CHAPTER- 1

INTRODUCTION

In the world of developing technologies, the success of sharing information will help the agriculturists in realizing and developing their potential. The information sharing is that the valuable and timely information is being shared between agriculturists, either formally or informally. The willingness of information sharing refers to the open attitude among agriculturists. This open attitude determines the degree and scope of information sharing. Using web-technologies like html and css we build the web application, we create dataset by gathering data from multiple resources and place them in place which is used to predict the price of the crop and results are subjected to non-linear test later priorities are set and rankings are given to the list of crops. Place information in our application and share that information to agriculturists whose data is collected and stored in the my sql server. we software to automatically send the updated information to the agriculturists in the form of text message. So that agriculturists no need to go to near by towns and cities to know the updated information. We will be machine learning algorithms to predict the price of the crop for the next two months. For prediction purpose we will be using Support vector machine (SVM), Naïve Baye's (NB) and K-Nearest Neighbour (KNN) algorithms to predict the cost of the crop production. Further, a ranking process is applied for decision making in order to select the classifiers results.

1.1 OBJECTIVE

- Data set collection from various sources.
- Data parsing and cleansing technique is applied to make the raw data into processing data.
- Usage of Ensemble of classifiers makes the model more robust and efficient.
- Ranking technique used in the project helps us to make efficient decisions.
- Creating a web application for user registrations and collection of data.
- The main objective is to obtain a better variety of crops that can be grown over the season. The proposed system would help to minimize the difficulties faced by farmers in choosing a crop and maximize the yield.
- The model predicts the crop yield by studying factors such as rainfall, temperature, area, pH value, soil type etc.

CHAPTER-2

LITERATURE SURVEY

A REVIEW ON DATA MINING TECHNIQUES FOR FERTILIZER RECOMMENDATION

Authors : Jignasha M. Jethva, Nikhil Gondaliya, Vinita Shah

To keep up nutrition levels in the soil in case of deficiency, fertilizers are added to soil. The standard issue existing among the Indian agriculturists choose approximate amount of fertilizers and add them manually. Excess or deficient extension of fertilizers can harm the plants life and reduce the yield. This paper gives overview of various data mining frameworks used on cultivating soil dataset for fertilizer recommendation.

A SURVEY ON DATA MINING TECHNIQUES IN AGRICULTURE

Authors: M.C.S.Geetha

Agriculture is the most critical application area especially in the developing nations like India. Use of information technology in agriculture can change the situation of decision making and farmers can yield in better way. This paper integrates the work of several authors in a single place so it is valuable for specialists to get data of current situation of data mining systems and applications in context to farming field.

AGRONUTRI ANDROID APPLICATION

Authors: S. Srija, R. Geetha Chanda, S.Lavanya, Dr. M. Kalpana Ph.D

This paper communicates the idea regarding the making of AgroNutri an android application that helps in conveying the harvest particular fertilizer amount to be applied. The idea is to calculate the measure of NPK composts to be applied depend on the blanked proposal of the crop of interest. This application works depend on the product chosen by the farmer and that is taken as input, thus providing the farmers. The future scope of the AgroNutri is that GPRS can be included so that according to location nutrients are suggested.

MACHINE LEARNING: APPLICATIONS IN INDIAN AGRICULTURE

Authors: Karandeep Kaur

Agriculture is a field that has been lacking from adaption of technologies and their advancements. Indian agriculturists should be up to the mark with the universal procedures. Machine learning is a native concept that can be applied to every field on all inputs and outputs. It has effectively settled its ability over ordinary calculations of software engineering and measurements. Machine learning calculations have improved the exactness of artificial intelligence machines including sensor based frameworks utilized in accuracy farming. This paper has evaluated the different uses of machine learning in the farming area. It additionally gives a knowledge into the inconveniences looked by Indian farmers and how they can be resolved using these procedures.

IMPACTS OF POPULATION GROWTH, ECONOMIC DEVELOPMENT, AND TECHNICAL CHANGE ON GLOBAL FOOD PRODUCTION AND CONSUMPTION

Authors: Uwe A. Schneider a,⇑, Petr Havlik b, Erwin Schmid c, Hugo Valin b, Aline Mosnier b,c, Michael Obersteiner b, Hannes Bottcher b, Rastislav Skalsky´ d, Juraj Balkovicˇ d, Timm Sauer a, Steffen Fritz b.

Throughout the following decades humanity will request more food from less land and water assets. This investigation evaluates the food production effects of four elective advancement situations

from the Millennium Ecosystem Assessment and the Special Report on Emission Scenarios. partialy and jointly considered are land and water supply impacts from population development, and specialized change, and forests and agriculture demand request shifts from population development and economic improvement. The income impacts on nourishment request are registered with dynamic flexibilities. Worldwide farming area increments by up to 14% somewhere in the range of 2010 and 2030.Deforestation restrictions strongly impact the price of land and water resources but have little consequences for the global level of food production and food prices. While projected income changes have the highest partial impact on per capita food consumption levels, population growth leads to the highest increase in total food production. The impact of technical change is amplified or mitigated by adaptations of land management intensities.

A SMART AGRICULTURAL MODEL BY INTEGRATING IOT, MOBILE AND CLOUD-BASED BIG DATA ANALYTICS

Authors: S.Rajeswari, K.Suthendran, K.Rajkumar.

In the cultivating field, the system models play a significant role to the enhancement of the agro-normal and money related conditions. In the proportions of benefits of the field and farm examinations to give the information and to recognize fitting and fruitful organization practices. It can recognize the organization to arrive managers and transversely over reality as long as the required soil, the board, environment, and money related information. Decision Support Systems (DSSs) use to make the information for the vermin the board, develop the officials. These systems are not using the impelled strategies to process the data. Thusly, use the adroit system thoughts to take the decisions for the issue. It expects a crucial activity in the comprehension of agronomic results, and their use as decision sincerely steady systems for farmers is extending.

