A Project report

on

CROP YIELD PREDICTION AND FERTILIZER ANALYSIS USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

in

Computer Science & Engineering

By

P. SUCHITHA	**(184G1A0598)**
D.VARSHA	**(184G1A05A8)**
Y.SREELAKSHMI	**(184G1A0596)**
D.VENKATA NARESH	**(194G5A0510)**

Under the Guidance of

Mr. B. Sreedhar **M.Tech, MBA**

Assistant Professor

Department of Computer Science & Engineering

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY

(Affiliated to JNTUA & Approved by AICTE)

(Accredited by NAAC with 'A' Grade &Accredited by NBA(EEE, ECE & CSE))
Rotarypuram Village, B K Samudram Mandal, Ananthapuramu-515701.

2021-2022

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY

(Affiliated to JNTUA & Approved by AICTE)
(Accredited by NAAC with ‘A’ Grade &Accredited by NBA (EEE, ECE & CSE)
Rotarypuram Village, B K Samudram Mandal, Ananthapuramu-515701.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Certificate

This is to certify that the Project report entitled **CROP YIELD PREDICTION AND FERTILIZER ANALYSIS USING MACHINE LEARNING** is the bonafide work carried out by **P. Suchitha** bearing Roll Number **184G1A0598, D. Varsha** bearing Roll Number **184G1A05A8 , Y. Sree Lakshmi** bearing Roll Number **184G1A0596, D. Venkata Naresh** bearing Roll Number **194G5A0510** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** during the academic year 2021 - 2022.

Signature of the Guide	**Head of the Department**
Mr. B. Sreedhar M.Tech., MBA	Mr. P.Veera Prakash M.Tech., (Ph.D.)
Assistant Professor	Assistant Professor & HOD
Date:	**EXTERNAL EXAMINER**
Place: Rotarypuram	

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we have now the opportunity to express our gratitude for all of them.

It is with immense pleasure that we would like to express our indebted gratitude to our Guide **Mr. B. Sreedhar, M.Tech,, MBA. ,Assistant Professor, Computer Science & Engineering**, who has guided us a lot and encouraged us in every step of the project work. We thank him for the stimulating guidance, constant encouragement andconstructive criticism which have made possible to bring out this project work.

We are very much thankful to **Mr. P. Veera Prakash, M.Tech., (Ph.D.), Assistant Professor & Head of the Department, Computer Science & Engineering,** for his kind support and for providing necessary facilities to carry out the work.

We wish to convey our special thanks to **Dr. G. Bala Krishna, Ph.D., Principal** of **Srinivasa Ramanujan Institute of Technology** for giving the required information in doingour project work. Not to forget, We thank all other faculty and non- teaching staff, and our friends who had directly or indirectly helped and supported us in completing our project in time.

We also express our sincere thanks to the Management for providing excellent facilities**.**

Finally, we wish to convey our gratitude to our families who fostered all the requirements and facilities that we need.

Project Associates

184G1A0598	SUCHITHA P
184G1A05A8	VARSHA D
184G1A0596	SREELAKSHMI Y
194G5A0510	VENKATA NARESH D

DECLARATION

We, Ms P. Suchitha with reg no: 184G1A0598 , Ms D. Varsha with reg no: 184G1A05A8, Ms Y. Sreelakshmi with reg no: 184G1A0596 , Mr D. Venkata Naresh with reg no: 194G5A0510 students of SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY , Rotarypuram , hereby declare that the dissertation entitled **"CROP YIELD PREDICTION AND FERTILIZER ANALYSIS "** embodies the report of our project work carried out by us during IV year Bachelor of Technology under the guidance of **Mr. B. Sreedhar M.Tech., MBA**. Department of CSE, **SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY**, and this work has been submitted for the partial fulfillment of the requirements for the award of the Bachelor of Technology degree.

The results embodied in this project have not been submitted to any other University of Institute for the award of any Degree or Diploma.

P. SUCHITHA	Reg no: 184G1A0598
D.VARSHA	Reg no: 184G1A05A8
Y.SREELAKSHMI	Reg no: 184G1A0596
D.VENKATA NARESH	Reg no: 194G5A0510

CONTENTS

Page No.

List of Figures

LIST OF ABBREVIATIONS

NPK	Nitrogen, Phosphorous, Potassium
EDA	Exploratory Data Analysis
ANN	Artificial Neural Networks
SVM	Support Vector Machine
CSV	Comma Separated Values
SRS	System Requirements Specification
RAM	Random Access Memory
UML	Unified Modelling Language

ABSTRACT

Agriculture is the keystone of a developing country such as India. For the revenue, the majority of their population depends on agriculture. Machine Learning is an imminent field of informatics that can be applied quite efficiently to the agricultural sector. Crop yield prediction and forecasting is essential for agricultural stakeholders which can be acquired through machine learning techniques. When the farmers are not aware of the soil nutrition and soil composition that results in minimal crop yield. Thus the proposed system developed, which in turn focuses on the macronutrients (NPK), pH and electrical conductivity in the soil and temperature for providing the most appropriate crop suggestions. The proposed system constructs a collaborative system of crop rotation, crop yield prediction and forecasting and fertilizer recommendation. In this project a system is developed which incorporates the agricultural dataset wherein voting based ensemble classifier algorithm is applied to suggest the appropriate crops. Crop yield prediction and forecasting will increase the agricultural production. Periodical crop rotation will improve the soil fertility. This system supports farmer friendly fertilization decision making. The accuracy of this system was around 95%.

Keywords

Nitrogen, Phosphorus, Potassium, soil nutrition, yield prediction, crop rotation, fertilize recommendation, Ensemble classifier, voting.

CHAPTER 1

INTRODUCTION

Agriculture plays an essential part in an economy's life. They are the backbone of our country's economy system. One of the key problems confronting farmers is selecting the right crop for cultivation. Selection of crops is determined by several factors such as temperature, soil composition, market prices etc.

