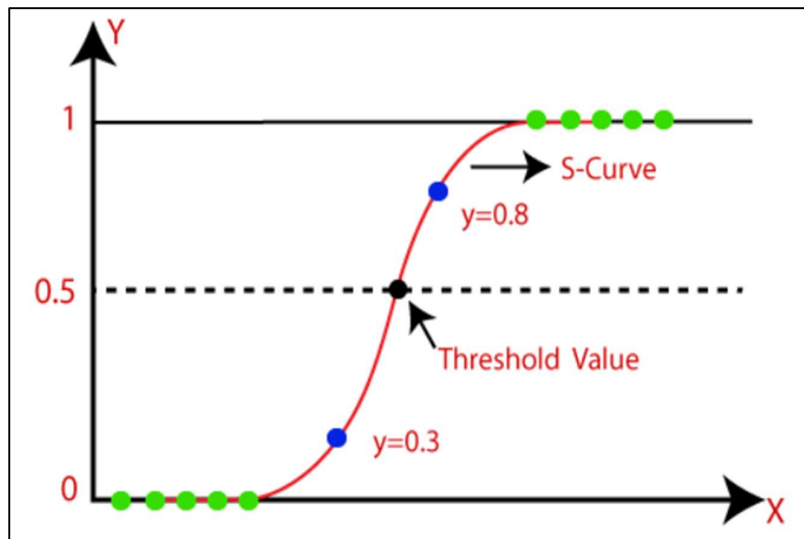
To improve the economic viability of the crop recommendation system, the following measures can be taken:

1. Cost-effective Deployment
2. Low Maintenance System
3. Improved Resource Management
4. Marketing and Outreach
5. Collaborations

#### 4. Algorithm

##### 1. Logistic Regression:

A categorical variable's outcome is predicted by logistic regression. As a result, the outcome should be a discrete or categorical value. However, rather than providing the exact value as 0 or 1, it provides the probabilistic properties that fall somewhere in the range of 0 and 1. It may very well be either Yes or No, 0 or 1, true or false, and so on. Instead of fitting a regression line in logistic regression, we fit a "S" shaped logistic function that anticipates the two most extreme values (0 or 1).



**Fig.3.2. Sigmoid Function for Logistic Regression [12]**

LoR equation:

The linear regression equation yields the logistic regression equation. The following are the mathematical steps to obtain Logistic Regression equations:

- The equation of straight line written as :

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

- Now, let's divide the equation by (1-y) because the value of 'y' ranges between 0 and 1.

$$\frac{y}{1-y} ; 0 \text{ for } y=0, \text{ and infinity for } y=1$$

- We also need the range to lie between -[infinity] to +[infinity] , therefore

$$\log \left[ \frac{y}{1-y} \right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

## 2. Decision Tree:

A yes/no "decision" is produced by a supervised ML algorithm given an item or circumstance characterized by a collection of attributes as input. The root node and the leaf node of a decision tree are the two nodes from which the splitting is to be carried out based on entropy and information gain.

The general formula of entropy for the dataset is:

$$E = - \sum_{i=1}^N P_i \log P_i$$

The general formula of Information Gain is the expected reduction in entropy caused by the splitting of data:

Information Gain = Entropy(S) - [(Weighted Avg.) \* Entropy (each feature)]

3. Naive Bayes:

It is a probabilistic classifier that solves classification algorithms and predicts the likelihood of an object based on the Bayes Theorem. It is a supervised machine learning technique.

The formula of bayes theorem is given as :

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Where,

$P(A|B)$  is Posterior probability,

$P(B|A)$  is Likelihood probability,

$P(A)$  is Prior probability,

$P(B)$  is Marginal probability.

4. Support Vector Machine(SVM):

It is a well-liked supervised machine learning approach that is applied to both classification and regression issues. The goal of the SVM algorithm is to establish the optimum decision boundary or line that may divide a multidimensional space into distinct classes, allowing us to quickly categorize the data points into those groups.

A hyperplane is the ideal boundary for making decisions.