AN OVERVIEW OF INTERNET OF THINGS AND DATA ANALYTICS IN AGRICULTURE: BENEFITS AND CHALLENGES

Authors: Olakunle Elijah, Tharek Abdul Rahman, Igbafe Orikumhi, Chee Yen Leow, Nour Hindia.

A blueprint of Iot and DA in agriculture has been shown in this paper. A couple of zones related to the association of Iot in agribusiness have been discussed in detail. The investigation of composing exhibits that there are clusters of work advancing being produced of Iot development that can be used to increase operational efficiency and gainfulness of plant and creatures. The benefits of Iot and DA, and open troubles have been identified and inspected in this paper. Iot is depended upon to offer a couple of benefits to the agribusiness division. Regardless, there are up 'til now different issues to be steered to make it moderate for close to nothing and medium-scale farmers. The key issues are security and cost. It is typical that as contention increases in the cultivating part.

MACHINE LEARNING FACILITATED RICE PREDICTION IN BANGLADESH

Authors: Mohammad Motiur Rahman, Naheena Haq, Rashedur M Rahman

In this examination, self-organising map (SOM) was utilized to group the information relationship between the information factors. After that chi-square test strategy was utilized to set up the level of reliance between the related variable qualities. It was discovered that the day by day outrageous climate conditions, for example, most extreme and least fluctuation in temperature, precipitation, dampness and wind speed were the principle drivers of product development, yield and wine quality.

SUPPORT VECTOR MACHINE-BASED CLASSIFICATION SCHEME OF MAIZE CROP
Authors: Suhas S Athani, CH Tejeshwar

This paper says about, advancement of a mechanized framework to distinguish and group weeds from the products would be of extraordinary help and we have proposed a set-up that decreases labour. We have considered pictures of maize edits as the informational index. Separating surface highlights of the picture and applying SVM (support vector machine) to arrange whether the given picture is a weed or a yield brought about a precision of 82%. The proposed framework gives a chance to investigate more about element extraction methods.

2.1 EXISTING SYSTEM

The computational and data demands of structural price forecasting generally far exceed than what is routinely available in developing countries. Consequently, researchers often rely on parsimonious representations of price processes for their forecasting needs. Contemporary parsimonious form of price forecasting relies heavily on time series modelling. In time series modelling, past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. During the past few decades, much effort has been devoted to the development and improvement of time series forecasting models. Time series modelling requires less onerous data input for regular and up-to date price forecasting. Hence there is a need for better classification which would be an ensemble or hybrid classification model.

2.2 DISADVANTAGES OF EXISTING SYSTEM

- Efficiency is low.
- The existing system which recommends crop yield is either hardware-based being costly to maintain, or not easily accessible.
- Despite many solutions that have been recently proposed, there are still open challenges in creating a user-friendly application with respect to crop recommendation.
- More number of repeated work.

2.3 PROPOSED SYSTEM

In proposed system, the data analysis technology is used to update the crop yield rate change. The concept of this project is to implement the crop selection method so that this method helps in solving many agriculture and farmers problems. This improves our Indian economy by maximizing the yield rate of crop production. Different types of land condition. So the quality of the crops are identified using ranking process. By this process the rate of the low quality and high quality crop is also intimated. The usage of ensemble of classifiers paves a path way to make a better decision on predictions due to the usage of multiple classifiers. Further, a ranking process is applied for decision making in order to select the classifiers results. This system is used to predict the cost of the fertilizers for further. This project uses Ensemble of classifiers such as Decision tree and Random forest classifier. In addition, this project uses Ranking technique.

2.4 ADVANTAGES OF PROPOSED SYSTEM

- Useful to people far away from towns/cities.
- Better time efficiency.
- Reduction of repeated work.

CHAPTER-3

AIM AND SCOPE OF THE PRESENT INVESTIGATION

3.1 AIM:

Our Aim from the project is to make a ML model which takes student data trains itself using various Machine Learning techniques and Algorithms (Random Forest, Decision Tree) and predict the yield and best fertilizer that suits for the crops in virtual environment by considering the overall factors that contribute in his overall yield.

Secondly, to learn the required tech stacks and use it to make model with an python application and lastly to execute it get output about yield and best fertilizer for the crop.

3.2 SCOPE:

This Project can be used to get the student performance with more accuracy than any other model published earlier and we can also make some mobile or web application based on the model.

Identifying the challenges farmers face in selecting suitable crops based on environmental, economic, and regional factors. Exploring the need for a system that reduces dependency on traditional knowledge and provides scientific, data-driven recommendations. Expanding the system to include pest and disease management recommendations. Incorporating precision agriculture techniques using IoT devices and remote sensing. Developing global or region-specific models for diverse agricultural conditions.

CHAPTER-4

EXPERIMENTAL OR MATERIALS AND METHODS ALGORITHMS USED.