Machine Learning is a technique that uses complex algorithms and a collection of predefined rules to operate intelligently. It uses past data to read the patterns and then perform the intended task according to the defined rules and algorithms based on the analysis it produces. Machine Learning is an imminent field of informatics that can be applied quite efficiently to the agricultural sector. Machine Learning is everywhere throughout the entire growing and harvesting cycle.

The major factors affecting crop yield are the soil type, land type and the macronutrients present in the soil. The purpose of this work is to categorize the soil samples according to the macro nutrients found therein and to predict the crops which can be grown in the soil. The crop recommendation system which is developed incorporates the agricultural dataset. The Nitrogen (N), Phosphorus (P) and Potassium (K), Soil type, Soil texture, Land type, pH and Electrical conductivity of the soil are taken as input to recommend the crops.

The system for crop yield prediction, forecasting and fertilizer recommendation are all separate and distinct in the existing system. The proposed system constructs a collaborative system of crop yield prediction and forecasting, crop rotation and fertilizer recommendation.

This system also proposes the required fertilizer to boost the nutrients contained in the soil and thus enhance the yield of the crop. Thus there arises a need for suggesting suitable crops and fertilizers using machine learning algorithm.

1.1 Problem Statement

The Problem Statement revolves around prediction of crop yield using Machine Learning Techniques. The goal of the project is to help the users choose a suitable crop to grow in order to maximize the yield and hence the profit.

The system proposed tries to overcome the drawbacks of existing systems and make predictions by analyzing structured data. The solution we are proposing is to design a system taking into consideration the most influencing parameters to grow a crop and to get a better selection of crops which can be grown over the season. This would help reduce the difficulties faced by the farmers in selecting the crop to get high yield and thus maximize profits which in turn will reduce the suicide rates.

The system consists of two main modules:

i. Yield Prediction Module – In this the user selects the crop and area and predicts the yield how much he can get.

ii. Fertilizer Module - This module helps the user decide whether a particular fertilizer is right for using.

1.2 Objectives

This project aims at predicting the crop yield at a particular weather condition and thereby recommending suitable crops for that field. It involves the following steps.

- Collect the weather data, crop yield data, soil type data and the rainfall data and merge these datasets in a structured form and clean the data. Data Cleaning is done to remove inaccurate, incomplete and unreasonable data that increases the quality of the data and hence the overall productivity.
- Perform Exploratory Data Analysis (EDA) that helps in analyzing the complete dataset and summarizing the main characteristics. It is used to discover patterns, spot anomalies and to get graphical representations of various attributes. Most importantly, it tells us the importance of each attribute, the dependence of each attribute on the class attribute and other crucial information.

- Divide the analyzed crop data into training and testing sets and train the model using the training data to predict the crop yield for given inputs.
- Compare various Algorithms by passing the analyzed dataset through them and calculating the error rate and accuracy for each. Choose the algorithm with the highest accuracy and lowest error rate.
- Implement a system in the form of a mobile application and integrate the algorithm at the back end.
- Test the implemented system to check for accuracy and failures.

1.3 Scope of the Project

Integrating farming and machine learning, we can lead to further advancements in agriculture by maximizing yield and optimizing the use of resources involved. Previous year‘s production data is an essential element for predicting the current yield.

The goal of this project is to help the farmers by combining agriculture and technology. The end result is an application that is available on the web as well as mobile.

The application has the following features:

i. Yield Prediction: This is one of the modules available in the application that enables the user to view the yield predictions of crops. The user is required to give the inputs of selected plants and predict the yield.

ii. Fertilizer: This is the second module available. The functionality that this module provides revolves around whether using fertilizers at a certain point of time would be recommended or not.

Agriculture is one of the main sources of income in India and there is an enormous need to maintain agricultural sustainability with the increasing rate of farmer suicides. Hence it is a significant contribution towards the economic and agricultural welfare of the countries across the world.

CHAPTER 2

LITERATURE SURVEY

Mansi Shinde et al [1], Designed a system which furnish farmers, the expert advice to identify the appropriate fertilizer and crops. The farmers can use this model using smart phones based on android. This method enhances the production of the crop. This software also allows people to buy the recommended fertilizers from the shopping portal.

V. Sellam et al [2], Evaluated environmental parameters such as Cultivation Area, Annual Rainfall and Food Price Index which affects the crop yield. The crop yield is a dependent variable that depends on all of these environmental factors. From the results produced factors like Weather Conditions, Soil Parameters etc have great effect on the crop production when compared to Cultivation Area and Food Price Index.

U.K. Diwan et al [3], Developed a climate based model to make a preliminary accurate crop yield forecast in advance. The focus of this work was on the crop yield forecast model through the use of weather parameters and crop yield history. Temperature (maximum and minimum) and relative humidity were found to play a major role among all the weather factors in all the districts.

Rushika Ghadge et al [4], presented a tool to help farmers for monitoring soil quality based on data mining techniques. The system focuses on the soil quality inspection to predict crops that are highly suitable for the soil type. This method also recommends an appropriate fertilizer for crop yield optimization. It recommends the crops by exploring soil nutrition contained in the agricultural land and suggest the crop that produces high productivity.

P. Priya et al [5], designed a framework that focuses on the prediction of crop yields based on the dataset of the Kharif and Rabi cropping seasons. This dataset is used to create the prediction model. The results obtained from this framework would be useful for farmers to forecast the yield before they are cultivated on agriculture land.

Vaneesbeer Singh et al [6], proposed a method to determine the yield class based on the macro and micro nutrients present in the soil, by using different machine learning algorithms. After analysis this system provides suggestions on the highly appropriate crops which results with the maximum profit.

Vrushal Milan Dolas et al [7] implemented an upgraded decision tree algorithm and the soil data set was incorporated by the classifier which is used in this application. The soil is with similar behaviour is grouped as one class, so that the farmers will be aware of the type of soil and can plant the crops appropriately.

