

CROP YIELD PREDICTION USING MACHINE LEARNING WITH PYTHON

An Internship Training and Summer Project Report submitted to

**DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING**

in partial fulfilment of the requirements for the
award of the Degree of

BACHELOR OF ENGINEERING

Submitted by

PANDI KAVYA (19101074)

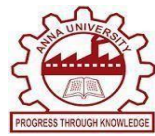
THATHIREDDY PUSHPALATHA (19101113)

YEGIREDDY DEEKSHITHA (19101124)

Under the guidance of

Mrs.B. ANANTHI, M.E., (Ph.D)., AP/CSE

Mr. A. SABARISHWAREN, MCA/ TC



**VIVEKANANDHA COLLEGE OF ENGINEERING FOR
WOMEN [Autonomous]**

Approved by AICTE, New Delhi and Accredited by NBA

(CSE, EEE, IT&BT)

Affiliated to Anna University, Chennai-25,

Elayampalayam, Tiruchengode, Namakkal Dt. – 637205

NOVEMBER & 2022



VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN

Elayampalayam, Tiruchengode, Namakkal Dt. – 637205

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Certificate

This is to certify that the Internship Training and Summer Project Report entitled "**CROP YIELD PREDICTION USING MACHINE LEARNING WITH PYTHON**", in partial fulfilment of the requirements for the award of the Degree of **BACHELOR OF ENGINEERING** is a record of original training undergone by **PANDI KAVYA(19101074),THATHIREDDY PUSHPALATHA (19101113), YEGIREDDY DEEKSHITHA(19101124)**, during the year 2022 of her study in the Department of **COMPUTER SCIENCE AND ENGINEERING, VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN** under my supervision and the report has not formed the basis for the award of any Degree/Fellowship or other similar title to any candidate of any University.

Place:

Signature of Guide

Date:

Mrs.B. ANANTHI, M.E., (Ph.D), AP/CSE

Countersigned


Head of the Department



Dr. C. POONGODI, M.E., Ph.D.,


Submitted to the Department of **COMPUTER SCIENCE AND ENGINEERING, VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN** for the examination held on _____

INTERNAL EXAMINER

COMPANY CERTIFICATES


**ATS ACCENT TECHNO SOFT**
Quality Matters...


An ISO 9001:2015 Certified



This is to certify that Mr. /Ms PANDI KAVYA (Reg.No 19101074)

Department of BE. CSE, Vivekanandha College of Engineering for Women, has successfully completed Internship on “Python with Machine Learning” and gained hands-on experience from 07.07.2022 to 06.08.2022 conducted by ATS.


Technical Head

info@accenttechnosoft.com 203, Nehru Street, Ram Nagar, Coimbatore - 641009. 0422 421 22 32 www.accenttechnosoft.com



This is to certify that Mr. /Ms **Thathireddy Pushpalatha (Reg.No : 19101113)**
Department of B.E CSE, Vivekanandha College of Engineering for Women, has successfully
completed Internship on “ **Python with Machine Learning** ” and gained
hands-on experience from 07-07-2022 to 06-08-2022 conducted by ATS.



Technical Head



This is to certify that Mr. /Ms PANDI KAVYA (Reg.No 19101074)
Department of BE. CSE, Vivekanandha College of Engineering for Women, has successfully
completed Internship on “ Python with Machine Learning ” and gained
hands-on experience from 07.07.2022 to 06.08.2022 conducted by ATS.

A handwritten signature in black ink, appearing to be "H. S. S.", written in a cursive style.

Technical Head

DECLARATION

We, **PANDI KAVYA, THATHIREDDY PUSHPALATHA, YEGIREDDY DEEKSHITHA** hereby declare that the Internship Training and Summer Project Report, entitled **“CROP YIELD PREDICTION USING MACHINE LEARNING WITH PYTHON”**, submitted to the **VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN** in partial fulfilment of the requirements for the award of the Degree of **BACHELOR OF ENGINEERING** is a record of original training undergone by me during the period **November & 2022** under the supervision and guidance of **Mrs.B. ANANTHI, AP/CSE**, Department of **COMPUTER SCIENCE AND ENGINEERING, VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN** and it has not formed the basis for the award of any Degree/ Fellowship or other similar title to any candidate of any University.

Place:

Signature of the Student

Date:

ACKNOWLEDGEMENT

We thank and praise the Lord Almighty for providing knowledge, Strength and all the necessary facilities to do this report successfully.

We are immensely grateful to our Honourable **Chairman and Secretary, “Vidya Rathna” Professor Dr. M. KARUNANITHI, B.Pharm., M.S., Ph.D., D.Litt.,** Vivekanandha Educational Institutions, who is our inspiration.

We are extremely grateful to our beloved Executive Director **Professor Dr. S.KUPPUSWAMY, B.E., M.Sc.(ENGG), Dr. Ing(France),** for his motivation and guidance of our report preparation.

We wish to express our profound thanks to our beloved **Principal, Dr.KCK.VIJAYAKUMAR, M.E., Ph.D., MIE,** for all the facilities and support provided during the period of our project successfully.

We would like to acknowledge our **Head of the Department Dr. C. POONGODI, M.E., Ph.D.,** Department of Computer Science and Engineering, for her encouragement and support for completing the report successfully.

We wish to thank our Internship Training and Summer Project Report Coordinators **Ms.S.SINDHUJA., M.E., AP/CSE** and **Mr.B.S.PRAKASH., M.E., (Ph.D) AP/CSE** for their kind support and guidance in completion of our report.

We wish to thank our project **Faculty Guide Mrs.B. ANANTHI, M.E., (Ph.D) AP/CSE** for her kind support and guidance in completion of our report.

We extend our sincere thanks to, Technical Consultant of **Accent Techno Soft Mr. A. SABARISHWAREN, MCA** for his kind support throughout my work

We are thankful and fortunate enough to get constant encouragement, support and guidance for our parents and all Teaching and non-Teaching staff members of the department of Computer Science and Engineering who helped us in successfully completing our report.

ABSTRACT

The impact of climate change in India, most of the agricultural crops are being badly affected in terms of their performance over a period of the last two decades. Predicting the crop yield in advance of its harvest would help the policy makers and farmers for taking appropriate measures for marketing and storage. This project will help the farmers to know the yield of their crop before cultivating on to the agricultural field and thus help them to make the appropriate decisions. It attempts to solve the issue by building a prototype of a prediction system. Implementation of such a system with an easy-to-use web based graphic user interface and the machine learning algorithm will be carried out. The results of the prediction will be made available to the farmer. Thus, for such kind of data analytics in crop prediction, there are different techniques or algorithms, and the help of those algorithms we can predict crop yield. Random forest algorithm is used. By analysing all these issues and problems like weather, temperature, humidity, rainfall, moisture, the situation faced by us. In India, there are many ways to increase the economic growth in the field of agriculture. Data mining is also useful for predicting crop yield production. Generally, data mining is the process of analysing data from various viewpoint and summarizing it into important information. Random forest is the most powerful supervised learning algorithm capable of performing both classification and regression tasks that operated by constructing a multitude of decision trees during training time and generating output of the class that is the mode of the classes (classification) or mean prediction (regression) of the individual.

TABLE OF CONTENTS

| CHAPTER NO. | CONTENTS | PAGENO |
|-------------|-------------------------------|-----------|
| | CERTIFICATE | i |
| | COMPANY CERTIFICATE | ii |
| | DECLARATION | v |
| | ACKNOWLEDGEMENT | vi |
| | ABSTRACT | vii |
| | LIST OF FIGURES | x |
| 1 | INTRODUCTION | 1 |
| | 1.1 IMPORTANCE | 2 |
| | 1.2 PURPOSE | 3 |
| | 1.3 OBJECTIVE | 4 |
| | 1.4 SCOPE | 4 |
| 2 | COMPANY PROFILE | 5 |
| 3 | LITERATURE SURVEY | 6 |
| 4 | SYSTEM ANALYSIS | 10 |
| | 4.1 EXISTING SYSTEM | 10 |
| | 4.2 PROPOSED SYSTEM | 10 |
| | 4.2.1 DATA COLLECTION | 10 |
| | 4.2.2 PREPROCESSING THE DATA | 11 |
| | 4.2.3 TRANSFORMING THE DATA | 11 |
| 5 | SYSTEM REQUIRMENTS AND | |
| | SPECIFICATION | 12 |
| | 5.1 SOFTWARE REQUIREMENTS | 12 |

| | | |
|----------|---------------------------------------|-----------|
| | 5.2 HARDWARE REQUIREMENTS | 12 |
| | 5.3 FUNCTIONAL REQUIREMENTS | 14 |
| | 5.4 NON-FUNCTIONAL REQUIREMENTS | 14 |
| | 5.5 BASIC OPERATIONAL REQUIREMENTS | 19 |
| 6 | SYSTEM DESIGN | 25 |
| 7 | IMPLEMENTATION | 28 |
| | 7.1 LIBRARIES USED | 35 |
| | 7.1.1 NUMPY | 35 |
| | 7.1.2 PANDAS | 36 |
| | 7.1.3 SKLEARN | 38 |
| 8 | TESTING | 40 |
| | 8.1 FUNCTIONAL TESTING | 40 |
| | 8.2 NON-FUNCTIONAL TESTING | 41 |
| 9 | CONCLUSION AND FUTURE SCOPE | 42 |
| | APPENDICES | 45 |
| | BIBLIOGRAPHY | 49 |
| | REFERENCES | |

LIST OF FIGURES

| FIGURE NO | FIGURE NAME | PAGENO |
|-----------|------------------------------|--------|
| 1.1 | CROP | 2 |
| 3.1 | MODULAR DIAGRAM | 7 |
| 4.1 | STAGES OF PURPOSED SYSTEM | 11 |
| 5.1 | TYPES OF REQUIREMENTS IN SRS | 13 |
| 6.1 | SYSTEM DESIGN | 27 |
| 6.2 | SYSTEM DESIGN | 28 |
| 7.1 | BASIC PROCESS OF CROP YIELD | 31 |
| 7.2 | LINEAR REGRESSION | 34 |
| 7.3 | DECISION TREE | 35 |
| 7.4 | CLEANING DATA | 38 |
| 7.5 | ADVANCED | 39 |
| 8.1 | TESTING LEVELS | 43 |
| 10.1 | PRDICTED CROP IS MANGO | 46 |
| 10.2 | PREDICTED CROP IS MUSKMELON | 48 |

CHAPTER 1

INTRODUCTION

Agriculture is the backbone of the Indian economy. In India, agricultural yield primarily depends on weather conditions. Rice cultivation mainly depends on rainfall. Timely advice to predict the future crop productivity and an analysis is to be made in order to help the farmers to maximize the crop production of crops. Yield prediction is an important agricultural problem. In the past farmers used to predict their yield from previous year yield experiences. Thus, for this kind of data analytics in crop prediction, there are different techniques or algorithms, and with the help of those algorithms we can predict crop yield. Random forest algorithm is used. Using all these algorithms and with the help of inter-relation between them, there are growing range of applications and the role of big data analytics techniques in agriculture. Since the creation of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are concentrated on cultivating artificial products that are hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops at the right time and at the right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. By analysing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India, there are several ways to increase the economic growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining is also useful for predicting crop yield production. The main objectives area.