4.1 RANDOM FOREST ALGORITHM:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps:

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points.
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

4.2 DECISION TREE:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
The complete process can be better understood using the below algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
Step-3: Divide the S into subsets that contains possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in step 3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

4.3 SYSTEM ARCHITECTURE

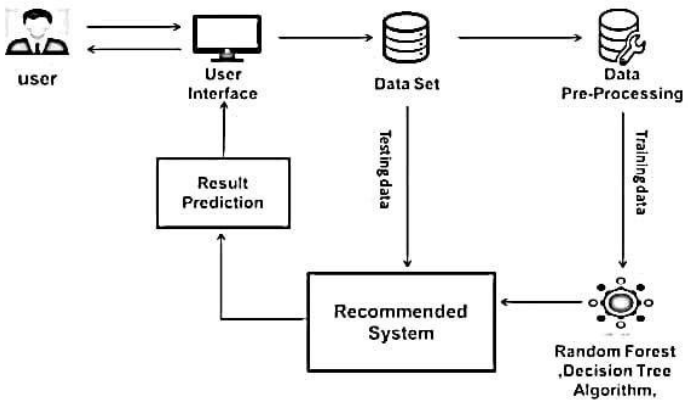

Fig 4.1: System Architecture

4.4 SYSTEM REQUIREMENTS:

4.4.1 HARDWARE REQUIREMENTS:

- System: Pentium Dual Core. ● Hard Disk: 40 GB.
- Monitor: 15'' LED

- Ram: 4 GB

4.4.2 SOFTWARE REQUIREMENTS:

- Operating system: Windows 7/10.
- Coding Language: Python.

4.5 SOFTWARE ENVIRONMENT

Python:

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

- Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

- Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

- Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

History of Python:

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Small Talk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

Python Features:

Python's features include –

- Easy-to-learn − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

- Easy-to-read − Python code is more clearly defined and visible to the eyes.
- Easy-to-maintain − Python's source code is fairly easy-to-maintain.
- A broad standard library − Python's bulk of the library is very portable and cross platform compatible on UNIX, Windows, and Macintosh.
- Interactive Mode − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- Portable − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- Extendable − You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- Databases − Python provides interfaces to all major commercial databases.
- GUI Programming − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- Scalable − Python provides a better structure and support for large programs than shell scripting.

4.6 MODULES:

- Metadata
- Data Pre-processing
- Crop Prediction Module
- Crop Recommendation Module

MODULES DESCRIPTION:

Metadata:

All the main data used in the data set are initialized with the number to use in the algorithm it is like initializing all the details. In this metadata, we are going to initialize all the parameters. This data makes us use the data easily in the algorithm. **Data Pre-processing:**

Hear the raw data in the crop data is cleaned and the metadata is appending to it by removing the things which are converted to the integer. So, the data is easy to train. Hear all the data. In this preprocessing, we first load the metadata into this and then this metadata will be attached to the data and replace the converted data with metadata. Then this data will be moved further and remove the unwanted data in the list and it will divide the data into the train and the test data.

Crop Prediction Module:

The obtained result will be helpful for the farmers to know the Yield of the crop so, he can go for the better crop which gives high yield and also say them the efficient use of agriculture field. This way we can help the farmers to grow the crop which gives them better yield.

Crop Recommendation Module:

In this module, we have proposed a model that addresses these issues. The novelty of the proposed system is to guide the farmers to maximize the crop yield as well as suggest the most profitable crop for the specific region.

4.7 UML DIAGRAMS:

UML is simply anther graphical representation of a common semantic model. UML provides a comprehensive notation for the full lifecycle of object-oriented development.

ADVANTAGES

- To represent complete systems using object oriented concepts
- To establish an explicit coupling between concepts and executable code
- To take into account the scaling factors that are inherent to complex and Critical end.
- To creating a modelling language usable by both humans and machines

UML defines several models for representing systems

- The class model captures the static structure
- The state model expresses the dynamic behaviour of objects
- The use case model describes the requirements of the user
- The interaction model represents the scenarios and messages flows
- The implementation model shows the work units
- The deployment model provides details that pertain to process.

4.7.1 DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

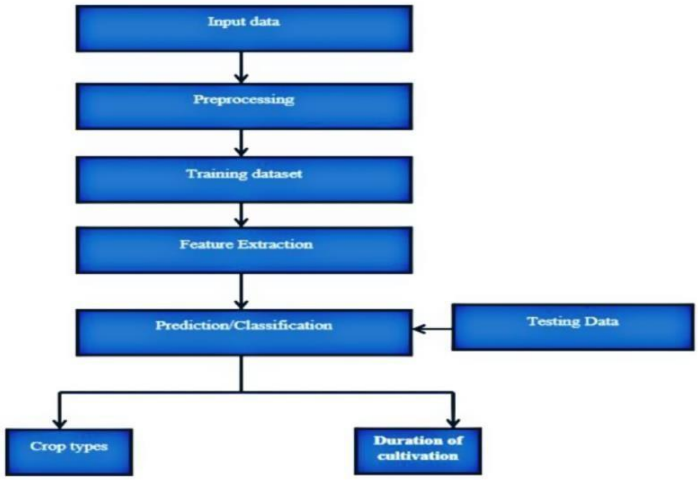

Fig4.2: Data Flow Diagram

4.7.2 USECASE DIAGRAM:

Use case diagrams overview the usage requirement for system. They are useful for presentations to management and/or project stakeholders, but for actual development you will find that use cases provide significantly more value because they describe "the meant" of the actual requirements. A use case describes a sequence of action that provides something of measurable value to an action and is drawn as a horizontal ellipse.