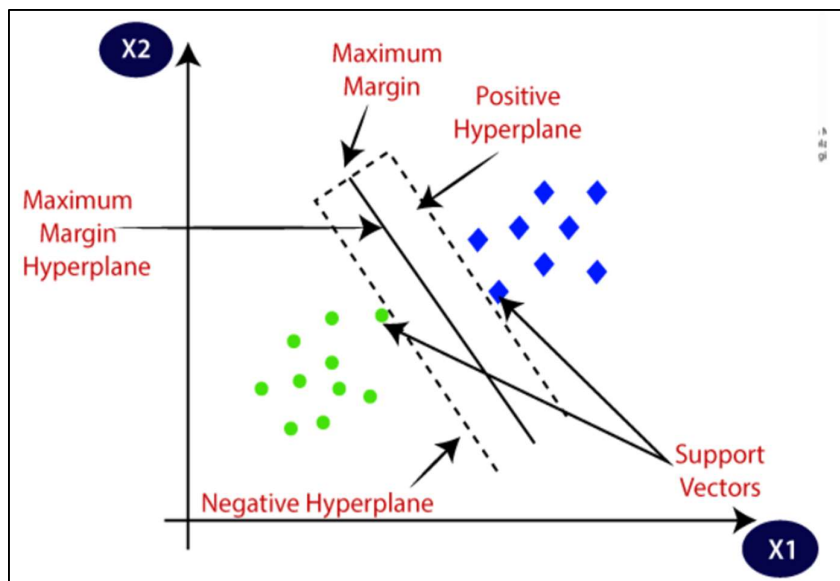


Fig. 3.3. SVM Hyperparameters [13]

SVM selects the extreme vectors that assist in the creation of the hyperplane, known as Support Vector Machines.

- SVM is two types:
  - Linear SVM: For data that can be separated linearly, use linear SVM.
  - Non-linear SVM: It's utilized with data that's been separated non-linearly.

The hyperplane is the line or optimal boundary that is found using the SVM technique. Margin is the separation between two vectors. The ideal hyperplane is referred to as having the largest margin.

#### 5. Random Forest (RF):

Random forest is a popular machine learning algorithm that belongs to the family of ensemble methods. It is a decision tree-based approach that combines multiple decision trees to create a robust and accurate prediction model. The algorithm works by creating a forest of decision trees, where each tree is trained on a random subset of the data and a random subset of the features.

In a random forest, the input data is split into multiple random subsets, and decision trees are trained on each subset. This process is repeated multiple times, creating a forest of decision trees. Each tree predicts the output, and the final output is obtained by averaging the predictions of all the trees in the forest.

Random forest is a powerful algorithm that has several advantages over other machine learning algorithms. Some of the key benefits of random forest include:

1. Robustness: Random forest is a robust algorithm that can handle noisy data and missing values.
2. Scalability: Random forest is a highly scalable algorithm that can handle large datasets with high-dimensional features.
3. Feature importance: Random forest provides a measure of feature importance, which can be used to identify the most important features in the dataset.
4. Nonlinear relationships: Random forest can capture nonlinear relationships between the features and the output.

Overall, random forest is a powerful algorithm that can be used for a wide range of machine learning tasks, including classification, regression, and feature selection.

Ensemble learning Algorithms:

Ensemble learning is a machine learning technique that combines multiple individual models to improve overall performance. Here are some common ensemble learning algorithms:

1. Bagging
2. Boosting
3. Random Forest
4. Stacking
5. AdaBoost
6. Gradient Boosting

Disadvantages of Ensemble Learning:

- Computational cost is more.
- Time Consuming.

## CHAPTER-4 EXPERIMENTS & RESULT ANALYSIS

This system has three units as shown in the above figure. In the soil inputs, the user has to enter the N,P,K and the pH value. Then in the weather inputs the user has to enter the numeric values of Rainfall, Temperature, and Humidity of the surrounding from which the final output will be the predicted crop

For the reasons for this undertaking we have utilized five popular algorithms:

Logistic Regression, Decision Tree, Naive Bayes, SVM, and Random Forest.

All these algorithms are based on supervised learning.

Our overall system is divided into two modules:

- Applying these ML algorithms into our model, and
- Predicting the best suitable crop.

Before applying ML algorithms, firstly data analyzing and pre-processing are to be done.