To use machine learning techniques

- a. To predict crop yield.
- b. To provide easy to use User Interface.
- c. To increase the accuracy of crop yield prediction.
- d. To analyse different climatic parameters (cloud cover, rainfall, temperature)



Figure 1.1 crop

1.1 IMPORTANCE

Crop yield prediction is an important agricultural problem. It becoming more important because of the growing concern about food security. Early crop yield prediction plays an important role in reducing famine by estimating the food availability for the growing world population. Hunger is one of the most devastating issues in the world and increasing crop yield production is a feasible solution to overcome this problem. The World Health Organization estimated that there is still an inadequate food supply for 820 million people around the world.

The target for the Sustainable Development Goals of the United Nations is to eliminate hunger, accomplish food security, and encourage sustainable agriculture by 2030. The Food and Agriculture Organization (FAO) estimated that there will be a 60 per cent demand for food to supply the world population of 9.3 billion by 2050. Therefore, crop yield prediction can offer crucial information required for developing a reasonable solution to achieve the target and end hunger.

Crop yield is influenced by various parameters, and it is difficult to build a reliable prediction model with traditional methods. However, with advancements in computational technology, the development and training of a novel approach for crop yield prediction have become feasible. Deep learning is a class of machine learning that is capable of learning from data that are unstructured and unlabelled, whereby the learning can be supervised, semi-supervised, or unsupervised. pointed out that deep learning techniques focus on learning abstract features of large datasets. To accurately predict crop yield requires primary knowledge of the association between functional attributes and interactive factors. To study such correlation requires both comprehensive datasets and high-efficiency algorithms, which can be achieved by using machine learning. The Agricultural yield primarily depends on weather conditions (rain, temperature, etc), pesticides. Accurate information about history

of crop yield is important for making decisions related to agricultural risk management and future predictions.

Factors effecting plant growth:

Rainfall

Soil fertility

Sunlight

Gases

Soil pH and water pH

Temperature

1.2 PURPOSE

It is been observed that farmers are facing the problem at the time of the yield of the crop because of the rapid changes in the weather where it effects the yield of the crop. Decrease the quality of the crop and which in turn provide less income to the farmers.

Crop yield simulations help to understand the cumulative effects of water and nutrient deficiencies, pests, diseases, the impact of crop yield variability, and other field conditions over the growing season.

Accurate yield prediction is pivotal to:

- 1) global food production, prices, and food security
- 2) Governments and companies make timely Import/Export decisions based on yield predictions.
- 3) Farmers and other stakeholders use the yield prediction to make informed financial and business decisions.

This project works on achieving the more quality of the crop that will help the farmers to gain more money. In this project we have collected the datasets of all the factors that are depends of the crops of several years. Using this data, the prediction is obtained to show that the harvest of the crop that is growth in that region.

1.3 OBJECTIVE

The main objectives are

- a. To use machine learning techniques to predict crop yield.
- b. To provide easy to use User Interface.
- c. To increase the accuracy of crop yield prediction.
- d. Efficient ways of production
- d. To analyse different climatic parameters (cloud cover, rainfall, temperature, humidity, soil type)

This project is based on the curbing the problem faced by the farmers as well as providing the accurate level of the harvest they can expect from the crop they have growth depending on the dependent factors like temperature, rainfall, etc. This project is mainly developed to help farmers so that this may help them in analysis of the harvest of the crop. Farmers are facing loses in the crop yield due to improper knowledge of the crop and the natural factor that are affecting them. In this project we analysis the factors and predict a graph that shows the crop's yield well before the harvest.

1.4 SCOPE

In the project, we introduce a scalable, accurate, and inexpensive method to predict crop yield using available climatic data and machine learning. Our machine learning approach can predict the crop yield with high spatial resolution (for particular regional). We have collected temperature, rainfall, crop yield and other datasets from various sources like Kaggle etc. Machine learning algorithms like Linear Regression, Decision Tree, SVM algorithm, Random Forest to predict crop yield based on factors like temperature, rainfall and pressure. The user interface gets the result of the various crop yield graph and displays it to the user.

CHAPTER 2

COMPANY PROFILE

ACCENT TECHNO SOFT (ATS)

Accent Techno Soft (ATS) provides a wide range of solutions in IT Consulting, technology and Operations space for the clients. To enhance the business value of the service offerings to the customers, they have formed strategic alliances with industry bodies, technologies vendors and system integrators. Through these partnerships they are able to deliver industry-best end –to-end solutions to the customers.

Accent specializes in the planning, analysis, and management of business, infrastructure and natural resources. ATS serve clients who share the belief that high-quality; objective analysis is a prerequisite to resolving complex problems.

Accent was founded by experienced software professionals and providing the foundation for the company's expertise in E-commerce/Web applications, custom application development, data warehousing, enterprise management solutions, and operations management (support, maintenance, implementation).

Accent Techno Soft, specialize in the business of software Training & HR Consultancy spotlighting in India.

Accent provides comprehensive and cost-effective training for individuals looking to inflate their IT skills in their current professions or looking to take the first step toward new careers. The success of the customers is realized through training sessions, but the foundation of Accent is based on inspiring students and companies to become more productive and successful in their daily activities

2.1 TECHNOLOGIES OFFERED BY ATS

- JAVA/J2EE
- Perl/Python
- .Net
- Web Designing
- PHP/MySQL
- Software Testing
- Hardware & Networking

CHAPTER 3

LITERATURE SURVEY

Literature Survey is a systematic and thorough search of all types of published literature as well as other sources including dissertation, these in order to identify as many items as possible that are relevant to a particular topic. Predicting agricultural products plays a very important role in agriculture. It helps in increasing net produce, better planning and gaining more profits. To achieve better results, we studied a few research papers related to our project topic.

3.1 Prediction of Crop Yield Using Machine Learning

Author: Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R L

Publication: International Research Journal of Engineering and Technology (IRJET)
Volume 05, Issue 02, Feb-2018

This paper states, most of the existing systems are hardware based which makes them expensive and difficult to maintain and lack to give accurate results. Some systems suggest crop sequence depending on yield rate and market price. In this paper, the system proposed tries to overcome these drawbacks and predicts crops by analysing structured data. Being a totally software solution, it does not allow maintenance factor to be considered much. Also the accuracy level would be high as compared to hardware based solutions, because components like soil composition, soil type, pH value, weather conditions all come into picture during the prediction process.

It can be achieved using unsupervised and supervised learning algorithms, like Kohonen Self Organizing Map (Kohonen's SOM) and BPN (Back Propagation Network). Dataset will then be trained by learning networks. It compares the accuracy obtained by different network learning techniques and the most accurate result will be delivered to the end user.

This paper proposes a system that will check soil quality and predict the crop yield accordingly along with it providing fertiliser recommendation if needed depending upon the quality of soil. The system takes inputs pH value and location from the user and result processing is done by two controllers. The result of the controller 1 and controller 2 are compared with a predefined —nutrients data store. These compared results are supplied to controller 3 wherein the combination of the above results and the predefined data set present

in the crop data store is compared. Finally, the results are displayed in the form of bar graphs along with accuracy percentage.