R.Sujatha et al [8], proposed the fundamental concepts for interpreting phenology and suitable planting dates for distinct genotypes. This paper describes the idea of estimating the crop yield in advance and to make the right choice of the most appropriate crop to increase the value and gain of the agricultural field.

Supriya D M et al [9], Developed a system where data mining techniques were used to predict class of the analyzed soil datasets. Thus the predicted class will indicate high profitable crop yield. Classification algorithms were designed to label the unknown samples using the information provided by a sequence of classified samples.

S. Veenadhari et al [10], developed a web application to forecast the impact of climate variables on crop production and to find the most effective weather factor on crop production for selected plants in particular areas of Madhya Pradesh. The application provided a summary of the potential impact of different weather factors which are responsible for crop yield.

CHAPTER 3

METHODOLOGY

The system uses machine learning to make predictions of the crop and Python as the programming language since Python has been accepted widely as a language for experimenting in the machine learning area. Machine learning uses historical data and information to gain experiences and generate a trained model by training it with the data. This model then makes output predictions. The better the collection of dataset, the better will be the accuracy of the classifier. It has been observed that machine learning methods such as regression and classification perform better than various statistical models [2].

Crop production is completely dependent upon geographical factors such as soil chemical composition, rainfall, terrain, soil type, temperature etc. These factors play a major role in increasing crop yield. Also, market conditions affect the crop(s) to be grown to gain maximum benefit. We need to consider all the factors altogether to predict the yield. Hence, using Machine Learning techniques in the agricultural field, we build a system that uses machine learning to make predictions of the production of crops by studying the factors such as rainfall, temperature, area, season, etc.

3.1 Machine Learning

Machine Learning is undeniably one of the most influential and powerful technologies in today‘s world. Machine learning is a tool for turning information into knowledge. In the past 50 years, there has been an explosion of data [10]. This mass of data is useless; we analyse it and find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover.

The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making. To learn the rules governing a phenomenon, machines have to go through a learning process**,** trying different rules and learning from how well they perform. Hence, why it‘s known as Machine Learning.

Basic Terminology

- Dataset: A set of data examples, which contain features important to solving the problem.
- Features: Important pieces of data that help us understand a problem. These are fed into a Machine Learning algorithm to help it learn.
- Model: The representation (internal model) of a phenomenon that a Machine Learning algorithm has learnt. It learns this from the data it is shown during training. The model is the output you get after training an algorithm. For example, a decision tree algorithm would be trained and produce a decision tree model.

Types of Machine Learning

There are multiple forms of Machine Learning; supervised, unsupervised, semi supervised and reinforcement learning. Each form of Machine Learning has differing approaches, but they all follow the same underlying process and theory.

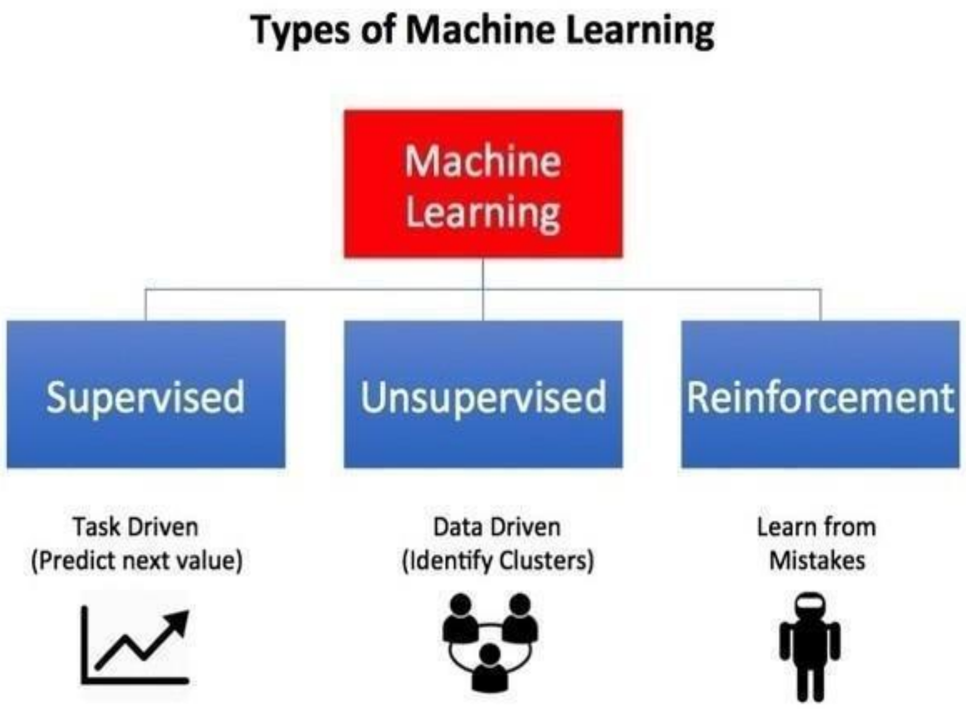


Fig 3.1: Types of Machine Learning

Supervised Learning: It is the most popular paradigm for machine learning. Given data in the form of examples with labels, we can feed a learning algorithm these example-label pairs one by one, allowing the algorithm to predict the label for each example, and giving it feedback as to whether it predicted the right answer or not. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully-trained, the supervised learning algorithm will be able to observe a new, never before-seen example and predict a good label for it.

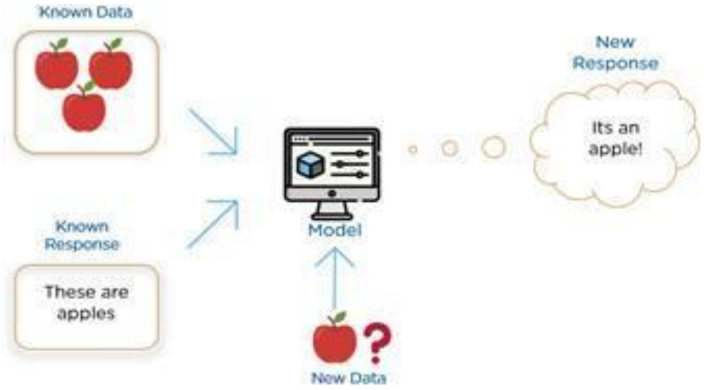


Fig 3.2: Process of Supervised Learning

Unsupervised learning: It is very much the opposite of supervised learning. It features no labels. Instead, the algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and organize the data in a way such that a human can come in and make sense of the newly organized data. Because unsupervised learning is based upon the data and its properties, we can say that unsupervised learning is data- driven. The outcomes from an unsupervised learning task are controlled by the data and the way it‘s formatted.