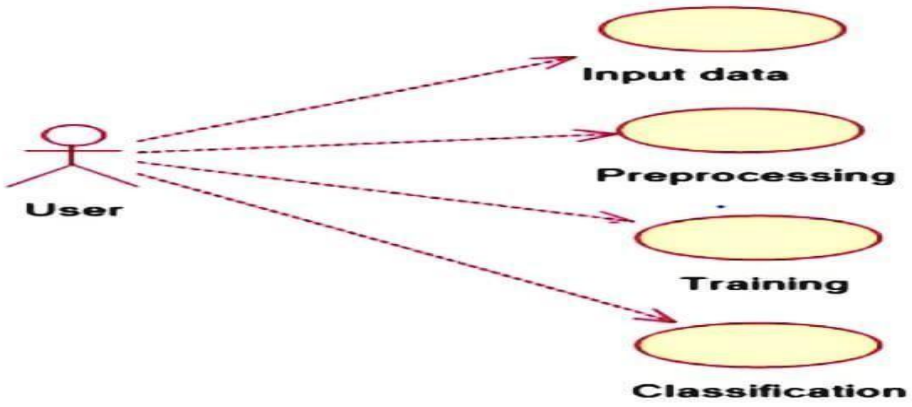

Fig 4.3: Use Case Diagram

4.7.3 ACTIVITY DIAGRAM:

Activity diagram are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. The activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. Activity diagram consist of Initial node, activity final node and activities in between.

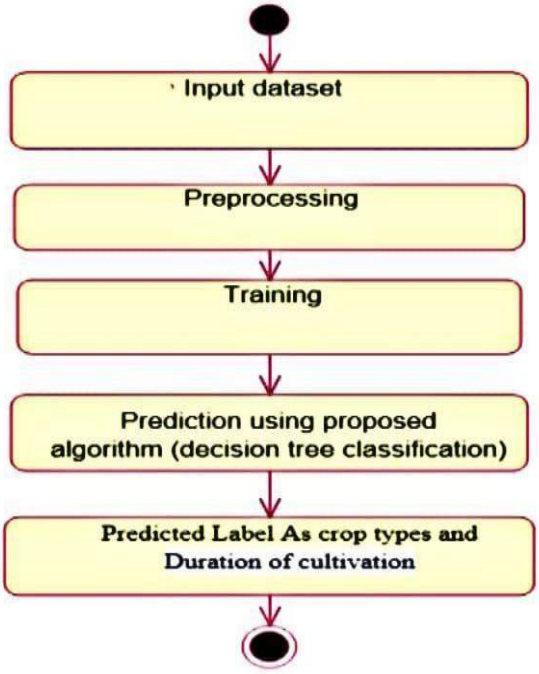

Fig4.4: Activity Diagram

4.8 INPUT DESIGN AND OUTPUT DESIGN:

4.8.1 INPUT DESIGN:

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?
How the data should be arranged or coded?
The dialog to guide the operating personnel in providing input.
Methods for preparing input validations and steps to follow when error occur.

4.8.2 OUTPUT DESIGN:

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

4.9 SYSTEM STUDY:

FEASIBILITY STUDY:

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.
Three key considerations involved in the feasibility analysis are

- ❖ ECONOMICAL FEASIBILITY
- ❖ TECHNICAL FEASIBILITY
- ❖ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TESTING:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS:

Unit testing:

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path

of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test:

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centred on the following items:

Valid Input : identified classes of valid input must be accepted.
Invalid Input : identified classes of invalid input must be rejected.
Functions : identified functions must be exercised.
Output : identified classes of application outputs must be exercised.
Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test:

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing:

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose.
It is used to test areas that cannot be reached from a black box level.

Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives:

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested:

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page

Integration Testing:

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results:

All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results:

All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER-5

RESULTS AND DISCUSSIONS

5.1 RESULTS

The result of the classification model prepared using Python may be divided into three parts in itself:

I. Pre-model Data Visualization
II. Model Information Visualization
III. Post-model Data Visualization

5.1.1 PRE-MODEL DATA VISUALIZATION

The pre-model data visualization section includes all data visualizations, such as visualizations created with the pure data set to understand the features and distribution of data in the data set.

Visualization:

Plots are essentially one-way circuits. It is also a method for determining the appropriation of each trademark in a size. The Density Plot is a simple histogram detail. Thickness addresses an unsteady measurement of dissemination. It only recognizes passages as a numerical list.

A density plot is used to track the distribution of at least one factor. The main thing to do when recovering new information is to independently check the dissemination of the factors. It provides a wealth of information.

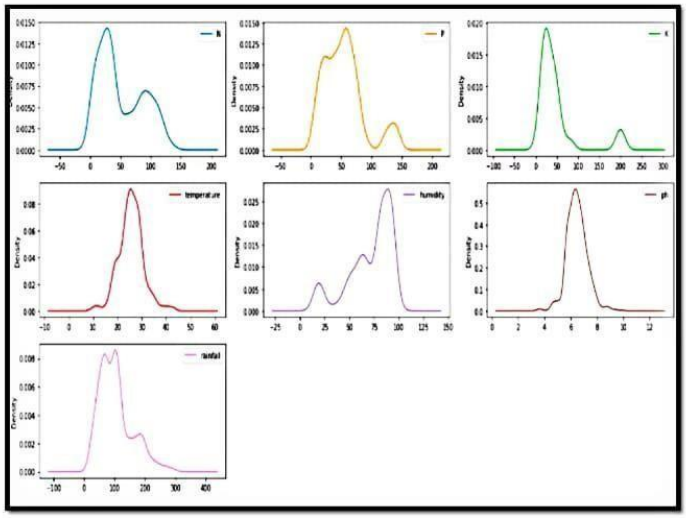

Fig 5.1: Density Diagram

The above figure shows a comparison between density and the seven parameters i.e., Nitrogen(N), Potassium(K), Phosphorus(P), Temperature, Humidity, Ph and rainfall.

5.1.2 MODEL INFORMATION VISUALIZATION

In this part of the chapter, the visualizations with respect to the model will be discussed.