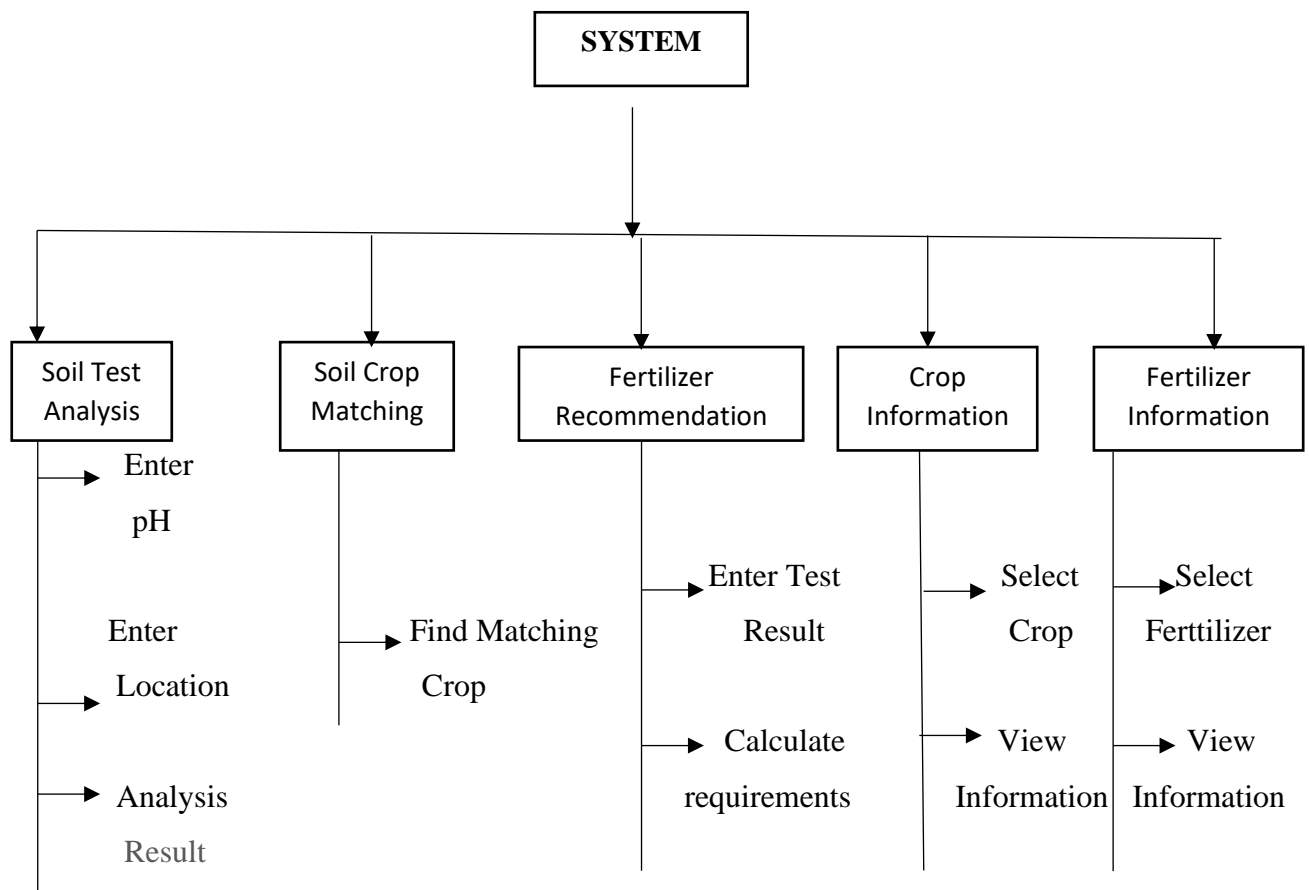


FIG 3.1 MODULAR DIAGRAM

The Fig 3.1, represents the architecture of the system. The system includes the following modules: soil test analysis, soil crop matching, fertilizer recommendation, crop information and fertilizer information. Each of the modules performs a specific function i.e., soil test analysis module on taking the ph and location as inputs analyses the soil and is categorized using the soil crop matching module. The fertilizer recommendation module analyses the test results obtained from the soil test analysis and soil crop matching modules and calculates the requirements. The crop information module on taking the crop as input

returns the information of the crop i.e., the production predicted. The fertilizer information module on taking the fertilizer as input returns the information of the fertilizer. The paper concludes that the system uses supervised and unsupervised Machine learning algorithms and gives best results based on accuracy. The results of the two algorithms will be compared and the one giving the best and accurate output will be selected. Thus the system will help reduce the difficulties faced by the farmers and stop them from attempting suicides. It will act as a medium to provide the farmers efficient information required to get high yield and thus maximize profits which in turn will reduce the suicide rates and lessen his difficulties. This paper includes the study of soil as well as fertilizers. The system proposed predicts the yield for the given crop. It also returns the information of the given fertilizer and also recommends one after calculating the requirements based on the soil properties and location. However, this paper does not take into consideration the climatic conditions of the location which have quite a large impact on the yield of the crops.

3.2 Predicting Yield of the Crop Using Machine Learning Algorithm

Author: P.Priya, U.Muthaiah & M.Balamurugan

Publication: International Journal of Engineering Sciences & Research Technology (IJESRT), April, 2018 This paper uses R programming with Machine Learning techniques. R is the leading tool for statistics, data analysis, and machine learning. It is more than a statistical package; it's a programming language, so you can create your own objects, functions, and packages. It's platform-independent, so it can be used on any operating system and it's free. R programs explicitly document the steps of our analysis and make it easy to reproduce and/or update analysis, which means it can quickly try many ideas and/or correct issues. All the datasets used in the research were sourced from the openly accessible records of the Indian Government. This was sourced for the years 1997 to 2013 for different seasons like Kharif and Rabi of rice production. From the vast initial dataset, only a limited number of important factors which have the highest impact on agricultural yield were selected for the present research. The dataset contains the following parameters: rainfall, season, and temperature and crop production. This paper also compares the two machine learning algorithms: decision trees and Random Forest.

Decision Tree: The Decision tree classifiers uses greedy approach hence an attribute chosen at the first step can't be used anymore which can give better classification if used in later steps. Also it over fits the training data which can give poor results for unseen data. So, to overcome this limitation ensemble model is used. In ensemble models results from different models are combined. The result obtained from an ensemble model is usually better than the result from any one of individual models.

Random Forest: Random Forests is an ensemble classifier which uses many decisions tree models to predict the result. A different subset of training data is selected, with replacement to train each tree. A collection of trees is a forest, and the trees are being trained on subsets which are being selected at random, hence random forests. This can be used for classification and regression problems. Class assignment is made by the number of votes from all the trees and for regression the average of the results is used. In this paper, the procedure they followed was as given below.

- Split the loaded data sets into two sets such as training data and test data in the split ratio of 67% and 33%.
- Then calculate Mean and Standard Deviation for needed tuples and then summarize the data sets. Compare the summarized data list and the original data sets & calculate the probability.
- Based on the result the largest probability produced is taken for prediction. The accuracy can be predicted by comparing the resultant class value with the test data set. The accuracy can range from 0% to 100%.

The paper concludes that the Results show that we can attain an accurate crop yield prediction using the Random Forest algorithm. The Random Forest algorithm achieves a largest number of crop yield models with the lowest models. It is suitable for massive crop yield prediction in agricultural planning. The dataset used for modelling here includes the climatic factors as well i.e., rainfall and temperature. The author did a comparative study of decision trees and random forest algorithms. But other algorithms were not considered and the dataset includes very few attributes that would not give accurate predictions.

CHAPTER 4

SYSTEM ANALYSIS

4.1 EXISTING SYSTEM

Existing system Other than blogging websites which provide information about the agriculture and agricultural accessories, there is no particular website for predicting the yield of the crop depending on the history in that specific geographical region

DISADVANTAGES OF EXISTING SYSTEM:

- The obtained result for the crop yield prediction using SMO classifier gives less accuracy when compared to naïve Bayes, multilayer perceptron and Bayesian network.
- Previously yield is predicted on the bases of the farmers prior experience but now weather conditions may change drastically so they cannot guess the yield.

4.2 PROPOSED SYSTEM

In the proposed system, supervised learning algorithms are used to form a model which will help us in providing is ready f choices of the most feasible crops that can be cultivated in that region along with its estimated yield. Two of the algorithms used here is Regression and Classification.

The main stages involved in the process are:

4.2.1 Dataset collection

4.2.2 Pre-processing the data

4.2.3 Transforming the data

4.2.1 Dataset collection

- The dataset used for this project is collected from various online sources like Kaggle.com and data.govt.in.
- We have taken the agricultural data of state names.
- Some important features or the parameters which has the highest impact on the agricultural yield considered in the project e are listed below:

Rainfall (in mm)

Humidity

Temperature

Area

Yield

Type of the soil

Location

Price

4.2.2 Pre-processing the data

- After the selection of the dataset, it has to be pre-processed into a form that you can work with. Some of the steps are formatting, cleaning and sampling.
- Initially the data have selected is converted into the format suitable for you to work with.
- Cleaning data is the removal or fixing of the mixed data. Sampling is taking a small representative sample of the selected data that may be much faster for exploring the solutions than electing the hole dataset.

4.2.3 Transforming the data

- The final step is transforming the selected data. The pre-processed data here is then transformed into data that is ready for machine learning algorithms by using various engineering features like scaling, feature aggregation and so on.
- There may be several features that can be combined into a single feature which would be more meaningful to the problem you are trying to solve.

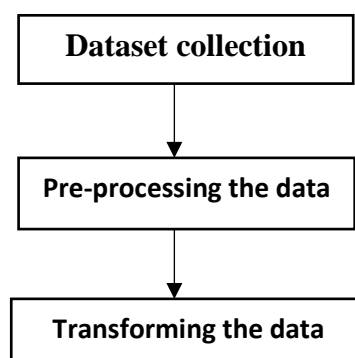


Fig 4.1 Stages of proposed system

CHAPTER 5

SYSTEM REQUIREMENTS SPECIFICATION

5.1 Software Requirements

- Operating system: Windows 7 and above
- Coding Language: Python
- Tools: Spyder/Colab/Visual Studio

5.2 Hardware Requirements

- Processors: Pentium IV and above
- Processor Speed: 3.00 GHZ
- RAM: 8 GB
- Ethernet /WIFI
- Hard Drive
- Storage: 20 GB

A software requirements specification (SRS) is a description of a software system to be developed. It lays out functional and non-functional requirements, and may include a set of use cases that describe user interactions that the software must provide. It is very important in an SRS to list out the requirements and how to meet them. It helps the team to save upon their time as they are able to comprehend how are going to go about the project. Doing this also enables the team to find out about the limitations and risks early on. An SRS can also be defined as a detailed description of a software system to be developed with its functional and non-functional requirements. It may include the use cases of how the user is going to interact with the software system. The software requirement specification document is consistent with all necessary requirements required for project development. To develop the software system, we should have a clear understanding of Software system. To achieve this,

we need continuous communication with customers to gather all requirements. A good SRS defines how the Software System will interact with all internal modules, hardware, and communication with other programs and human user interactions with a wide range of real-life scenarios. It is very important that testers must be cleared with every detail specified in this document in order to avoid faults in test cases and its expected results.

Qualities of SRS

- Correct
- Unambiguous
- Complete
- Consistent
- Ranked for importance and/or stability
- Verifiable
- Modifiable
- Traceable



Fig:5.1. Types of Requirements in SRS

Some of the goals an SRS should achieve are to:

- Provide feedback to the customer, ensuring that the IT Company understands the issues the software system should solve and how to address those issues.
- Help to break a problem down into smaller components just by writing down the requirements
- Speed up the testing and validation processes.
- Facilitate reviews.

5.3 Functional Requirements

A Functional Requirement is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behaviour, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform. In software engineering and systems engineering, a Functional Requirement can range from the high-level abstract statement of the sender's necessity to detailed mathematical functional requirement specifications. Functional software requirements help you to capture the intended behaviour of the system.