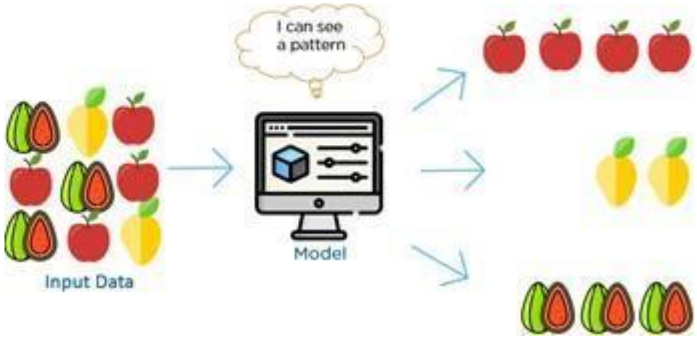


Fig 3.3: Process of Unsupervised Learning

Reinforcement learning: It is fairly different when compared to supervised and unsupervised learning. Reinforcement learning is very behaviour driven. It has influences from the fields of neuroscience and psychology. For any reinforcement learning problem, we need an agent and an environment as well as a way to connect the two through a feedback loop. To connect the agent to the environment, we give it a set of actions that it can take that affect the environment. To connect the environment to the agent, we have it continually issue two signals to the agent: an updated state and a reward (our reinforcement signal for behaviour).

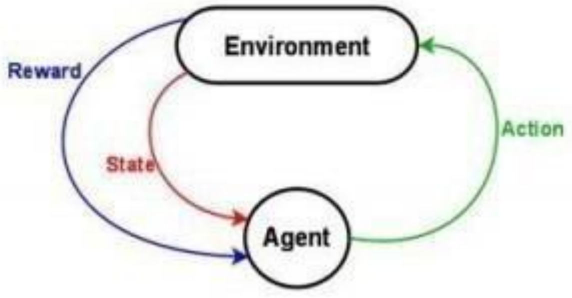


Fig 3.4: Reinforcement Learning

3.2 Basic Process of machine learning

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

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

iii. 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.

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

v. Tuning**:** Fine tune the model to maximize its performance.

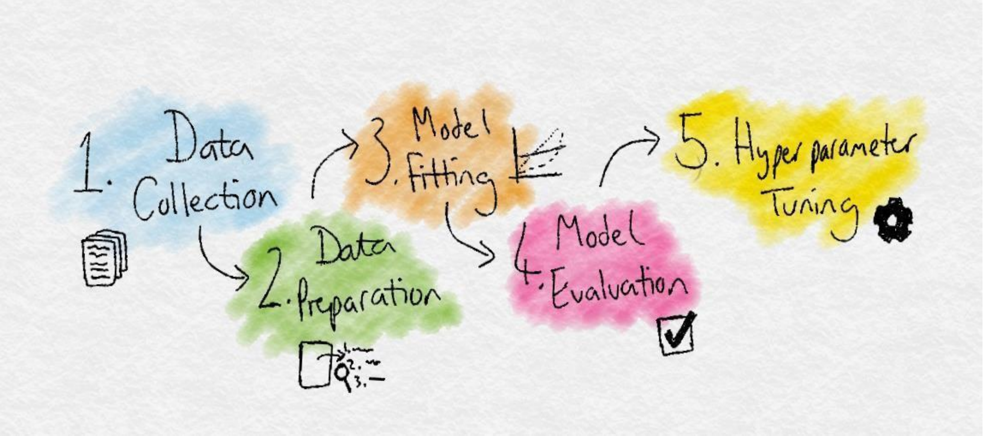


Fig 3.5: Basic process of Machine learning

3.3 Algorithms Used

Machine Learning offers a wide range of algorithms to choose from. These are usually divided into classification, regression, clustering and association. Classification and regression algorithms come under supervised learning while clustering and association comes under unsupervised learning.

- Classification: A classification problem is when the output variable is a category, such as —red‖ or —blue‖ or —disease‖ and —no disease‖. Example: Decision Trees
- Regression: A regression problem is when the output variable is a real value, such as dollars or weight. Example: Linear Regression
- Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior Example: k means clustering.
- Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

A few algorithms can come under multiple types. Considering the problem statement and the desired output of the project, the most suitable type of algorithm would come under regression.

Before choosing an algorithm and working with it further, many algorithms were explored and the error rates and accuracy were checked for each.

ANN vs. Random Forest: Random Forest is less computationally expensive and does not require a GPU to finish training. A random forest can give you a different interpretation of a decision tree but with better performance. Neural Networks will require much more data than an everyday person might have on hand to actually be effective. The neural network will simply decimate the interpretability of your Features to the point where it becomes meaningless for the sake of performance.

- SVM vs. Random Forest: Random forests are probably THE worry-free‖ approach. There are no real hyper parameters to tune (maybe except for the number of trees; typically, the more trees we have the better). On the contrary, there are a lot of knobs to be turned in SVM. SVMs: Choosing the right‖ kernel, regularization penalties, the slack variable, etc. Random forests are much simpler to train and easier to find a good, robust model. In SVMs, we typically need to do a fair amount of parameter tuning, and in addition to that, the computational cost grows linearly with the number of classes as well.
- Linear Regression vs. Random Forest: Random forests very often outperform linear regression. Random forests fit data better from the get-go without transforms. They‘re more forgiving in almost every way. You don‘t need to scale your data, you don‘t need to do any monotonic transformations (log, etc.). You often don‘t even need to remove outliers. You can throw in categorical features, and it‘ll automatically partition the data if it aids the fit. You don‘t have to spend any time generating interaction terms. And perhaps most important: in most cases, it‘ll probably be notably more accurate.
- KNN vs. Random Forest: Random Forest is faster due to KNN‘s expensive real time Execution. KNN should be wisely selected while there is no such

decision to be made in Random Forest. KNN has large computation cost during runtime if sample size is large.