Confusion Matrix:

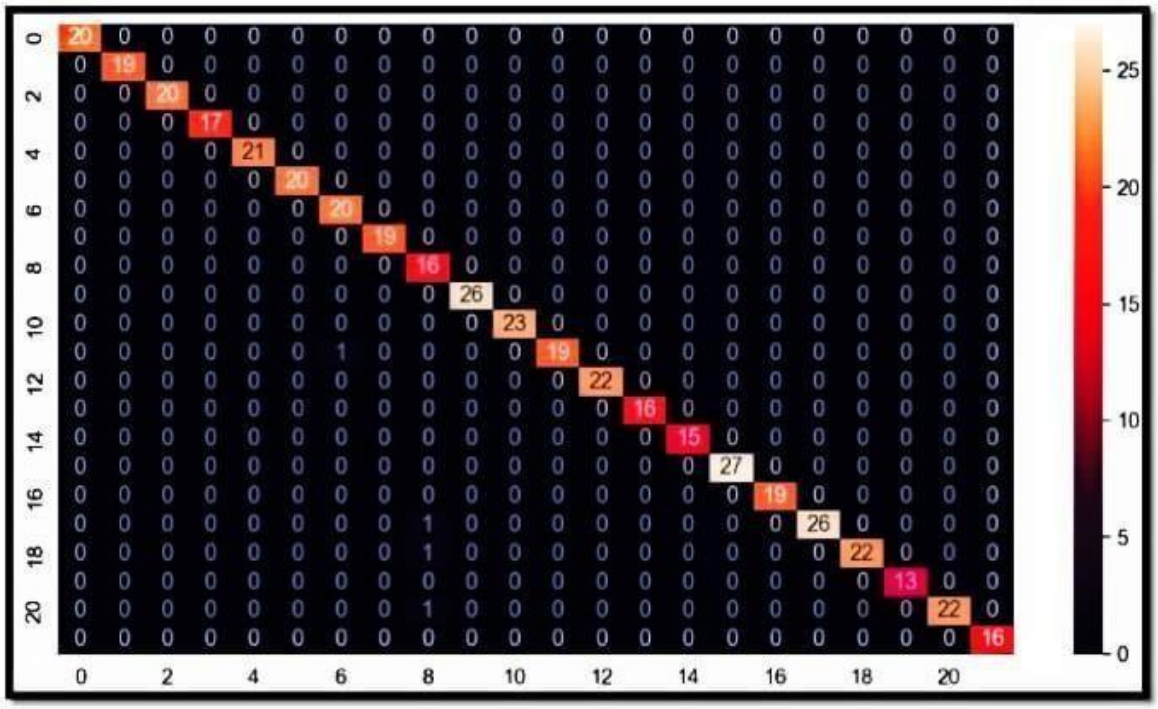

Fig 5.2: Confusion Matrix for Random Forest

The graph above shows the correct and incorrect classifications. It misclassified 22 times out of the 22 instances of training data used. This is clearly due to the bias that was previously mentioned. There have been no other misclassifications, demonstrating the model's integrity and reliability.

Correlation:

Correlation between attributes allows us to determine how strongly or weakly they are related to one another. The positive relationship between the variables is represented by the numeric value 1. The darker the colour of the numeric values 0, the more negative the relationship between the variables.

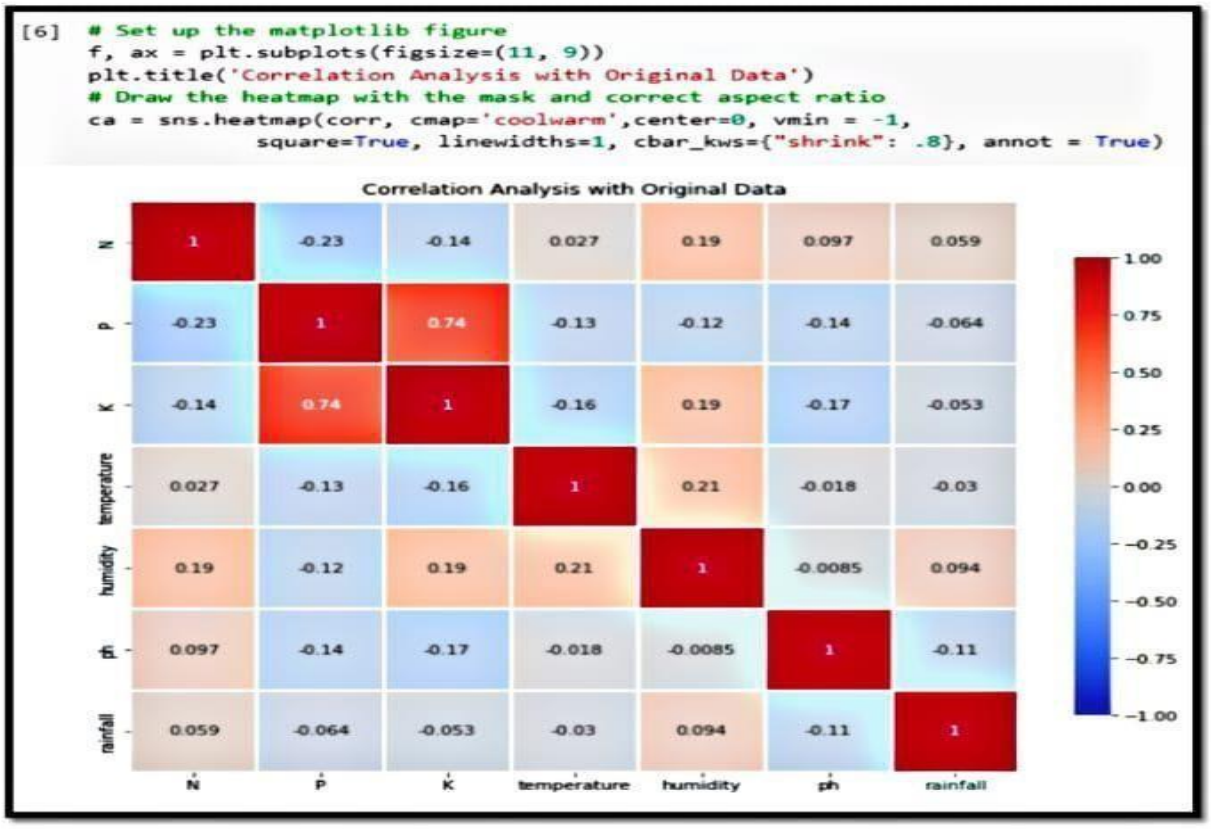

Fig 5.3: Correlation Diagram

Correlation Matrix Plot:

A correlation plot matrix can be formed for a collection of variables with each other variables will be plotted against each other. Here have seven columns where normally distributed with random values and column names are: Nitrogen(N), Potassium(K), Phosphorus(P), Temperature, Humidity, Ph and rainfall.

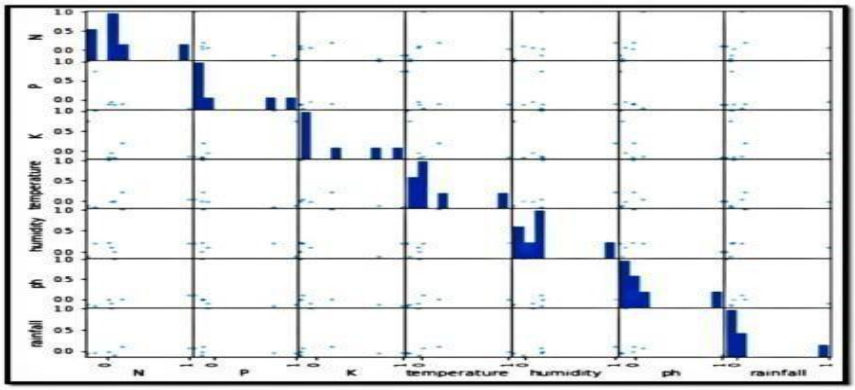

Fig 5.4: Correlation Matrix Plot Diagram

Classification Report:

- **Accuracy**

$$ACCURACY = \frac{NUMBER\ OF\ CORRECT\ PREDICTIONS}{TOTAL\ NUMBER\ OF\ PREDICTIONS}$$

Accuracy is a classification problem metric that indicates the percentage of correct predictions. We compute it by dividing the total number of predictions by the number of correct predictions. This formula provides a simple definition based on a binary classification problem. (In the second part of this article, we discuss multiclass and multilabel problems.) In the case of binary classification, accuracy can be expressed as True/False Positive/Negative values.

- **Precision:**

$$PRECISION = \frac{TRUE\ POSITIVE}{TOTAL\ POSITIVE + FALSE\ POSITIVE}$$

$$PRECISION = \frac{TRUE\ POSITIVE}{TOTAL\ PREDICTED\ POSITIVE}$$

Precision is defined as the fraction of positive examples that are actually positive among all positive examples predicted by us. It can also be defined as the number of true positives divided by the total number of true positives plus false positives. False positives occur when the model incorrectly labels something as positive when it is actually negative, or in our case, when the model incorrectly labels someone as a terrorist when they are not.

- **Recall:**

$$RECALL = \frac{TRUE\ POSITIVE}{TOTAL\ POSITIVE + FALSE\ NEGATIVE}$$

$$RECALL = \frac{TRUE\ POSITIVE}{TOTAL\ ACTUAL\ POSITIVE}$$

In statistics, the metric our intuition tells us we should maximize is known as recall, or a model's ability to find all relevant cases within a dataset. The number of true positives divided by the number of true positives plus the number of false negatives is the precise definition of recall. True positives are data points classified as positive by the model that are actually positive (meaning they are correct), whereas false negatives are data points classified as negative by the model that are actually positive (meaning they are correct) (incorrect).

- **F1-Score:**

$$F1 - SCORE = 2 * \frac{PRECISION * RECALL}{PRECISION + RECALL}$$

It is traditionally defined as the harmonic mean of precision and recall. It's also known as the F Score or the F Measure. In other words, the F1 score conveys the balance between precision and recall. It is thought to be a better measure than Precision and Recall separately because the trade-off between the two is difficult to achieve.

	Random Forest	KNN
Accuracy	0.995	0.988
Precision	0.996	0.996
Recall	0.995	0.989
F1-score	0.995	0.989

Table 5.1: Algorithm Comparison

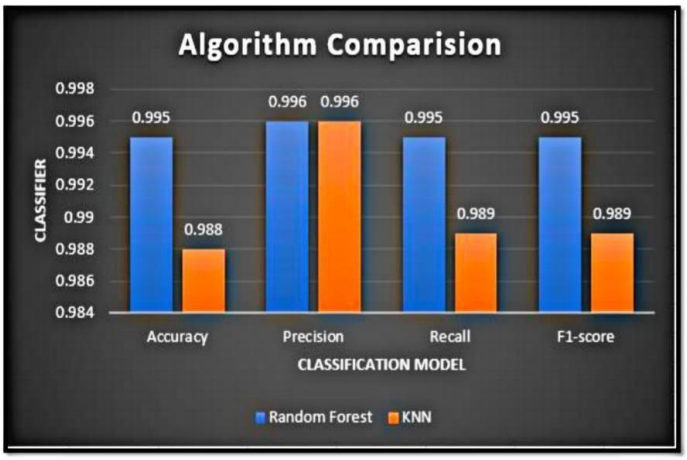

Fig 5.5.: Algorithm Comparison

The above figure represents all the necessary factors taken to determine whether a model is good or bad. They are technically known as Model Evaluation Metrics.