The dataset is taken from the Kaggle [10] named as ‘Crop Recommendation’ csv file containing 2200 rows and 8 columns. Fig.4.1 below depicts the dataset that is taken:

	<b>N</b>	<b>P</b>	<b>K</b>	<b>temperature</b>	<b>humidity</b>	<b>ph</b>	<b>rainfall</b>	<b>label</b>
<b>0</b>	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
<b>1</b>	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
<b>2</b>	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
<b>3</b>	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
<b>4</b>	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
<b>5</b>	69	37	42	23.058049	83.370118	7.073454	251.055000	rice
<b>6</b>	69	55	38	22.708838	82.639414	5.700806	271.324860	rice
<b>7</b>	94	53	40	20.277744	82.894086	5.718627	241.974195	rice
<b>8</b>	89	54	38	24.515881	83.535216	6.685346	230.446236	rice
<b>9</b>	68	58	38	23.223974	83.033227	6.336254	221.209196	rice
<b>10</b>	91	53	40	26.527235	81.417538	5.386168	264.614870	rice

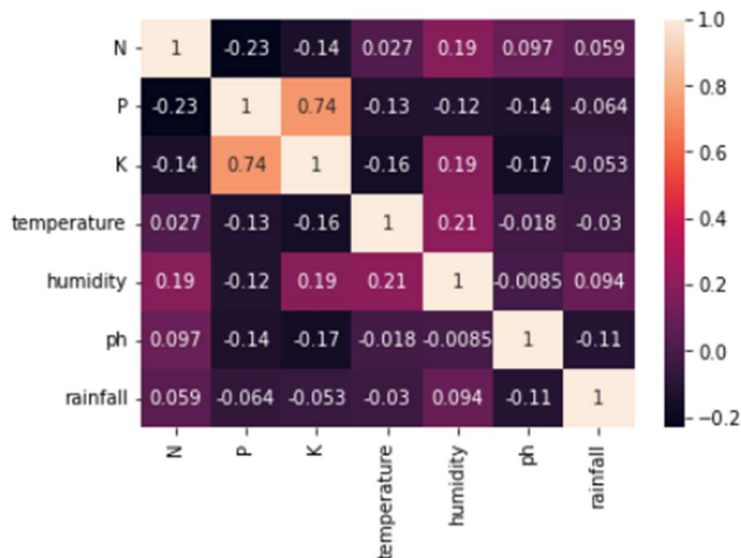
**Fig.4.1. Dataset**

The input variables are described in the figure where the mean values, standard deviation, minimum value, maximum value, count of the rows and the percentage with respect to each feature is shown.

	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

**Fig.4.2. Description of features**

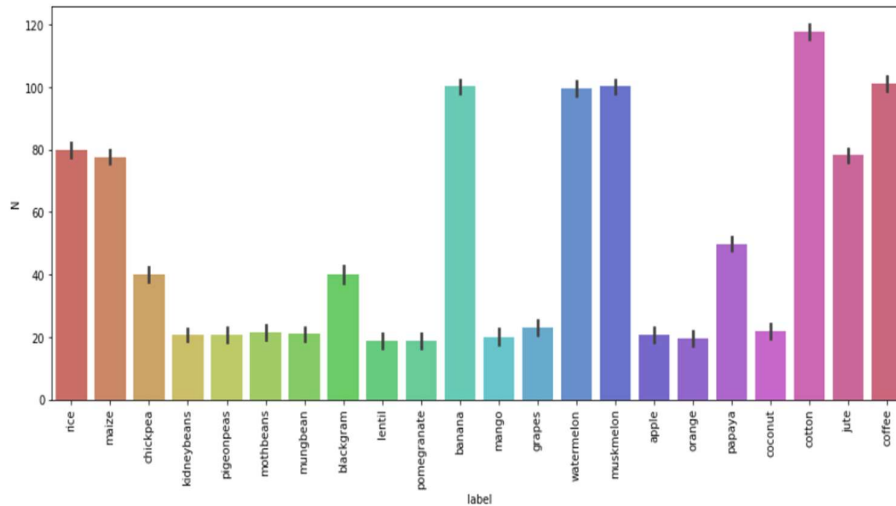
The strength of correlation among the variables are visualized using the heat map that represents the correlation between the feature variables, so that we can find the feature that is best for ML model building. The heat map further transforms this correlation matrix into color coding.