Benefits of functional requirements:

- Helps to check whether the application is providing all the functionalities that were mentioned in the functional requirement of that application
- A functional requirement document helps to define the functionality of a system or one of its subsystems.
- Functional requirements along with requirement analysis help identify missing requirements. They help clearly define the expected system service and behaviour.
- Errors caught in the Functional requirement gathering stage are the cheapest to fix.
- Support user goals, tasks, or activities

5.4 Non-Functional Requirements

Non-functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviours. They may relate to emergent system properties such as reliability, response time and store occupancy. Non-functional requirements arise through the user needs, because of budget constraints, organizational

policies, the need for interoperability with other software and hardware systems or because of external factors such as: -

- Product Requirements
- Organizational Requirements
- User Requirements
- Basic Operational Requirements

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviours. This should be contrasted with functional requirements that define specific behaviour or functions. The plan for implementing non-functional requirements is detailed in the system architecture. Broadly, functional requirements define what a system is supposed to do and non-functional requirements define how a system is supposed to be. Functional requirements are usually in the form of system shall do requirement, an individual action of part of the system, perhaps explicitly in the sense of a mathematical function, a black box description input, output, process and control functional model or IPO Model. In contrast, non-functional requirements are in the form of system shall be requirement, an overall property of the system as a whole or of a particular aspect and not a specific function. The systems' overall properties commonly mark the difference between whether the development project has succeeded or failed.

Non-functional requirements include:

- **Response time** - The time the system takes to load and the time for responses on any action the user does.
- **Processing time** - How long is acceptable to perform key functions or export / import data?
- **Throughput**- The number of transactions the system needs to handle must be kept in mind.
- **Storage** - The amount of data to be stored for the system to function.
- **Growth Requirements** - as the system grows it will need more storage space to keep up with the efficiency.

- **Locations of operation** - Geographic location, connection requirements and the restrictions of a local network prevail.
- **Architectural Standards** - The standards needed for the system to work and sustain.

Product Requirements

- **Correctness:** It follows a well-defined set of procedures and rules to compute and also rigorous testing is performed to confirm the correctness of the data.
- **Ease of Use:** The front end is designed in such a way that it provides an interface which allows the user to interact in an easy manner.
- **Modularity:** The complete product is broken up into many modules and well- defined interfaces are developed to explore the benefit of flexibility of the product.
- **Robustness:** This software is being developed in such a way that the overall performance is optimized and the user can expect the results within a limited time with utmost relevancy and correctness

Non-functional requirements are also called the qualities of a system. These qualities can be divided into execution quality & evolution quality. Execution qualities are security & usability of the system which are observed during run time, whereas evolution quality involves test ability, maintainability, extensible or scalability.

Hardware Requirements:

The hardware requirements include the requirements specification of the physical computer resources for a system to work efficiently. The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system.

The Hardware Requirements are listed below:

1.Processor:

A processor is an integrated electronic circuit that performs the calculations the at run a computer. A processor performs arithmetical, logical, input/output (I/O) and other basic instructions that are passed from an operating system (OS). Most other processes are dependent on the operations of a processor. A minimum 1 GHz processor should be used,

although we would recommend S2GHz or more. A processor includes an arithmetical logic and control unit (CU), which measures capability in terms of the following:

- Ability to process instructions at a given time
- Maximum number of bits/instructions
- Relative clock speed



The proposed system requires a 2.4 GHz processor or higher.

2.Ethernet connection (LAN) OR a wireless adapter (Wi-Fi):

Wi-Fi is a family of radio technologies that is commonly used for the wireless local area networking (WLAN) of devices which is based around the IEEE 802.11 family of standards. Devices that can use Wi-Fi technologies include desktops and laptops, smartphones and tablets, TV 's and printers, digital audio players, digital cameras, cars and drones. Compatible devices can connect to each other over Wi- Fi through a wireless access point as well as to connected Ethernet devices and may use it to access the Internet. Such an access point (or hotspot) has a range of about 20 meters (66 feet) indoors and a greater range outdoors. Hotspot coverage can be as small as a single room with walls that block radio waves, or as large as many square kilometres achieved by using multiple overlapping access points.



3.Hard Drive:

A hard drive is an electro-mechanical data storage device that uses magnetic storage to store and retrieve digital information using one or more rigid rapidly rotating disks, commonly known as platters, coated with magnetic material. The platters are paired with magnetic heads, usually arranged on a moving actuator arm, which reads and writes data to the platter surfaces. Data is accessed in a random-access manner, meaning that individual blocks of data can be stored or retrieved in any order and not only sequentially. HDDs are a type of non-volatile storage, retaining stored data even when powered off. 32 GB or higher is recommended for the proposed system.



3.Memory (RAM):

Random-access memory (RAM) is a form of computer data storage that stores data and machine code currently being used. A random-access memory device allows data items to be read or written in almost the same amount of time irrespective of the physical location of data inside the memory. In today's technology, random-access memory takes the form of integrated chips. RAM is normally associated with volatile types of memory (such as DRAM modules), where stored information is lost if power is removed, although non-volatile RAM has also been developed. A minimum of 2 GB RAM is recommended for the proposed system.



5.5 Basic Operational Requirements

Operational requirement is the process of linking strategic goals and objectives to tactic goals and objectives. It describes milestones, conditions for success and explains how, or what portion of, a strategic plan will be put into operation during a given operational period, in the case of, a strategic plan will be put into operation during a given operational period, in the case of commercial application, a fiscal year or another given budgetary term. An operational plan is the basis for, and justification of an annual operating budget request. Therefore, a five-year strategic plan would typically require five operational plans funded by five operating budgets. Operational plans should establish the activities and budgets for each part of the organization for the next 1-3 years. They link the strategic plan with the activities the organization will deliver and the resources required to deliver them.

An operational plan draws directly from agency and program strategic plans to describe agency and program missions and goals, program objectives, and program activities. Like a strategic plan, an operational plan addresses four questions:

- Where are we now?
- Where do we want to be?
- How do we get there?

The customers are those that perform the eight primary functions of systems engineering, with special emphasis on the operator as the key customer. Operational requirements will define the basic need and, at a minimum, will be related to these following points:

Mission profile or scenario: It describes about the procedures used to accomplish mission objective. It also finds out the effectiveness or efficiency of the system.

Performance and related parameters: It point out the critical system parameters to accomplish the mission.

Utilization environments: It gives a brief outline of system usage. Finds out appropriate environments for effective system operation. **Operational life cycle:** It defines the system lifetime.

4.5 Software Requirements

- Operating system: Windows 7 and above
- Coding Language: Python
- Tools: Spyder/Colab/Visual Studio

Python

It is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options. Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers. Additionally, Python supports the use of modules and a package, which means that programs can be designed in a modular style and code can be reused across a variety of projects.

What can you do with python?

Some things include:

Data analysis and machine learning

Web development

Automation or scripting

Software testing and prototyping

Everyday Tasks



Data analysis and machine learning

Python has become a staple in data science, allowing data analysts and other professionals to use the language to conduct complex statistical calculations, create data visualizations, build machine learning algorithms, manipulate and analyse data, and complete other data-related tasks

Web development

python is often used to develop the back end of a website or application—the parts that a user doesn't see. Python's role in web development can include sending data to and from servers, processing data and communicating with databases, URL routing, and ensuring security. Python offers several frameworks for web development. Commonly used ones include Django and Flask. Some web development jobs that use Python include back-end engineers, full stack engineers, Python developers, software engineers, and DevOps engineers

Automation or scripting

If you find yourself performing a task repeatedly, you could work more efficiently by automating it with Python. Writing code used to build these automated processes is called scripting. In the coding world, automation can be used to check for errors across multiple files, convert files, execute simple math, and remove duplicates in data. Python can even be used by relative beginners to automate simple tasks on the computer—such as renaming files, finding and downloading online content or sending emails or texts at desired intervals.

- **Notebook documents:** Self- contained documents that contain a representation of all content visible in the notebook web application, including inputs and outputs of the

Software testing and prototyping

In software development, Python can aid in tasks like build control, bug tracking, and testing. With Python, software developers can automate testing for new products or features. Some Python tools used for software testing include Green and Requestium.

Everyday tasks

Python isn't only for programmers and data scientists. Learning Python can open new possibilities for those in less data-heavy professions, like journalists, small business owners, or social media marketers. Python can also enable non-programmers to simplify certain tasks in their lives. Here are just a few of the tasks you could automate with Python:

- Keep track of stock market or crypto prices
- Send yourself a text reminder to carry an umbrella anytime it's raining
- Update your grocery shopping list
- Renaming large batches of files
- Converting text files to spreadsheets
- Randomly assign chores to family members
- Fill out online forms automatically

SOFTWARE TOOLS

The software requirements are description of features and functionalities of the target system. Requirements convey the expectations of users from the software product. The requirements can be obvious or hidden, known or unknown, expected or unexpected from client's point of view.

1.Jupyter Notebook:

The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use. The Jupyter Notebook combines three components:

- **The notebook web application:** An interactive web application for writing and running code interactively and authoring notebook documents.
- **Kernels:** Separate processes started by the notebook web application that runs users' code in a given language and returns output back to the notebook web application. The kernel also handles things like computations for interactive widgets, tab completion and introspection.

computations, narrative text, equations, images, and rich media representations of objects. Each notebook document has its own kernel.



PyCharm

The PyCharm editor is the main part of the IDE that you use to create, read and modify code. PyCharm is the most popular IDE for Python, and includes great features such as excellent code completion and inspection with advanced debugger and support for web programming and various frameworks. The intelligent code editor provided by PyCharm enables programmers to write high quality Python code. The editor enables programmers to read code easily through colour schemes, insert indents on new lines automatically, pick the appropriate coding style, and avail context-aware code completion suggestions. At the same time, the programmers can also use the editor to expand a code block to an expression or logical block, avail code snippets, format the code base, identify errors and misspellings, detect duplicate code, and auto-generate code. PyCharm offers some of the best features to its users and developers in the following aspects

- Code completion and inspection
- Advanced debugging
- Support for web programming and frameworks such as Django and Flask



GOOGLE COLAB

- Google Colab makes data science, deep learning, neural network, and machine learning accessible to individual researchers who cannot afford costly computational infrastructure.
- Colab is a Jupyter Notebook-like product from Google Research. A Python program developer can use this notebook to write and execute random Python program codes just using a web browser.
- In a nutshell, Colab is a cloud-hosted version of Jupyter Notebook. To use Colab, you do not need to install and runtime or upgrade your computer hardware to meet Python's CPU/GPU intensive work
- load requirements. Furthermore, Colab gives you free access to computing infrastructure like storage, memory, processing capacity, graphics processing units (GPUs), and tensor processing units (TPUs).
- Google has specially programmed this cloud-based Python coding tool keeping in mind the needs of machine learning programmers, big data analysts, data scientists, AI researchers, and Python learners.
- The best part is one code notebook for all the components needed to present a complete machine learning or data science project to program supervisors or sponsors. For example, your Colab notebook can contain executable codes, live Python codes, rich text, HTML

LaTeX, images, data visualizations, charts, graphs, tables, and more. Google Colab is simply an online representation of Jupyter Notebook. While Jupyter Notebook needs installation on a computer and can only use local machine resources, Colab is a full-fledged cloud app for Python coding.