5.1.3 POST-MODEL DATA VISUALIZATION

Data visualisation is the depiction of interpreting data by displaying it in a graphical environment, allowing for the detection and exposure of patterns, inclinations, and connections.

The most popular charting libraries are:

- Matplotlib: Low-level, gives the user a lot of freedom.
- Pandas Visualization: This border is easy to use. Matplotlib can be used to create it.

- Plotly: Allows you to make an interactive plot for visualisation.

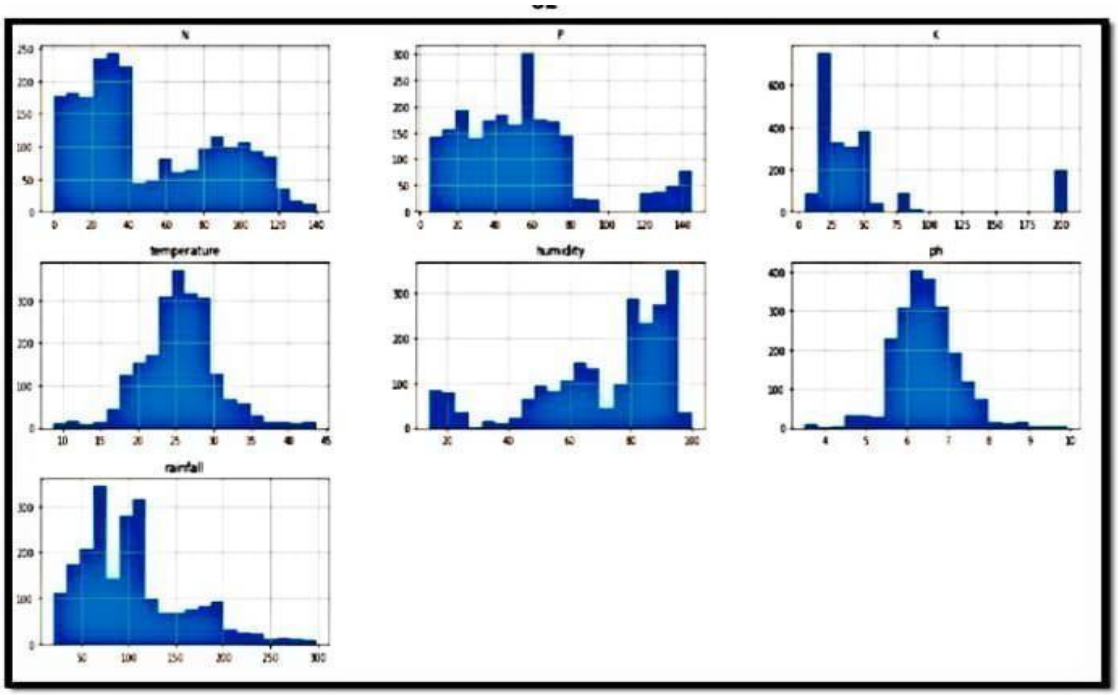

Fig 5.6: Hist Diagram

Phosphorus(P), Potassium(K) and rainfall is easy to notice the distribution is skewed very much to left. The attribute humidity is easy to notice the distribution is skewed very much to right. The attribute ph. has a Gaussian or nearly The attributes Nitrogen(N) and temperature may have an exponential distribution. The attribute Gaussian distribution.

5.2 ANALYSIS

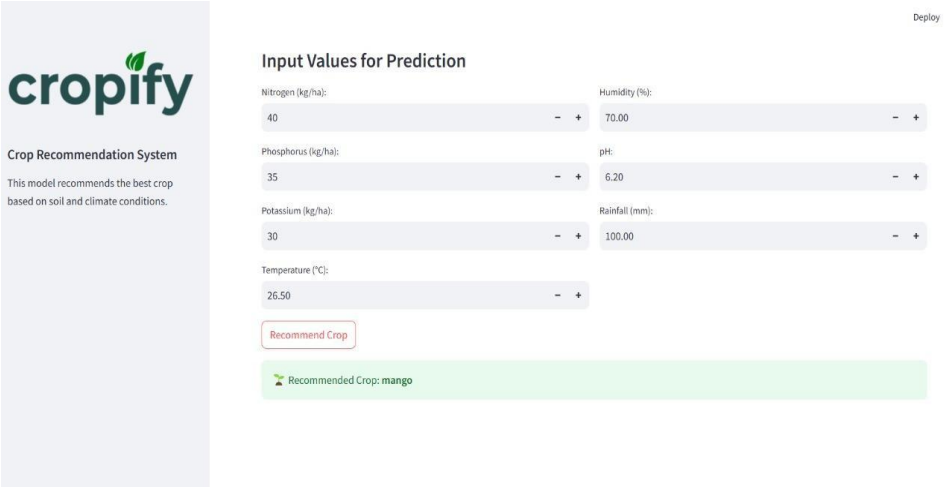

Fig 5.7: Crop Prediction

The above figure shows to discover, we will utilize seven unique fixings: nitrogen (N), phosphorus (P), potassium (K), heat, temperature, pH, and precipitation. Presently you want to get a decent gather to get a decent reap.