**Fig.4.3. Confusion Matrix**

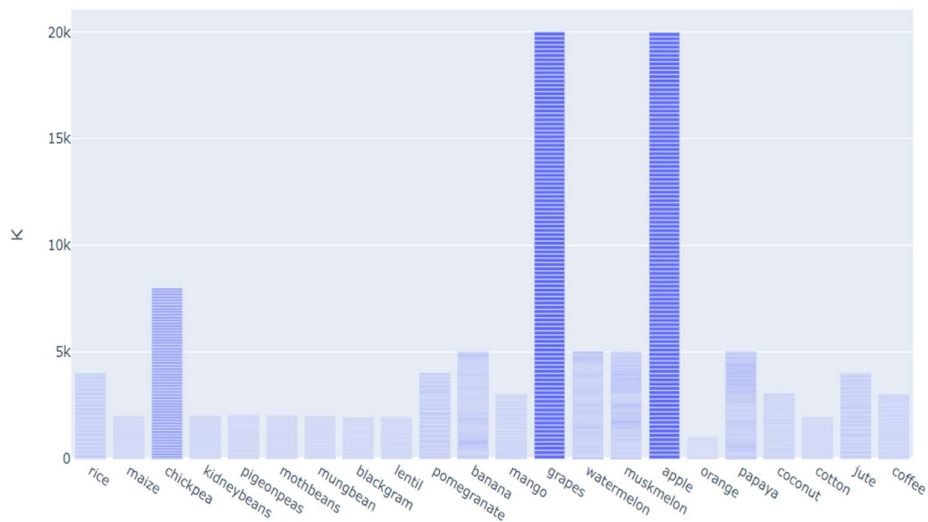


Figure 4.4 shows the relationship between the Nitrogen present in the soil with respect to different crops. As seen in the figure the value of the nitrogen depicts the best suitable crop that needs to be sown.



**Fig. 4.4 Nitrogen value determining different Crops**

Figure 4.5 shows the relationship between the Potassium present in the soil with respect to different crops. As seen in the figure the value of the Potassium depicts the best suitable crop that needs to be sown.



**Fig. 4.5. Potassium(K) value with respect to different crops**

After all the preprocessing and the visualizing of the data, we have implemented different Machine Learning Models which includes Logistic Regression, SVM, Decision Tree, Naive Bayes, and Random Forest so that we can easily come to know that which crop is to be grown on particular land based on the accuracy of the different models.

The first model that we have implemented is Logistic Regression which is shown in Fig.7. and the input features are the soil factors which includes N, P, K and the environmental factors which includes temperature, rainfall, and humidity. Based on these factors the suitable crop is recommended. After this the accuracy came to be 96.81 percent and based on the accuracy classification report is generated and different evaluation parameters are shown for a particular crop on a particular land.

Further, the second model that we have implemented is the Decision Tree which is shown in Fig.8. and the input features are the soil factors which includes N, P, K and the environmental factors which includes temperature, rainfall, and humidity. Based on these factors the suitable crop is recommended. After this the accuracy came to be 90.90 percent and based on the accuracy classification report is generated and different evaluation parameters are shown for a particular crop on a particular land.

Now, the next model that we have implemented is Naive Bayes shown in Fig.9. and the input features are the soil factors which includes N, P, K and the environmental factors which includes temperature, rainfall, and humidity. Based on these factors the suitable crop is recommended. After this the accuracy came to be 99.77 percent and based on the accuracy classification report is generated and different evaluation parameters are shown for a particular crop on a particular land.

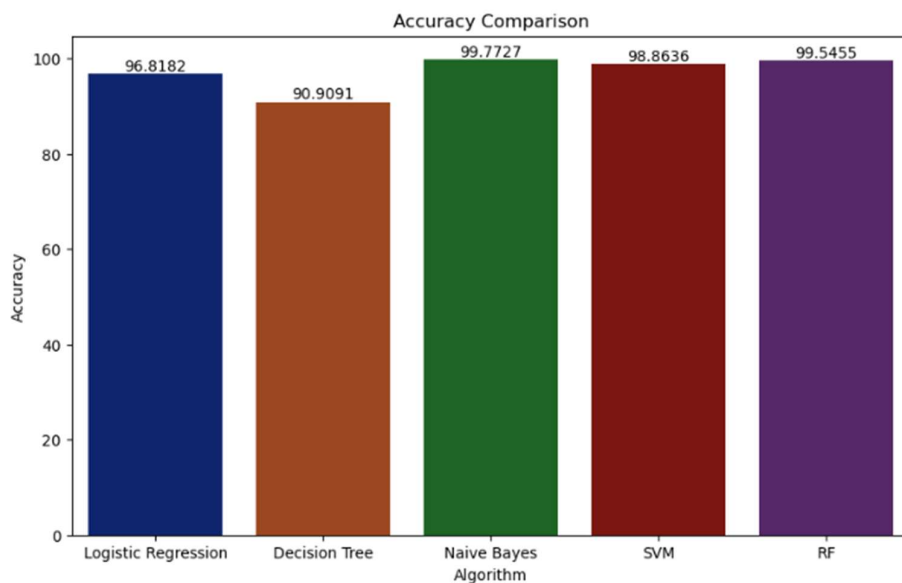
Further, the next model that we have implemented is the Support Vector Machine presented in Fig.10. and the input features are the soil factors which