Random Forest

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity. It can be used for both classification and regression tasks. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sk-learn provides a great tool for this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results so the sum of all importance is equal to one.

The hyper parameters in random forest are either used to increase the predictive power of the model or to make the model faster. Python offers some built in random Forest functions which have the following hyper parameters.

i. To increase the predictive power:

- Firstly, there is the n_estimators hyper parameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions.
- Another important hyper parameter is max_features , which is the maximum number of features random forest considers to split a node.
- The last important hyper parameter is min_sample_leaf . This determines the minimum number of leafs required to split an internal node.

ii To Increase the model's speed:

- The n jobs hyper parameter tells the engine how many processors it is allowed to use. If it has a value of one, it can only use one processor.

A value of means that there is no limit.

- The random state hyper parameter makes the model‵s output replicable. The model will always produce the same results when it has a definite value of random state and if it has been given the same hyper parameters and the same training data.
- The oob score (also called oob sampling), which is a random forest cross validation method. In this sampling, about one-third of the data is not used to train the model and can be used to evaluate its performance. These samples are called the out-of-bag samples.

The working of Random Forest is as follows:

● **Step 1** − First, start with the selection of random samples from a given dataset.

● **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

● **Step 3** − In this step, voting will be performed for every predicted result.

● **Step 4** − At last, select the most voted prediction result as the final prediction result.

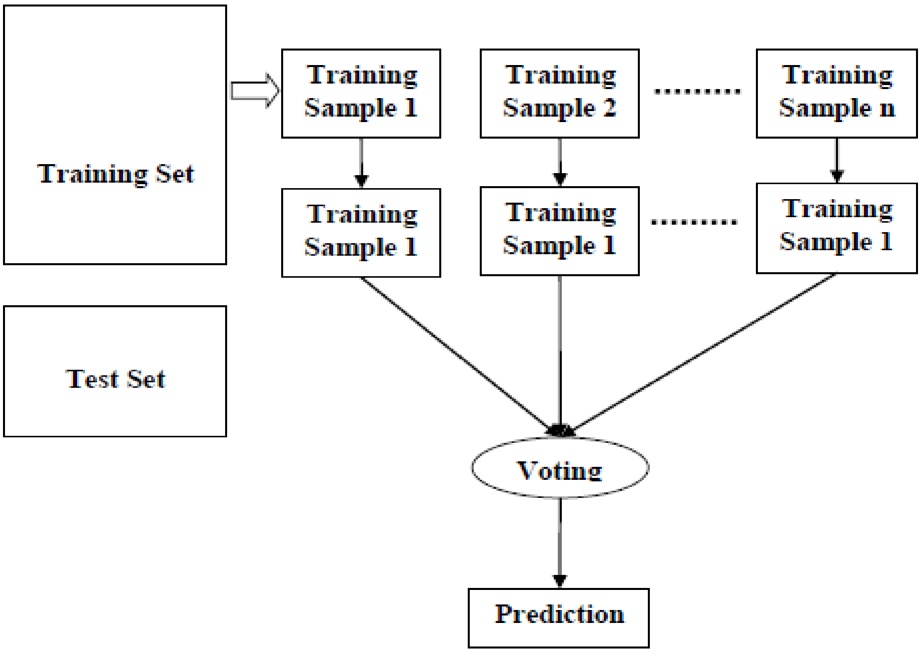


Fig 3.6: Working of Random Forest algorithm

Another Reason for choosing Random Forest:

The data of a particular crop was taken and passed through two algorithms .i.e., Random Forest and another Algorithm that is said to give best results for that crop. The accuracy achieved in both the algorithms were compared. Rice and Groundnut were chosen based on the research papers that were found.

- Rice: According to the paper, the best algorithm for rice yield prediction is Linear Regression . After running both the algorithms, we found a very high difference between actual value and predicted value in Linear Regression while Random Forest continued to maintain an accuracy of 90+.
- Groundnut: The Research Paper stated that KNN works best for Groundnut yield prediction on running the algorithms, we did not find much difference in the results in both the algorithms.

Hence, we can conclude that Random Forest can be used as a general algorithm which gives a considerably high accuracy with Good predictions.

CHAPTER 4
SYSTEM REQUIREMENTS SPECIFICATION

A software requirements specification (SRS) is a description of a software system to be developed. It lays out functional and nonfunctional requirements, and may include a set of use cases that describe user interactions that the software must provide. It is very important in a 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.

A 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

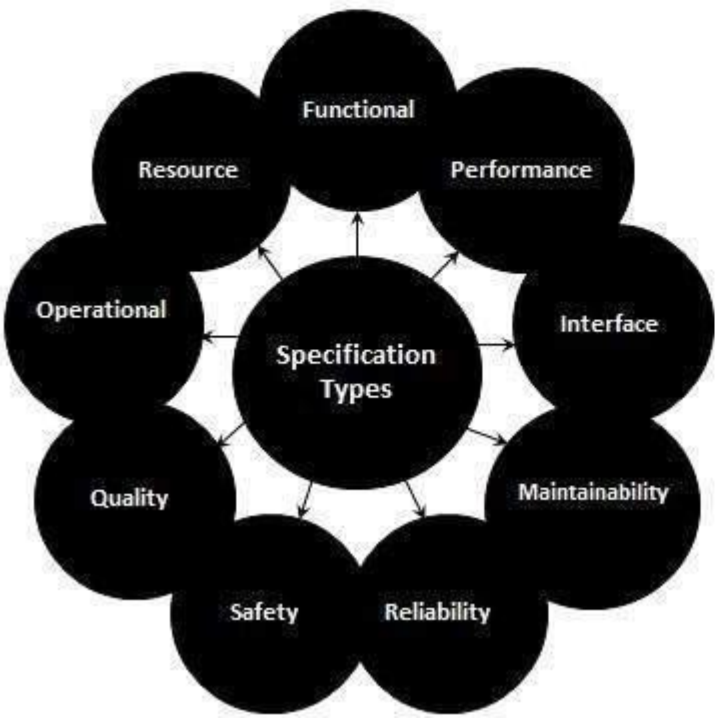


Fig 4.1: Types of Requirements in SRS

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.