CHAPTER-6
CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

Machine learning techniques are applied in various fields of agriculture. Implement a system to predict crop production from the collection of past data. Carry out an arrangement to conjecture generally upheld rural creation. Plant creation is gotten ready for data handling strategies. Today, random forest is utilized to characterize a fascinating yield as a genuine harvest. In agribusiness, it is for the most part consistent with anticipate yields. The higher the yield, the higher the yield. The proposed procedures assist ranchers with being careful about various harvest necessities and costs. This assists ranchers with picking what harvests to plant. This work is frequently used to distinguish extra plants that can be gathered monetarily and proficiently. This innovation can grow a wide assortment of yields. Indian ranchers might enjoy the benefit of precisely foreseeing yields in various pieces of India.

6.2 FUTURE SCOPE

This research work can be enhancing to the high level by building a recommender system of agriculture production and distribution for farmer. India may be a country wherever agriculture is extremely vital. The prosperity of the farmers ends up in the prosperity of the state. Thus, our work would assist farmers in sowing the acceptable seed supported soil necessities so as to extend productivity and exploit such a way. As a result, farmers will plant the acceptable crop, increasing their yield and therefore the nation's overall productivity. Our future work can concentrate on associate degree improved knowledge set with an oversized variety of attributes.

REFERENCES

[1] Aruvansh Nigam, Saksham Garg, Archit Agrawal ―Crop Yield Prediction using ML Algorithms ―, 2019 Fifth International Conference on Image Information Processing (ICIIP)

[2] LeoBrieman, ―Random Forests‖, 2019 IEEE Transactions on Instrumentation and Measurement

[3] Priya, P., Muthaiah, U., Balamurugan, M.‖ Predicting Yield of the Crop Using Machine Learning Algorithm‖,2020 International Journal of Engineering Sciences & Research Technology (IJESRT)

[4] Mishra, S., Mishra, D., Santra, G. H.,―Applications of machine learning techniques in agricultural crop production‖,2020 Indian Journal of Science and Technology

[5] Dr.Y Jeevan Kumar,‖Supervised Learning Approach for Crop Production‖, 2020 5th International Conference on Communication and Electronics Systems (ICCES)

[6] RameshMedar,Vijay S, Shweta, ―Crop Yield Prediction using Machine Learning Techniques‖, 2018 International Journal of Engineering Sciences & Research Technology (IJESRT)

[7] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa,‖Machine Learning Methodologies for Paddy Yield Estimation in India: A CASE STUDY‖, IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium

[8] https://en.wikipedia.org/wiki/Crop_yield

[9] https://en.wikipedia.org/wiki/Random_forest

APPENDIX SOURCE CODE

```
import streamlit as st
import numpy as np
import pandas as pd
from PIL import Image
import joblib
import os

# Set page configuration
st.set_page_config(
    page_title="Crop Recommendation System",
    layout="wide",
    initial_sidebar_state="expanded",
)

# Define dataset & model paths
model_path = r"C:\Cropify-Crop-Recommendation-System-main (1)\Cropify-Crop-Recommendation-System-main\Model\RDF_model.pkl"

# Load trained model
try:
    rdf_clf = joblib.load(model_path)
except Exception as e:
    st.error(f"Error loading model: {e}")
    st.stop()

# Sidebar UI
with st.sidebar:
    try:
        image = Image.open('logo_cropify.png')
        st.image(image, width=300)
    except Exception:
        st.warning("Logo not found!")

    st.markdown("## Crop Recommendation System")
    st.write("This model recommends the best crop based on soil and climate conditions.")

# Feature inputs
st.markdown("### Input Values for Prediction")

col1, col2 = st.columns(2)
```

```
with col1:
    n_input = st.number_input('Nitrogen (kg/ha):', min_value=0, max_value=140, step=1)
    p_input = st.number_input('Phosphorus (kg/ha):', min_value=5, max_value=145, step=1)
    k_input = st.number_input('Potassium (kg/ha):', min_value=5, max_value=205, step=1)
    temp_input = st.number_input('Temperature (°C):', min_value=9.0, max_value=43.0, step=0.1, format="%.2f")

with col2:
    hum_input = st.number_input('Humidity (%):', min_value=15.0, max_value=99.0, step=0.1, format="%.2f")
    ph_input = st.number_input('pH:', min_value=3.6, max_value=9.9, step=0.1, format="%.2f")
    rain_input = st.number_input('Rainfall (mm):', min_value=21.0, max_value=298.0, step=0.1, format="%.2f")

# Pass only 7 features to the model
predict_inputs = [[n_input, p_input, k_input, temp_input, hum_input, ph_input, rain_input]]

# Prediction button
if st.button('Recommend Crop'):
    try:
        prediction = rdf_clf.predict(predict_inputs)[0]
        st.success(f"🌱 Recommended Crop: **{prediction}**")
    except Exception as e:
        st.error(f"Prediction Error: {e}")
import os
import joblib
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

# Define dataset path
dataset_path = r"C:\Cropify-Crop-Recommendation-System-main (1)\Cropify-Crop-Recommendation-System-main\Dataset\Crop_recommendation.csv"

# Verify if dataset exists
if not os.path.exists(dataset_path):
    raise FileNotFoundError(f"❌ Dataset not found at {dataset_path}. Please check the path.")
```

```
# Load dataset
df = pd.read_csv(dataset_path, encoding='utf-8')

# Select only the required 7 features
required_features = ["N", "P", "K", "temperature", "humidity", "ph", "rainfall"]
X = df[required_features]  # Use only the required columns
y = df["label"]

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a new Random Forest model
rdf_clf = RandomForestClassifier(n_estimators=100, random_state=42)
rdf_clf.fit(X_train, y_train)

# Save the trained model
model_path = r"C:\Cropify-Crop-Recommendation-System-main (1)\Cropify-Crop-Recommendation-System-main\Model\RDF_model.pkl"
joblib.dump(rdf_clf, model_path)

print("☑ Model retrained and saved successfully!")
```