includes N, P, K and the environmental factors which includes temperature, rainfall, and humidity. Based on these factors the suitable crop is recommended. After this the accuracy came to be 98.86 percent and based on the accuracy classification report is generated and different evaluation parameters are shown for a particular crop on a particular land.

The final model that we have implemented is Random Forest shown in Fig.11. and the input features are the soil factors which includes N, P, K and the environmental factors which includes temperature, rainfall, and humidity. Based on these factors the suitable crop is recommended. After this the accuracy came to be 99.54 percent and based on the accuracy classification report is generated and different evaluation parameters are shown for a particular crop on a particular land.

#### **Accuracy Comparison:**

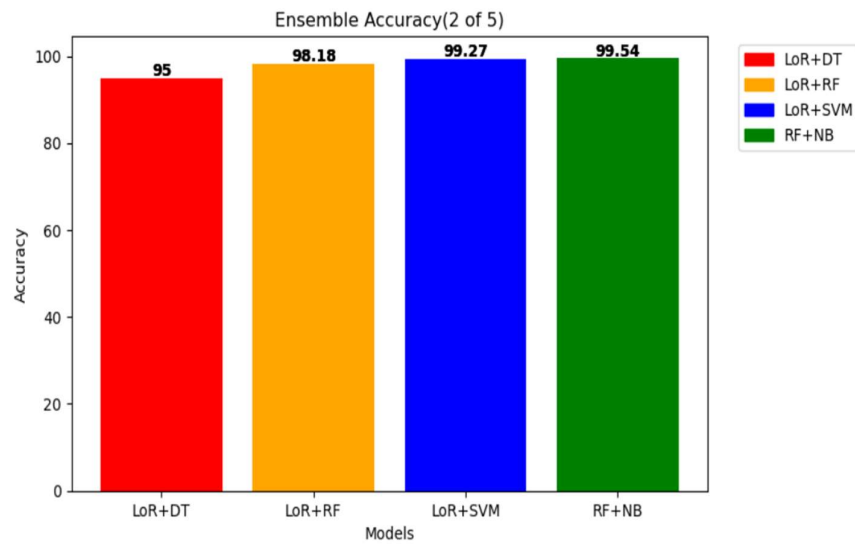
After all the model implementation, we have compared the accuracy of all the models shown in Fig.4.6. The accuracy of Naive Bayes is maximum among all the models, and we finally came to the conclusion to use the Naive Bayes algorithm to build the predictive system as the accuracy of the Naive Bayes is 99.77 percent.



**Fig.4.6 Accuracy Comparison of all Models**

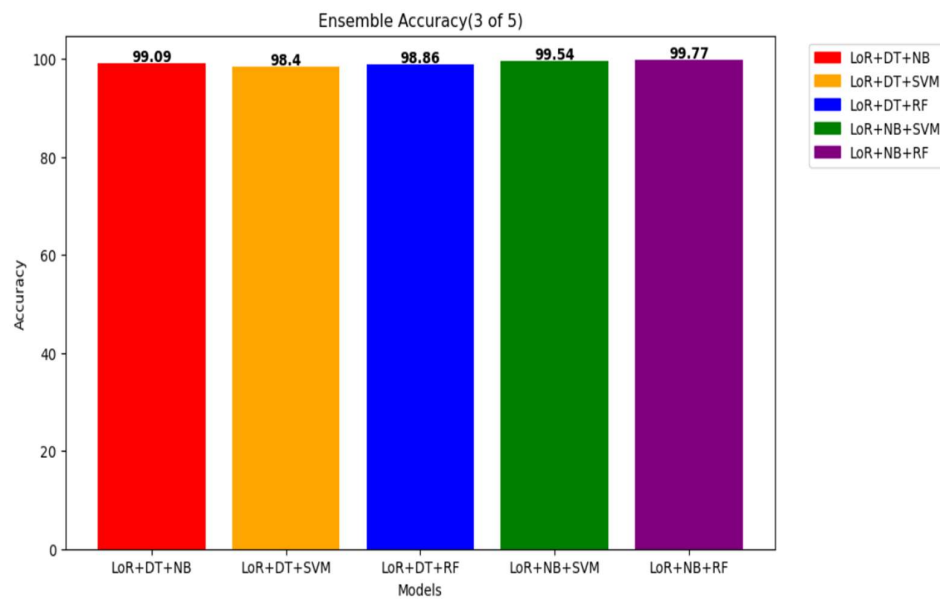
## Accuracy comparison using Ensemble learning technique:

### A. Using 2 out of 5 models



**Fig.4.7 Accuracy Comparison of Ensemble Models (2 of 5)**

### B. Using 3 out of 5 models

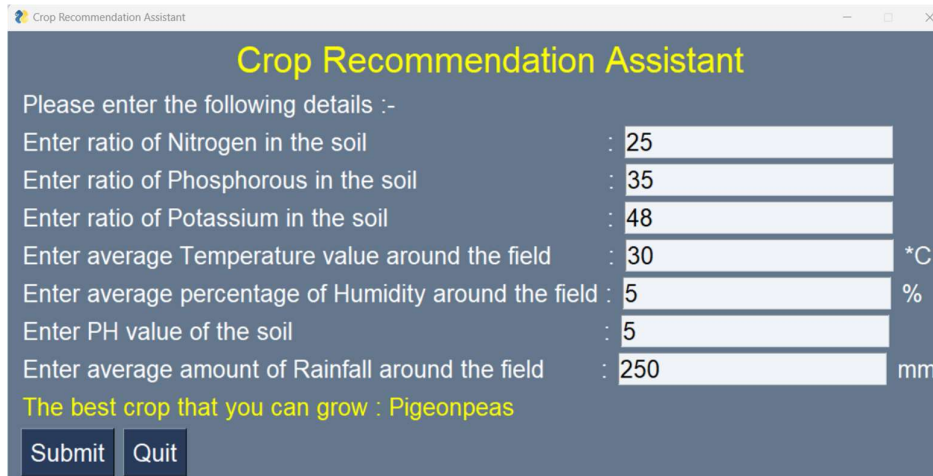


**Fig.4.8 Accuracy Comparison of Ensemble Models (3 of 5)**

### Graphic User Interface:

The next step was to create an easy to GUI [1] with python tkinter module where the user can give the details about the soil and the climate and can get the best suitable crop for the land as an output.

We used a voice assistant to make the model more effective, so that the user can access the speech feature along with the display.



Crop Recommendation Assistant

**Crop Recommendation Assistant**

Please enter the following details :-

Enter ratio of Nitrogen in the soil	:	25	
Enter ratio of Phosphorous in the soil	:	35	
Enter ratio of Potassium in the soil	:	48	
Enter average Temperature value around the field	:	30	*C
Enter average percentage of Humidity around the field	:	5	%
Enter PH value of the soil	:	5	
Enter average amount of Rainfall around the field	:	250	mm

The best crop that you can grow : Pigeonpeas

Submit Quit

**Fig. 4.9 Crop Recommendation Assistant**

## CHAPTER-5 CONCLUSIONS

### 5.1 Conclusion

Since agriculture is the key to our country's economy, therefore any effort done in this sector will put a great effect altogether. Crop yield forecasting is still a difficult task for farmers nationwide., so we need to be more sincere about it. Also, the farmers in our country are not well aware about the different factors that are affecting crops because of this they face a lot of challenges in growing a crop. In order to accomplish this we have proposed a system that will help the farmers to choose the right crop that will increase their productivity. This system can be extended to the web and also the GUI [1] where millions of farmers can access this. This project has expressed the best suitable recommendation of various crops in India using different ML algorithms like Logistic Regression, Decision Tree, Naive Bayes, SVM, and Random Forest Out of these algorithms, Random Forest algorithm achieved the best accuracy result.

### 5.2 Future Scope

Making farming supportable and versatile to the continuous change in environment and social construction is really difficult for researchers and specialists across the globe. The agricultural framework requests change and a multidisciplinary approach. Also, accurate agricultural methodologies were given due significance for expanding creation and efficiency from exactly the same restricted assets. The methodology needs data from different sources and effective utilization of them in significant fields. This need led to a developing interest in information disclosure from immense heaps of information created out of different exploration and overview works. The rise of Information Mining strategies altered the field of data age and example acknowledgment. Information mining, however, is a young science that has numerous applications in agriculture and other allied fields and has great potential for the future of agriculture.