Best Features of Google Colab

- GPUs and TPUs
- Notebook Sharing
- Special Library Installation
- Pre-Installed Libraries
- Collaborative Coding
- Cloud Storage
- Cloud Storage
- Multiple Data Sources
- Automatic Version Control

CHAPTER 6

SYSTEM DESIGN

Agriculture is the backbone of the Indian economy. The productivity of agriculture is very low. So as the demand of food is increasing, the researchers, farmers, agricultural scientists and government are trying to put extra effort and techniques for more production. Data mining can be used for predicting the future trends of agricultural processes. Data mining is the process to discover interesting knowledge from large amounts of data. Data mining is the process that results in the discovery of new patterns in large data sets. The goal of the data mining process is to extract knowledge from an existing data set and transform it into a human understandable formation for advance use. It is the process of analysing data from different perspectives and summarizing it into useful information. There is no restriction to the type of data that can be analysed by data mining. We can analyse data contained in a relational database, a Datawarehouse, a web server log or a simple text file. Analysis of data in effective way requires understanding of appropriate techniques of data mining. The intention of this paper is to give details about different data mining techniques in perspective of agriculture domain so researchers can get details about appropriate data mining technique in context to their work area. Data mining tasks can be classified into two categories: Descriptive data mining and Predictive data mining. Descriptive data mining tasks characterize the general properties of the data in the database while predictive data mining is used to predict explicit values based on patterns determined from known results. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. As far as datamining technique is concern; in the most of cases predictive data mining approach is used. Predictive data mining technique is used to predict future crop, weather forecasting, pesticides and fertilizers to be used, revenue to be generated and so on.

Fig:6.1 SYSTEM DESIGN

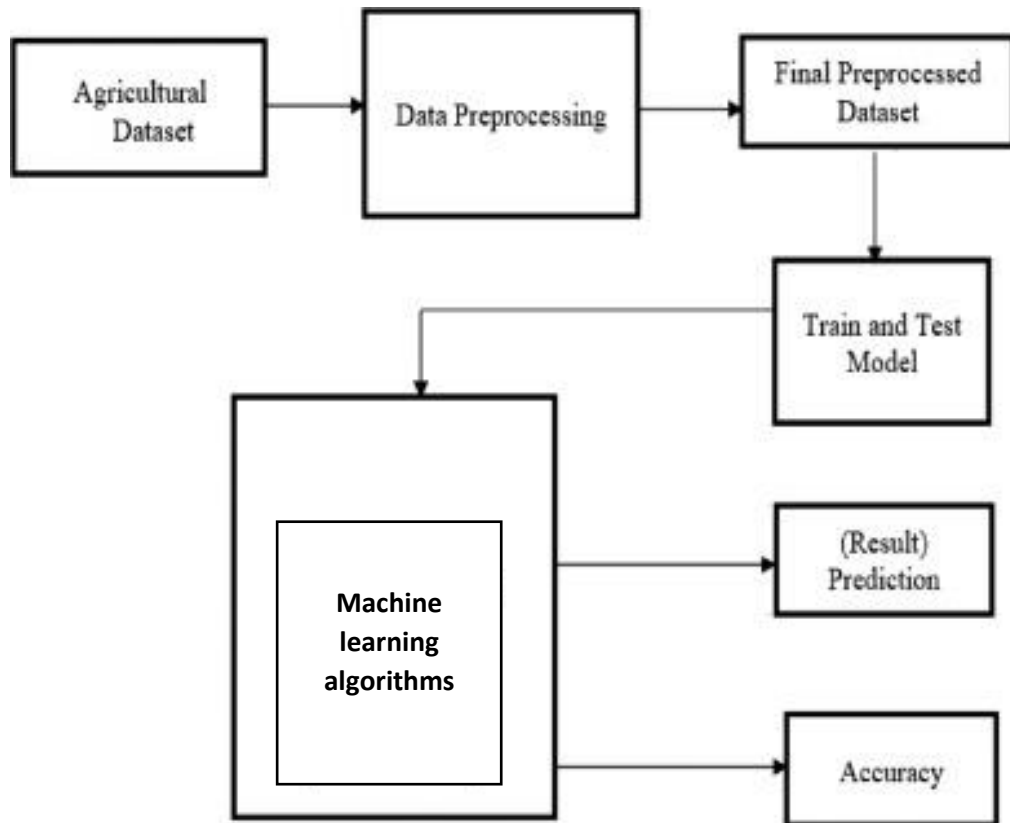
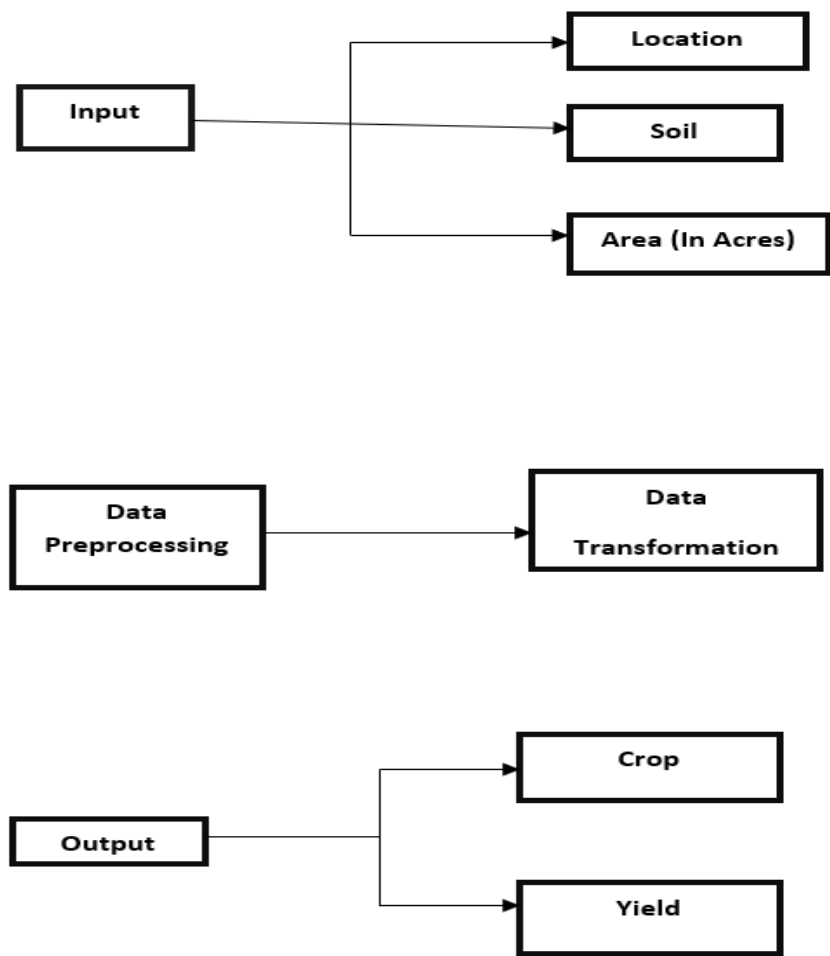


Fig:6.2 SYSTEM DESIGN



CHAPTER 7

IMPLEMENTATIONS

Data Mining Technique

In this we have used the data mining techniques include Classification to predict the exact pattern for the best crop yield prediction. With the help of these data mining techniques farmers can grow the crop yield. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. It is a process in which a model learns to predict a class label from a set of training data which can then be used to predict discrete class labels on new samples. To maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training is one of the major goals of classification algorithm.

Data mining classification algorithms can follow three different learning approaches: supervised learning, unsupervised learning, semi-supervised learning.

The different classification techniques for discovering knowledge are Rule Based Classifiers, Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbour (NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, Genetic Algorithms.

Machine learning

Machine learning is a method of data analysis that automates analytically model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. Re-ensuring interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyse bigger, more complex data and deliver faster, more accurate results even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities or avoiding unknown risks. Two of the most widely

adopted machine learning methods are supervised learning and unsupervised learning but there are also other methods of machine learning.

Crop yield prediction models

The dependable attributes can be difficult to find. Several methods of predicting and modelling crop yields have been used in the past with varying success. Farmer has to face the different problems due to various factors which affect the planning made by him in advance. These factors do not have the fixed type of impact, it varies time to time, year to year depends on the situation, climatic nature, increase in costs of various constraints under uncertain environment, ambiguity and vagueness. Fuzzy logic modelling provides the formulation of mathematical modelling to find the interface results in uncertain situations. Statistical models often do not take into account characteristics of the plants, the weather, or the soil attributes limiting their usefulness. Some models are based on information from just a single year or location. When a model is developed using single location or year data, it will have limited practical applications, therefore variability from multiple environments must be included.

Basic process

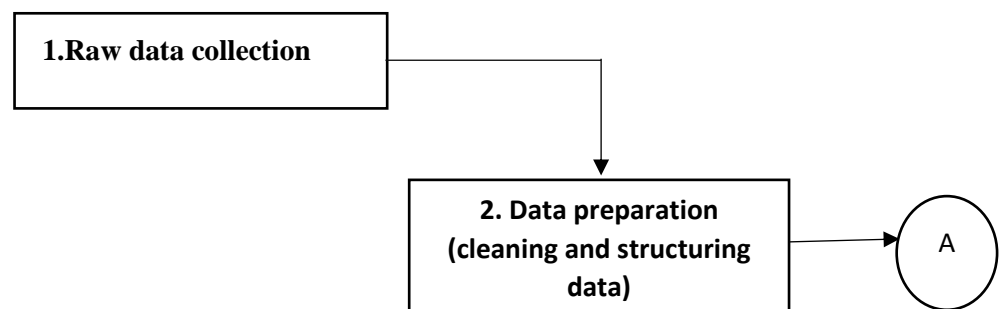
Data Collection: Collect the data that the algorithm will learn from.