4.1 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 you 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 you 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 behavior.
- Errors caught in the Functional requirement gathering stage are the cheapest to fix.
- Support user goals, tasks, or activities

4.2 Basic Requirements

1. **Data collection**: The dataset used in this project is the data collected from reliable websites and merged to achieve the desired data set. The sources of our datasets are: https://en.tutiempo.net/ for weather data and https://www.kaggle.com/srinivas1/agricuture-crops-production-in-india for crop yield data. It consists of names of the crops, production, area, average temperature, average rainfall (mm), season, year, name of the states and the districts. 'Production' is the dependent variable or the class variable. There are eight independent variables and 1 dependent variable.

2. Data Preprocessing: The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling. Here, data pre-processing focuses on finding the attributes with null values or invalid values and finding the relationships between various attributes as well. Data Pre-processing also helps in finding out the impact of each parameter on the target parameter. To preprocess our datasets we used EDA methodology. All the invalid and null values were handled by removing that record or giving the default value of that particular attribute based on its importance.

3. Dataset splitting: A dataset used for machine learning should be partitioned into two subsets — training and test sets. We split the dataset into two with a split ratio of 80% i.e., in 100 records 80 records were a part of the training set and remaining 20 records were a part of the test set.

4. Model training: After a data scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entails ―feeding‖ the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data an answer you want to get a predictive analysis. The purpose of model training is to develop a model. We trained our model using the random forest algorithm. On training the model it predicts the yield on giving the other attributes of the dataset as input.

5. Model evaluation and testing: The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That‘s the optimization of model parameters to achieve an algorithm‘s best performance.

Application Requirements

1. Users must be able to register as a new user.
2. Users should be able to login if they already have an account.
3. The location inputs must be read correctly by the application.
4. The weather Prediction algorithm must give accurate prediction of the average rainfall and average temperature.
5. The user should be able to give the required inputs like soil type and area.
6. All the modules of the application must work in a proper manner.
7. The yield prediction module should provide two options. One is if the user is familiar with what crop is to be grown and other is when the user is not sure.
8. The predictions must be accurate.
9. Users must be able to access the Fertilizers module as well.
10. The fertilizer module must help the farmers decide whether to use the fertilizers or not.
11. The user must be able to logout.

4.3 Non-Functional Requirements

Non-Functional Requirement (NFR) specifies the quality attribute of a software system. They judge the software system based on Responsiveness, Usability, Security, Portability and other non-functional standards that are critical to the success of the software system.

Failing to meet non-functional requirements can result in systems that fail to satisfy user needs. Non-functional Requirements allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. Example, the site should load in 3 seconds when the number of simultaneous users are > 10000. They specify the criteria that can be used to judge the operation of a system rather than specific behaviour. 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 Requirement, etc.

Benefits of Non-Functional Requirements:

- The nonfunctional requirements ensure the software system follows legal and compliance rules.
- They ensure the reliability, availability, and performance of the software system.
- They ensure good user experience and ease of operating the software.
- They help in formulating security policy of the software system.

Requirements

1. The access permissions for system data may only be changed by the system‘s data administrator.
2. Passwords shall never be viewable at the point of entry or at any other time.
3. Apps should be able to adapt themselves to increased usage or be able to handle more data as time progresses.
4. Application should be responsive to the user Input or to any external interrupt which is of highest priority and return back to the same state.

5. Users should be able to understand the flow of the App easily i.e. users should be able to use the App without any guideline or help from experts/manuals.
6. All the app data should be secured and be encrypted with minimum needs so that it‘s protected.
7. There should be a common plan where the user can access the application to install and look for regular updates to give feedback.
8. The application should be able to render it‘s layout to different screen sizes. Along with automatic adjustment of Font size and image rendering.
9. The application should run at a speed that is desirable by the users. A slow application can lead to frustration and hence, will not be preferred over other faster applications.
10. The application must be stable. It should never crash or force close in the case of many users using it simultaneously.
11. The application must be easy to maintain.
12. It must be user-friendly. Having a user-friendly application is of key importance for the success of the application.

4.4 Python Libraries:

Normally, a library is a collection of books or is a room or place where many books are stored to be used later. Similarly, in the programming world, a library is a collection of precompiled codes that can be used later on in a program for some specific well-defined operations. Other than pre-compiled codes, a library may contain documentation, configuration data, message templates, classes, and values, etc.

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don’t need to write the same code again and again for different programs. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc.

Working of Python Library

As is stated above, a Python library is simply a collection of codes or modules of codes that we can use in a program for specific operations. We use libraries so that we don't need to write the code again in our program that is already available. But how it works. Actually, in the MS Windows environment, the library files have a DLL extension (Dynamic Load Libraries). When we link a library with our program and run that program, the linker automatically searches for that library. It extracts the functionalities of that library and interprets the program accordingly. That's how we use the methods of a library in our program. We will see further, how we bring in the libraries in our Python programs.

Python standard library

The Python Standard Library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. Most of the Python Libraries are written in the C programming language. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. Python Standard Library plays a very important role. Without it, the programmers can't have access to the functionalities of Python. But other than this, there are several other libraries in Python that make a programmer's life easier. Let's have a look at some of the commonly used libraries:

1. **Pandas:** Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.
2. **Numpy:** The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like

TensorFlow use Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

3. **Scikit-learn:** It is a famous Python library to work with complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc. This library works in association with Numpy and SciPy.