Agriculture is encountering a progress stage driven by populace strain and environmental change. More creation and efficiency are being normal from Restricted assets. New concentrated research is being finished to investigate ways of expanding creation with ideal utilization of assets keeping up with supportability. This led to the utilization of present day complex PCs that helped advancements in the farming examination. Because of broad utilization of PC and reasonable storage spaces, there is a gigantic abundance of information implanted in immense information bases of various agric-unified undertakings. A new improvement in Data Innovation for the agriculture field has turned into a fascinating examination region to foresee/predict the yield. The amount of data stored in the modern world has been massively increasing over time, and most of it is unstructured and unable to be processed in any way to separate useful data using mining techniques. The Multiple Linear Regression (MLR) methods are used to provide a succinct investigation of crop yield expectation in this paper. This project's core focus is on deconstructing the agriculture investigation of natural shaping and inorganic framing, time development of the plant, benefit and loss of knowledge, and deconstructing the land business land in a specific region. investigating correlations between flooded and unirrigated land The expectation in agribusiness is done by focusing on natural, inorganic, and land informational collections.. The object is to gauge distinction in proficiency and expectation among natural and inorganic framing.

The objective of the system is to make a robust model, so we can work with a bigger dataset in future. We can also add more features in this project such as plant disease predictions and this research can be extended by applying different prediction techniques like SVR, Neural Networks, Fuzzy logic etc for predicting the yield of various crops. The system can be enhanced further by adding (i) image processing to detect crop disease where users can upload pictures of the diseased crop and can obtain recommended pesticides.

(ii) Implementation of a smart irrigation system to check the soil and weather conditions.

Apart from tkinter GUI [1], a website or webpage can be created giving the direct access to the user for predicting the best suitable crop.

### **5.3 Applications in Agriculture**

There are various Applications in agriculture include some of the following:

#### **5.3.1 Crop Selection and Crop Yield Prediction**

The choice of the appropriate yield that will be planted plays a key role in increasing the harvest yield. It depends on a number of factors, including the kind of soil and its composition, the environment, the district's geology, crop output, market prices, and so forth. In terms of crop choice, which depends on various factors, techniques like Counterfeit brain organizations, K-closest neighbors, and Decision Trees have carved out a niche for themselves. Crop selection has been done with consideration of ML and the effects of common calamities like starvations. The utilization of fake brain organizations to pick the harvests in view of soil and environment has been shown by scientists. A plant supplement to the board framework has been proposed in view of ML techniques to address the issues of soil, keep up with its ripeness levels, and consequently further develop the harvest yield.

#### **5.3.2 Weather Forecasting**

Indian agriculture basically depends on occasional rains for the water system. Accordingly, an exact gauge of weather conditions can decrease the huge work looked at by ranchers in India including crop determination, watering, and collecting. As the farmers have unfortunate admittance to the Web because of computerized partition, they need to depend on the little data accessible with respect to meteorological forecasts. Cutting-edge, as well as precise climate data, is still not accessible as the weather conditions change progressively over the long haul. Specialists have been dealing with working on the exactness of climate forecasts by utilizing different calculations. Fake Brain networks have been embraced widely for this reason. Moreover, a climate forecast in view of ML strategy called Support Vector Machines had been proposed. These calculations have shown improved results over traditional calculations.



### 5.3.3 Smart Irrigation System

India's farming regions consume a huge amount of water. Environmental changes are occurring when groundwater levels gradually decrease and Earth's average temperature rises. The stream water for water system is a hotly contested topic in several Indian states. Many companies have developed sensor-based technology for clever farming that uses sensors to monitor the soil temperature, water level, supplement content, weather forecasts, and speculation reports in order to combat the water crisis. However, the high cost of such equipment dissuades the small landowners and ranchers in India from using them. These innovative devices are being developed in accordance with AI standards. Using the sensors and tools, it is also possible to record the supplement content of soil.