Data Preparation: Format and engineer the data into the optimal format, extracting important features and performing dimensionality reduction.

Training: Also known as the fitting stage, this is where the Machine Learning algorithm actually learns by showing it the data that has been collected and prepared.

Evaluation: Test the model to see how well it performs.

Tuning: Fine tune the model to maximize its performance.



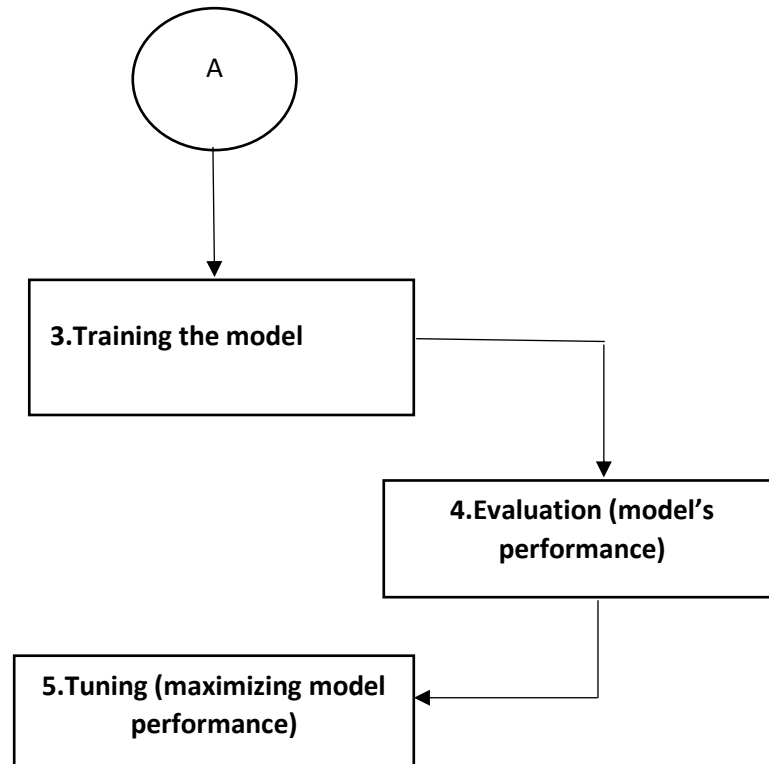


Fig:7.1 BASIC PROCESS OF CROP YIELD

Regression:

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc. Regression is a supervised_learning_technique

It helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modelling, and determining the causal-effect relationship between variables.**

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, “Regression shows a line or curve that passes through all the data points on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum. The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Terminologies Related to the Regression Analysis:

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Use of Regression Analysis

Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

5.5 Linear Regression:

- Linear regression is a statistical regression method which is used for predictive analysis.

- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.

$$Y = aX + b$$

Here, Y=dependent variables(target variables),
 X=Independent variables(predictor variables),
 a and b are the linear coefficients

Some popular applications of linear regression are:

- **Analysing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**
- **Arriving at ETAs in traffic.**

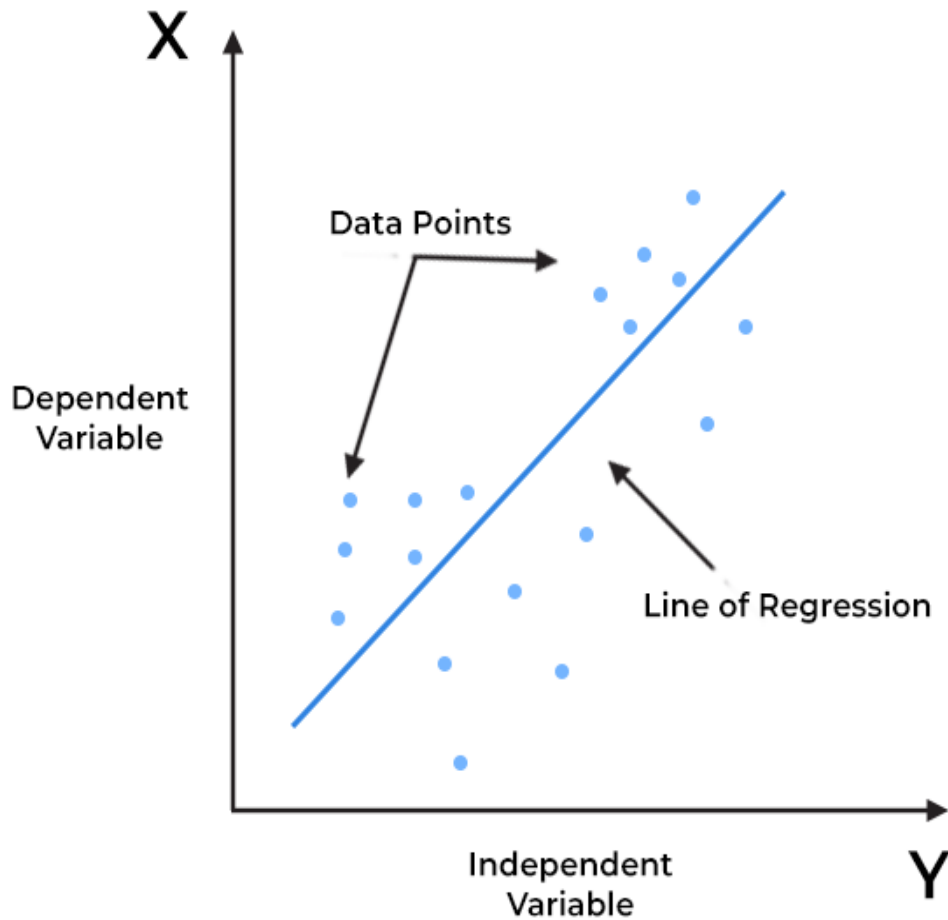


Fig:7.2 LINEAR REGRESSION

5.6 Decision Tree Classification Algorithm

- 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 splits the tree into subtrees.

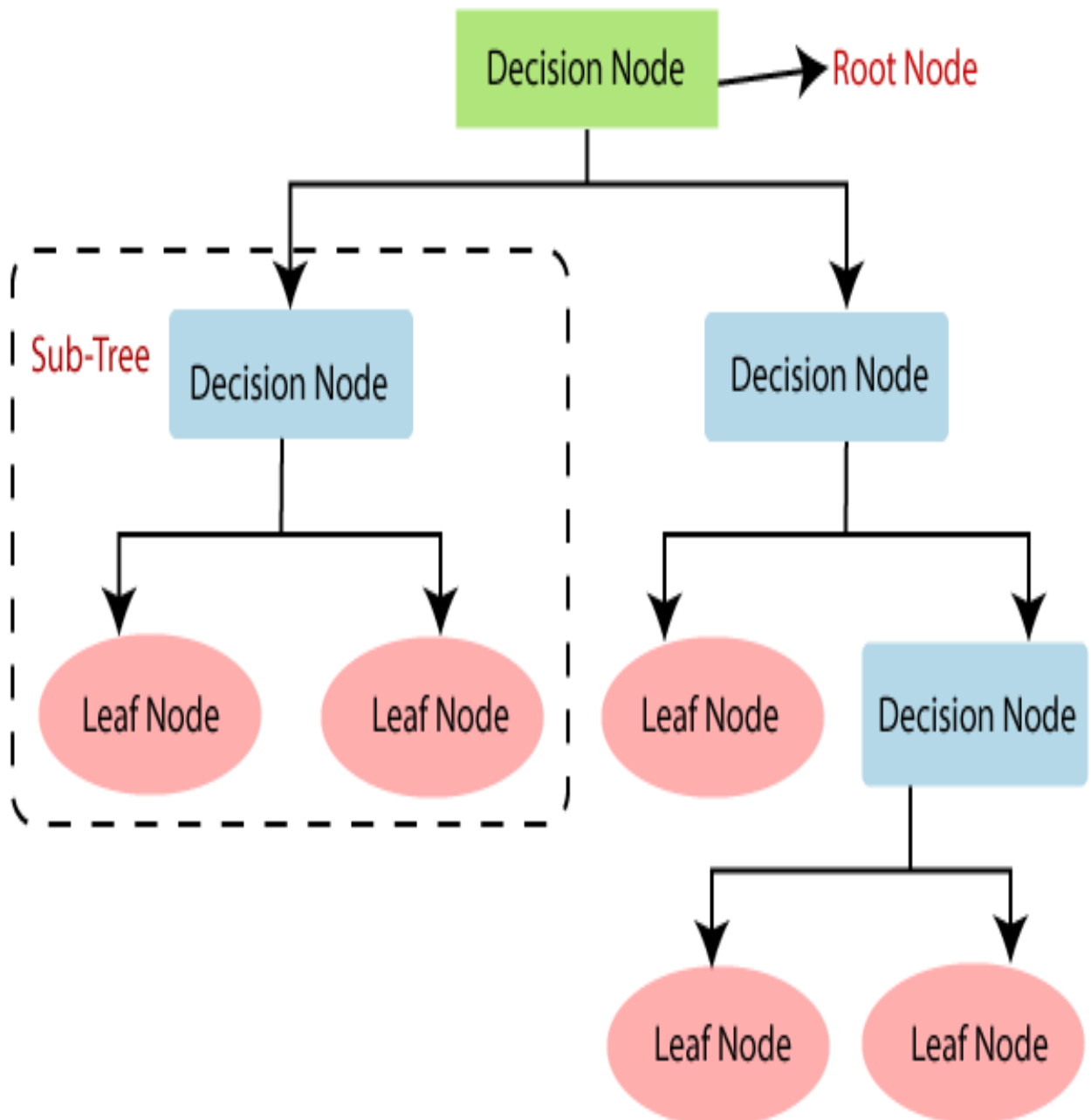


Fig:7.3 DECISION TREE

Decision Tree Terminologies

- **Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

7.1 LIBRARIES THAT ARE USED:

7.1.1 NumPy:

NumPy is a library for the Python programming language, and it's specifically designed to help you work with data. With NumPy, it is easy to create arrays, which is a data structure that allows you to store multiple values in a single variable.