4. **Streamlit:** Streamlit is a open source python library for building fast, performant and beautiful data apps. It is a great tool to develop interactive apps quickly for demonstrating a solution. It turns data scripts into shareable web apps in minutes. No front-end experience required.

   Streamlit is an open-source python framework for building web apps for

Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app. Streamlit builds upon three simple principles.

- Embrace Scripting
- Weave in interaction
- Deploy instantly

We can build a streamlit in 4 simple steps:

1. Make sure you have python 3.6+ installed in your system.
2. Install streamlit using PIP:
   ```
   $pip install streamlit
   ```
3. Create a python script by naming it hello.py , enter the following code and save it.
   ```
   Import streamlit as stst.title('Hello World')
   st.write('Welcome to my first app.')
   ```
4. Run the app by running the following command in a terminal:
   ```
   $streamlit run hello
   ```
   we can install the streamlit by using following commands:
   ```
   $pip install source streamlit
   $pip install streamlit==0.62.0
   ```

5. **Django:** Django is an extremely popular and fully featured server-side web framework, written in Python. This module shows you why Django is one of the most popular web server frameworks, how to set up a development environment, and how to start using it to create your own web applications.

Django is a Python-based web framework, free and open-source, that follows the model–template–views (MTV) architectural pattern. It is maintained by the Django Software Foundation (DSF), an independent organization established in the US.

Django's primary goal is to ease the creation of complex, database-driven websites. The framework emphasizes reusability and "pluggability" of components, less code, low coupling, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings, files, and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models.

Components:

- Lightweight and standalone web server for development and testing
- A form serialization and validation system that translate between HTML forms and values suitable for storage in the database.
- A template system that utilizes the concept of inheritance borrowed from object-oriented programming.
- A caching framework that can use any of several cache methods.
- Support for middleware classes that can intervene at various stages of request processing and carry out custom functions.
- An internal dispatcher system that allows components of an application to communicate events to each other via pre-defined signals.
- An internationalization system, including translations of Django's own components into a variety of languages.
- A serialization system that can produce read XML and/or read XML and/or JSON representations of Django model instances.
- A system for extending the capabilities of the template engine.

- An interface to Python's built-in unit test framework.

Django can be run in conjunction with Apache, Nginx using WSGI, Gunicorn, or Cherokee using flup (a Python module). Django also includes the ability to launch a FastCGI server, enabling use behind any web server which supports FastCGI, such as Lighttpd or Hiawatha. It is also possible to use other WSGI-compliant web Servers. Backends: PostgreSQL, MySQL, MariaDB, SQLite, and Oracle. Microsoft SQL Server can be used with django-mssql while similarly external backends exist for IBM Db2, SQL Anywhere and Firebird. There is a fork named django-nonrel, which supports NoSQL databases, such as MongoDB and Google App Engine's Datastore.

Django may also be run in conjunction with Jython on any Java EE application server such as GlassFish or JBoss. In this case django-jython must be installed in order to provide JDBC drivers for database connectivity, which also can provide functionality to compile Django in to a .war suitable for deployment.

Google App Engine includes support for Django version 1.x.x as one of the bundled frameworks.

Use of Libraries in Python Program

As we write large-size programs in Python, we want to maintain the code’s modularity. For the easy maintenance of the code, we split the code into different parts and we can use that code later ever we need it. In Python, modules play that part. Instead of using the same code in different programs and making the code complex, we define mostly used functions in modules and we can just simply import them in a program wherever there is a requirement. We don’t need to write that code but still, we can use its functionality by importing its module. Multiple interrelated modules are stored in a library. And whenever we need to use a module, we import it from its library. In Python, it’s a very simple job to do due to its easy syntax. We just need to use **import**.

4.5 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 that 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

Fig 4.2: Processor

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.

Fig 4.3: Ethernet Connection

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 ofnon volatile storage, retaining stored data even when powered off. 32 GB or higher is recommended for the proposed system.

Fig 4.4: Hard Disk

4. 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.

Fig 4.5: RAM

4.6 Software Requirements

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.
- **Notebook documents:** Self- contained documents that contain a representation of all content visible in the notebook web application, including inputs and outputs of the computations, narrative text, equations, images, and rich media representations of objects. Each notebook document has its own kernel.

Fig 4.6: Jupter notebook icon

2. 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

Fig 4.7: Python icon

3. Pycharm: Py Charm 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

Fig 4.8: Pycharm image

3.Visual Studio Code

Visual Studio Code, also commonly referred to as VS Code, isa source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality. Visual Studio Code is a source-code editor that can be used with a variety of languages, including Java, JavaScript, Go, Node.js, Python, C++. It is based on the Electron framework, which is used to develop Node.js Web applications that run on the Blink layout engine. Visual Studio Code employs the same editor component (codenamed "Monaco") used in Azure DevOps (formerly called Visual Studio Online and Visual Studio Team Services).

Out of the box, Visual Studio Code includes basic support for most common programming languages. This basic support includes syntax highlighting, bracket matching, code folding, and configurable snippets. Visual Studio Code also ships with IntelliSense for JavaScript, TypeScript, JSON, CSS, and HTML, as well as debugging support for Node.js. Support for additional languages can be provided by freely available extensions on the VS Code Marketplace.

Visual Studio Code can be extended via extensions, available through a central repository. This includes additions to the editor and language support.

A notable feature is the ability to create extensions that add support for new languages, themes, and debuggers, perform static code analysis, and add code linters using the Language Server Protocol.

Source control is a built-in feature of Visual Studio Code. It has a dedicated tab inside of the menu bar where you can access version control settings and view changes made to the current project. To use the feature you must link Visual Studio Code to any supported version control system (Git, Apache Subversion, Perforce, etc.).