## REFERENCES

- [1] “Python - GUI Programming (Tkinter)”, [www.tutorialspoint.com/https://www.tutorialspoint.com/python/python\\_gui\\_programming.htm](http://www.tutorialspoint.com/https://www.tutorialspoint.com/python/python_gui_programming.htm) (accessed March, 2023).
- [2] “tkinter — Python interface to Tcl/Tk”, [docs.python.org/https://docs.python.org/3/library/tkinter.html](https://docs.python.org/3/library/tkinter.html) (accessed Feb, 2023).
- [3] “Python Pandas Tutorial”, [www.javatpoint.com/https://www.javatpoint.com/python-pandas](http://www.javatpoint.com/https://www.javatpoint.com/python-pandas) (accessed July, 2022).
- [4] “Python NumPy Tutorial”, [www.javatpoint.com/https://www.javatpoint.com/numpy-tutorial](http://www.javatpoint.com/https://www.javatpoint.com/numpy-tutorial) (accessed Nov. 3, 2022).
- [5] “Scikitlearn” <https://scikit-learn.org/stable/> (accessed July, 2022).
- [6] “SciPy tutorial”, [www.javatpoint.com/https://www.javatpoint.com/python-scipy](http://www.javatpoint.com/https://www.javatpoint.com/python-scipy) (accessed Dec, 2022).
- [7] “PyTorch”, [en.wikipedia.org/https://en.wikipedia.org/wiki/PyTorch](https://en.wikipedia.org/wiki/PyTorch) (accessed July, 2023).
- [8] “Matplotlib (Python Plotting Library)”, [www.javatpoint.com/https://www.javatpoint.com/matplotlib](http://www.javatpoint.com/https://www.javatpoint.com/matplotlib) (accessed July, 2022).
- [9] “seaborn: statistical data visualization” <https://seaborn.pydata.org/>
- [10] “Kaggle” <https://www.kaggle.com/> (accessed July, 2022).
- [11] “Wireless Sensor Network”, [en.wikipedia.org/https://en.wikipedia.org/wiki/Wireless\\_sensor\\_network](https://en.wikipedia.org/wiki/Wireless_sensor_network) (accessed Jan, 2023).
- [12] “Logistic Regression in Machine Learning”, [www.javatpoint.com/https://www.javatpoint.com/logistic-regression-in-machine-learning](http://www.javatpoint.com/https://www.javatpoint.com/logistic-regression-in-machine-learning) (accessed Aug, 2022).
- [13] “Support Vector Machine”, [www.javatpoint.com/https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm](http://www.javatpoint.com/https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm) (accessed Nov 2022).
- [14] Gosai, D., Raval, C., Nayak, R., Jayswal, H., & Patel, A. (2021). Crop Recommendation System using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 558–569.
- [15] Shariff, S., & B, S. R. (n.d.). *Crop Recommendation using Machine Learning Techniques*.
- [16] Lokhande, A., & Dixit, M. (2022). Crop Recommendation System Using Machine Learning. *International Research Journal of Engineering and Technology*.

- [17] Manjula, E., & Djodiltachoumy, S. (2017). A Model for Prediction of Crop Yield. *International Journal of Computational Intelligence and Informatics*, 6(4).
- [18] Shastry, A., Sanjay, H. A., & Bhanusree, E. (n.d.). Prediction of Crop Yield Using Regression Techniques. *International J Ownal of Soft Computing*, 12(2), 7.
- [19] Kumar, Rs., Research Scholar, Mp., & Professor, A. (2016). AGRICULTURAL ANALYSIS FOR NEXT GENERATION HIGH TECH FARMING IN DATA MINING. *International Journal of Scientific Development and Research*, 1.
- [20] Majumdar, J., Naraseeyappa, S., & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. *Journal of Big Data*, 4(1).
- [21] Chaudhary Farhana Kausar, K. (2020). PREDICTION OF CROP YIELD USING MACHINE LEARNING. In *International Journal of Engineering Applied Sciences and Technology* (Vol. 4).
- [22] Singh, V., Sarwar, A., & Sharma, V. (n.d.). Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. *International Journal of Advanced Research in Computer Science*, 8(5).
- [23] Preethaa, M. M., Professor, A., & Shree, Kv. (2016). CROP YIELD PREDICTION. *International Journal On Engineering Technology and Sciences-IJETS*, III.
- [24] Ojha, T., Misra, S., & Raghuwanshi, N. S. (2015). Wireless sensor networks for agriculture: The state-of-the-art in practice and future challenges. *Computers and Electronics in Agriculture*, 118, 66–84. <https://doi.org/10.1016/j.compag.2015.08.011>