NumPy arrays provide an efficient way of storing and manipulating data. NumPy also includes a number of functions that make it easy to perform mathematical operations on arrays. This can be really useful for scientific or engineering applications. And if when working with data from a Python script, using NumPy can make the life a lot easier.

Arrays are different from Python lists in several ways. First, NumPy arrays are multi-dimensional, while Python lists are one-dimensional. Second, NumPy arrays are homogeneous, while Python lists are heterogeneous. This means that all the elements of a NumPy array must be of the same type. Third, NumPy arrays are more efficient than Python lists. NumPy arrays can be created in several ways. One way is to create an array from a Python list. Once created a NumPy array, it can manipulate in various ways. For example, can change the shape of an array, or can index into an array to access its elements. And also perform mathematical operations on NumPy arrays, such as addition, multiplication, and division.



Fig: 6.1

7.1.2 PANDAS:

Pandas is a Python library, and it is used to analyse data.

It has functions for analysing, cleaning, exploring, and manipulating data.

Pandas gives the answers about the data Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called cleaning the data.

Cleaning Data

Clean Data:

Data cleaning means fixing bad data in your data set.

Bad data could be:

Empty cells

Data in wrong format

Wrong data Duplicates

Cleaning Empty Cells:

Empty cells can potentially give you a wrong result when you analyse data.

One way to deal with empty cells is to remove rows that contain empty cells.

Cleaning Data of Wrong Format:

Cells with data of wrong format can make it difficult, or even impossible, to analyse data.

To fix it, you have two options: remove the rows, or convert all cells in the columns into the same format.

Cleaning Wrong Data:

Wrong data" does not have to be "empty cells" or "wrong format", it can just be wrong, like if someone registered "199" instead of "1.99".

Remove duplicates:

Duplicate rows are rows that have been registered more than one time.

To discover duplicates, we can use the duplicated () method

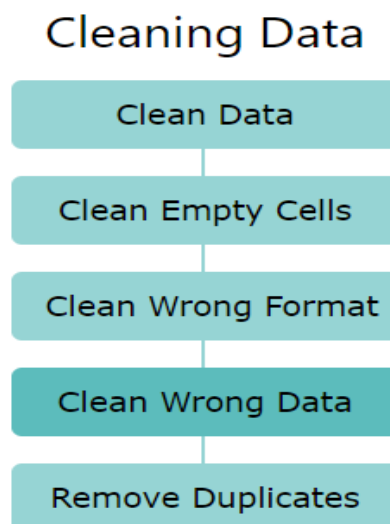


Fig:7.4 CLEANING DATA

Advanced

Correlations:

A great aspect of the Pandas module is the `corr()` method.

The `corr()` method calculates the relationship between each column in your data set

Plotting:

Pandas uses the `plot()` method to create diagrams.

We can use Pyplot, a submodule of the Matplotlib library to visualize the diagram on the screen.

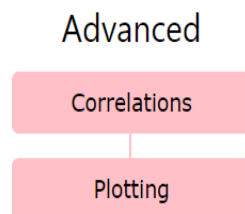


Fig :7.5 ADVANCED

7.1.3 Scikit-learn (Sklearn):

Sklearn is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Features

Rather than focusing on loading, manipulating and summarising data, Scikit-learn library is focused on modelling the data. Some of the most popular groups of models provided by Sklearn are as follows –

Supervised Learning algorithms – Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.

Unsupervised Learning algorithms – On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.

Clustering – This model is used for grouping unlabelled data.

Cross Validation – It is used to check the accuracy of supervised models on unseen data.

Dimensionality Reduction – It is used for reducing the number of attributes in data which can be further used for summarisation, visualisation and feature selection.

Ensemble methods – As name suggest, it is used for combining the predictions of multiple supervised models.

Feature extraction – It is used to extract the features from data to define the attributes in image and text data.

Feature selection – It is used to identify useful attributes to create supervised models.

Open Source – It is open-source library and also commercially usable under BSD license.

CHAPTER 8

TESTING

System testing is actually a series of different tests whose primary purpose is to fully exercise the computer-based system. Although each test has a different purpose, all work to verify that all the system elements have been properly integrated and perform allocated functions. The testing process is actually carried out to make sure that the product exactly does the same thing what is supposed to do. In the testing stage following goals are tried to achieve:

- To affirm the quality of the project.
- To find and eliminate any residual errors from previous stages.
- To validate the software as a solution to the original problem.
- To provide operational reliability of the system

Testing is the most important part of the software development process. Some of the reasons for its importance are as follows:

- Testing helps find the bugs in the software which prevent the program from performing the required tasks.
- Bug fixing in the early stages helps to save a lot of time.
- Testing is essential to ensure that the product will work well once deployed.
- Testing improves the quality of the software.

Validation is the process of ensuring that the software built is in accordance with the business requirements. It assures customer satisfaction.

6.1 Functional testing

This type of testing is done against the functional requirements of the project. Types:

- Unit testing: Each unit /module of the project is individually tested to check for bugs. If any bugs found by the testing team, it is reported to the developer for fixing.
- Integration testing: All the units are now integrated as one single unit and checked for bugs. This also checks if all the modules are working properly with each other.

- System testing: This testing checks for operating system compatibility. It includes both functional and non-functional requirements.
- Sanity testing: It ensures change in the code doesn't affect the working of the project.
- Smoke testing: this type of testing is a set of small tests designed for each build.
- Interface testing: Testing of the interface and its proper functioning.
- Regression testing: Testing the software repetitively when a new requirement is added, when bug fixed etc.
- Beta/Acceptance testing: User level testing to obtain user feedback on the product.

6.2 Non-functional testing

This type of testing is mainly concerned with the non-functional requirements such as performance of the system under various scenarios.

- Performance testing: Checks for speed, stability and reliability of the software, hardware or even the network of the system under test.
- Compatibility testing: This type of testing checks for compatibility of the system with different operating systems, different networks etc.
- Localization testing: This checks for the localized version of the product mainly concerned with UI.
- Security testing: Checks if the software has vulnerabilities and if any, fix them.
- Reliability testing: Checks for the reliability of the software
- Stress testing: This testing checks the performance of the system when it is exposed to different stress levels.
- Usability testing: Type of testing checks the easily the software is being used by the customers.
- Compliance testing: Type of testing to determine the compliance of a system with internal or external standards.

Testing Methodologies:

White box testing

Black box testing

Advantages

- The test is unbiased as the designer and the tester are independent of each other.
- The tester does not need knowledge of any specific programming languages.
- The test is done from the point of view of the user, not the designer.
- Test cases can be designed as soon as the specifications are complete

Testing levels:

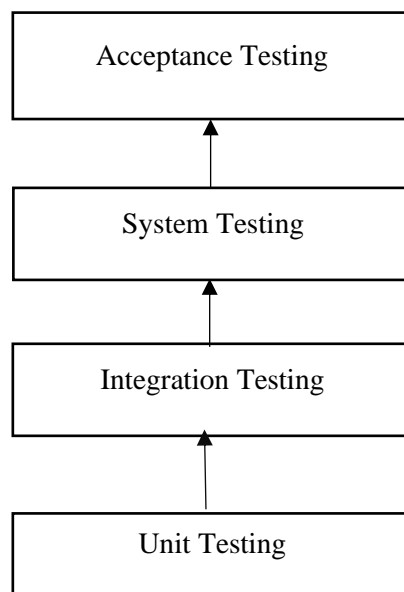


Fig:8.1 TESTING LEVELS

CHAPTER 9

CONCLUSION AND FUTURE SCOPE

CONCLUSION

This system is proposed to deal with the increasing rate of farmer suicides and to help them to grow financially stronger. The Crop Recommender system helps the farmers to predict the yield of a given crop and also helps them to decide which crop to grow. Moreover, it also. Appropriate datasets were collected, studied and trained using machine learning tools. The system tracks the user's location and fetches needed information from the backend based on the location. Thus, the user needs to provide limited information like the soil type and area.

This system contributes to the field of agriculture. One of the most important and novel contributions of the system is suggesting the user the right time to use the fertiliser, this is done by predicting the different parameters. Also, the system provides a list of crops with their productions based on the climatic conditions.

The future work is focused on providing the sequence of crops to be grown depending on the soil and weather conditions and to update the datasets time to time to produce accurate predictions. The Future Work targets a fully automated system that will do the same. Another functionality that we are trying to implement is to provide the correct fertiliser for the given crop and location. To implement this through study of parameters that are required.

Future scope

In future, this model can be implemented throughout the India by adding the data points for all the region. According to our analysis model will give more accuracy as the data points increases, so to get better accuracy model data points can be increased. Our system can be integrated with messaging module so that registered farmers can get the notification of the prediction directly.

APPENDIX

CODING

```
#Importing the required libraries

import pandas as pd

import numpy as np
from sklearn.model_selection import train_test_split
#Reading the csv file
data=pd.read_csv('/content/cpdata.csv')
print(data.head(1))
#Creating dummy variable for target i.e label
label= pd.get_dummies(data.label).iloc[:, 1:]
data= pd.concat([data,label],axis=1)
data.drop('label', axis=1,inplace=True)
print("The data present in one row of the dataset is")
print(data.head(1))
train=data.iloc[:, 0:4].values
test=data.iloc[:, 4:].values
#Dividing the data into training and test set
X_train,X_test,y_train,y_test=train_test_split(train,test,test_size=0.3)

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
#Importing Decision Tree classifier
from sklearn.tree import DecisionTreeRegressor
clf=DecisionTreeRegressor()
#Fitting the classifier into training set
clf.fit(X_train,y_train)
```

```

pred=clf.predict(X_test)
from sklearn.metrics import accuracy_score
# Finding the accuracy of the model
a=accuracy_score(y_test,pred)
print("The accuracy of this model is: ", a*100)
#Using firebase to import data to be tested
from firebase import firebase
firebase =firebase.FirebaseApplication('https://cropit-eb156.firebaseio.com/')
tp=firebase.get('/Realtime',None)
ah=tp['Air Humidity']
atemp=tp['Air Temp']
shum=tp['Soil Humidity']
pH=tp['Soil pH']
rain=tp['Rainfall']

l=[]
l.append(ah)
l.append(atemp)
l.append(pH)
l.append(rain)
predictcrop=[]
# Putting the names of crop in a single list
crops=['wheat','mungbean','Tea','millet','maize','lentil','jute','cofee','cotton','ground
nut','peas','rubber','sugarcane','tobacco','kidney beans','moth
beans','coconut','blackgram','adzuki beans','pigeon peas','chick
peas','banana','grapes','apple','mango','muskmelon','orange','papaya','watermelon','pomegran
ate']
cr='rice'
#Predicting the crop
predictions = clf.predict(predictcrop)
count=0

```

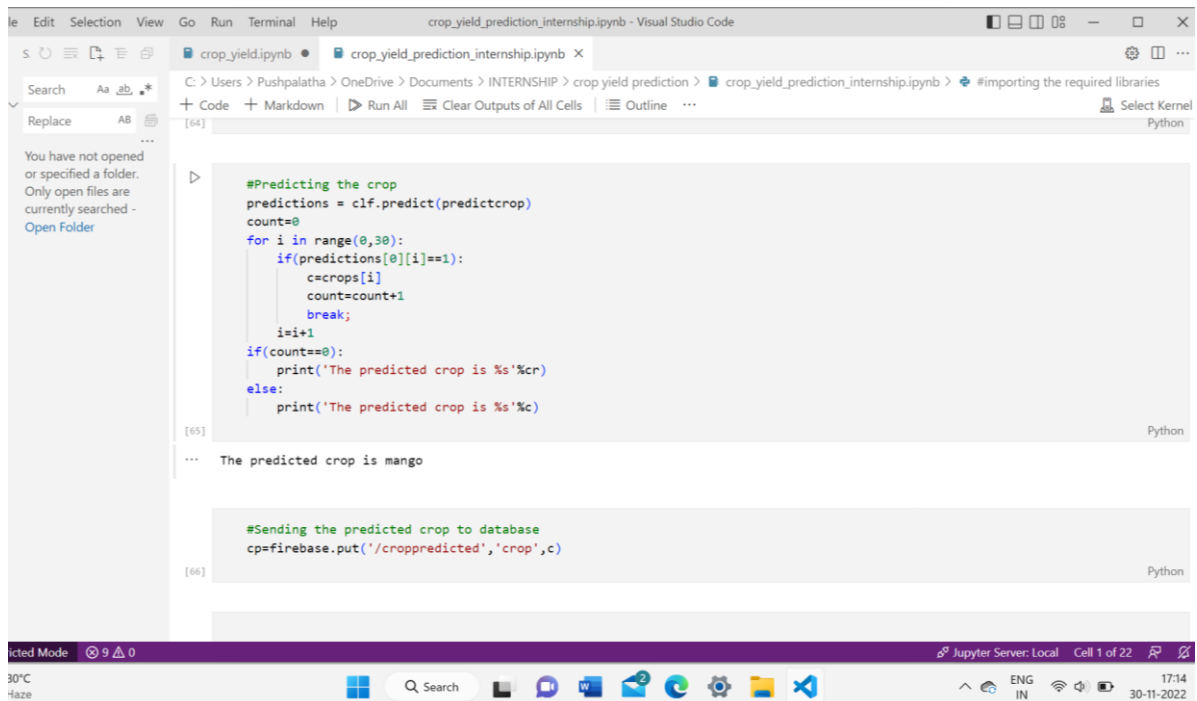
```

for i in range(0,30):
    if(predictions[0][i]==1):
c=crops[i]
        count=count+1

        break;
    i=i+1
if(count==0):
    print("The predicted crop is %s"%cr)
else:
    print("The predicted crop is %s"%c)
#Sending the predicted crop to database
cp=firebase.put('/croppredicted','crop',c)

```

RESULT:



The screenshot shows a Jupyter Notebook with two cells. The first cell contains the following Python code:

```

#Predicting the crop
predictions = clf.predict(predictcrop)
count=0
for i in range(0,30):
    if(predictions[0][i]==1):
        c=crops[i]
        count=count+1
        break;
    i=i+1
if(count==0):
    print('The predicted crop is %s'%cr)
else:
    print('The predicted crop is %s'%c)

```

The output of the first cell is:

```

The predicted crop is mango

```

The second cell contains the following Python code:

```

#Sending the predicted crop to database
cp=firebase.put('/croppredicted','crop',c)

```

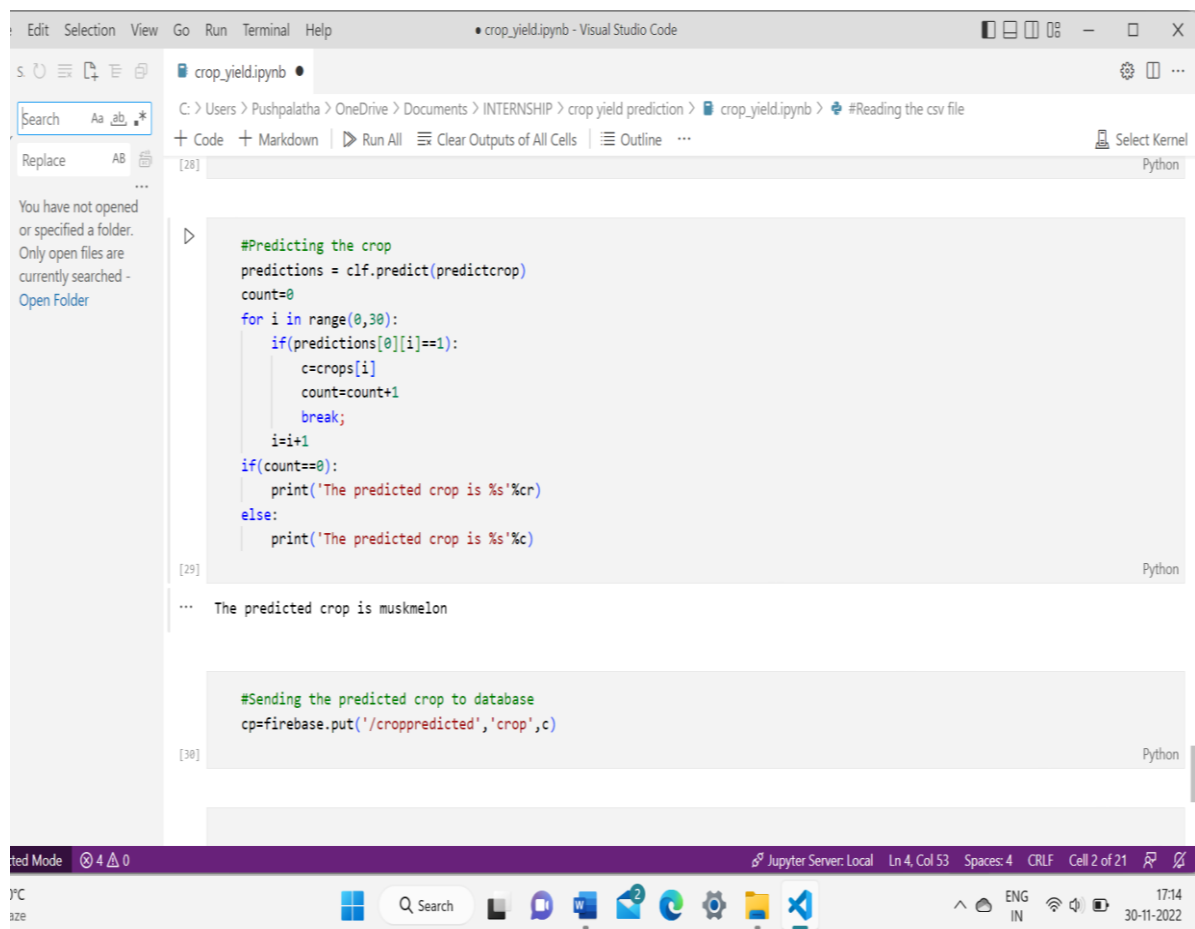
The output of the second cell is:

```


```

The bottom status bar of the Visual Studio Code window shows the temperature as 30°C, the time as 17:14, and the date as 30-11-2022.

Fig:10.1 THE PREDICTED CROP IS MANGO



The screenshot shows a Jupyter Notebook titled 'crop_yield.ipynb' in Visual Studio Code. The notebook is running on a local Jupyter Server. The code in the notebook is as follows:

```
[28] #Predicting the crop
      predictions = clf.predict(predictcrop)
      count=0
      for i in range(0,30):
          if(predictions[0][i]==1):
              c=crops[i]
              count=count+1
              break;
          i=i+1
      if(count==0):
          print('The predicted crop is %s'%cr)
      else:
          print('The predicted crop is %s'%c)

[29] ... The predicted crop is muskmelon

[30] #Sending the predicted crop to database
      cp=firebase.put('/croppredicted','crop',c)
```

The output of the notebook shows the predicted crop is muskmelon. The status bar at the bottom indicates the notebook is in 'Jupyter Server: Local' mode, with the current cell being Cell 2 of 21. The system tray shows the time as 17:14 on 30-11-2022.

Fig :10.2 THE PREDICTED CROP IS MUSKMELON

BIBLIOGRAPHY

REFERENCES

1. <https://www.ijert.org/crop-yield-prediction-using-machine-learning-algorithms>
2. <https://www.sciencedirect.com/science/article/pii/S0168169920302301>
3. <https://www.frontiersin.org/articles/10.3389/fpls.2019.00621/full>
4. https://www.researchgate.net/publication/343730263_Crop_yield_prediction_using_machine_learning_A_systematic_literature_review
5. <https://towardsdatascience.com/predicting-crops-yield-machine-learning-nanodegree-capstone-project-e6ec9349f69>
6. https://www.academia.edu/43293164/Crop_Yield_Prediction_Analysis_using_Feed_Forward_and_Recurrent_Neural_Network
7. data.govt.in
8. <https://www.kaggle.com/>