This allows you to create repositories as well as to make push and pull requests directly from the Visual Studio Code program. Visual Studio Code includes multiple extensions for FTP, allowing the software to be used as a free alternative for web development. Code can be synced between the editor and the server, without downloading any extra software.

Visual Studio Code allows users to set the code page in which the active document is saved, the newline character, and the programming language of the active document. This allows it to be used on any platform, in any locale, and for any given programming language.

Working with Python in Visual Studio Code, using the Microsoft Python extension, is simple, fun, and productive. The extension makes VS Code an excellent Python editor, and works on any operating system with a variety of Python interpreters. It leverages all of VS Code's power to provide auto complete and IntelliSense, linting, debugging, and unit testing, along with the ability to easily switch between Python environments, including virtual and conda environments.

Autocomplete and IntelliSense

The Python extension supports code completion and IntelliSense using the currently selected interpreter. IntelliSense is a general term for a number of features, including intelligent code completion (in-context method and variable suggestions) across all your files and for built-in and third-party modules. IntelliSense quickly shows methods, class members, and documentation as you type, and you can trigger completions at any time with ctrl+Space. You can also hover over identifiers for more information about them.

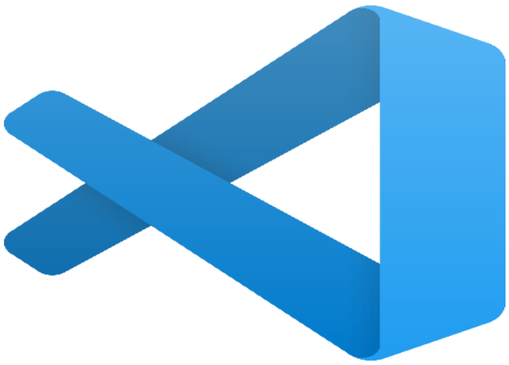


Fig 4.9: Visual Studio Code image

CHAPTER 5

SYSTEM ANALYSIS AND DESIGN

Systems development is a systematic process which includes phases such as planning, analysis, design, deployment, and maintenance. System Analysis is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose. Analysis specifies what the system should do.

System Design is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently. System Design focuses on how to accomplish the objective of the system.

5.1 UML DIAGRAMS:

UML represents Unified Modelling Language. UML is an institutionalized universally useful showing dialect in the subject of article situated programming designing. The fashionable is overseen, and become made by way of, the Object management Group.

The goal is for UML to become a regular dialect for making fashions of item arranged PC programming. In its gift frame UML is contained two noteworthy components: a Meta-show and documentation. Later on, a few type of method or system can also likewise be brought to; or related with, UML. The Unified Modeling Language is a popular dialect for indicating, Visualization, Constructing and archiving the curios of programming framework, and for business demonstrating and different non-programming frameworks. The UML speaks to an accumulation of first-rate building practices which have verified fruitful in the showing of full-size and complicated frameworks. The UML is a essential piece of creating gadgets located programming and the product development method. The UML makes use of commonly graphical documentations to specific the plan of programming ventures.

GOALS:

The Primary goals inside the plan of the UML are as in step with the subsequent:

1. Provide clients a prepared to utilize, expressive visual showing Language on the way to create and change massive models.
2. Provide extendibility and specialization units to make bigger the middle ideas.
3. be free of specific programming dialects and advancement manner.
4. Provide a proper cause for understanding the displaying dialect.
5. Encourage the improvement of OO gadgets exhibit.
6. Support large amount advancement thoughts, for example, joint efforts, systems, examples and components.
7. Integrate widespread procedures.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

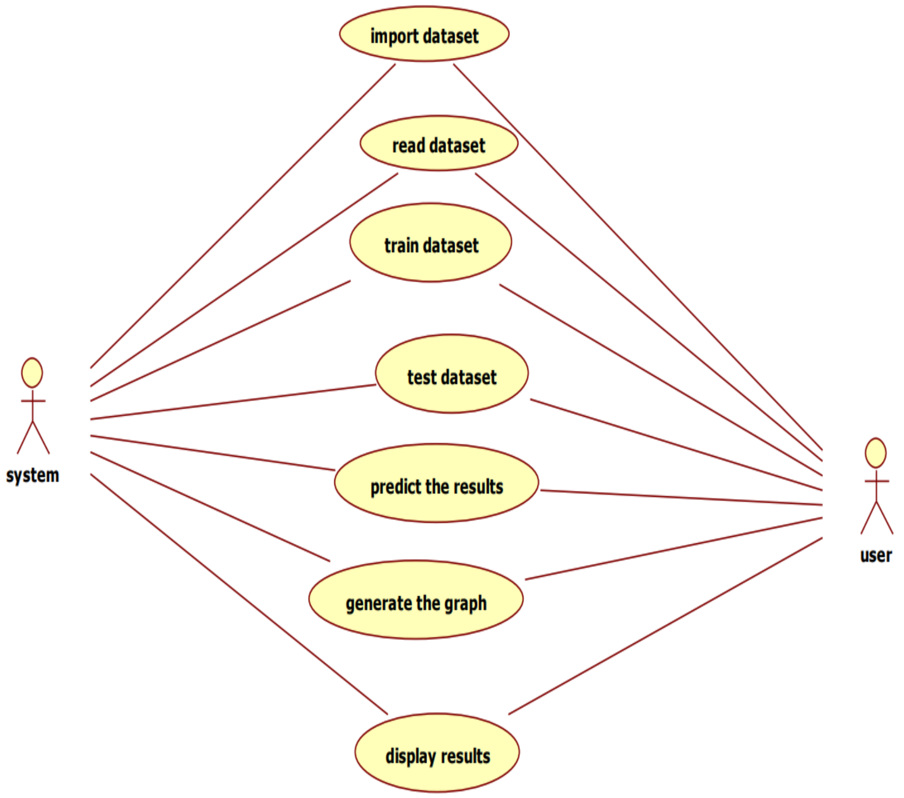


Fig 5.1: Usecase Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

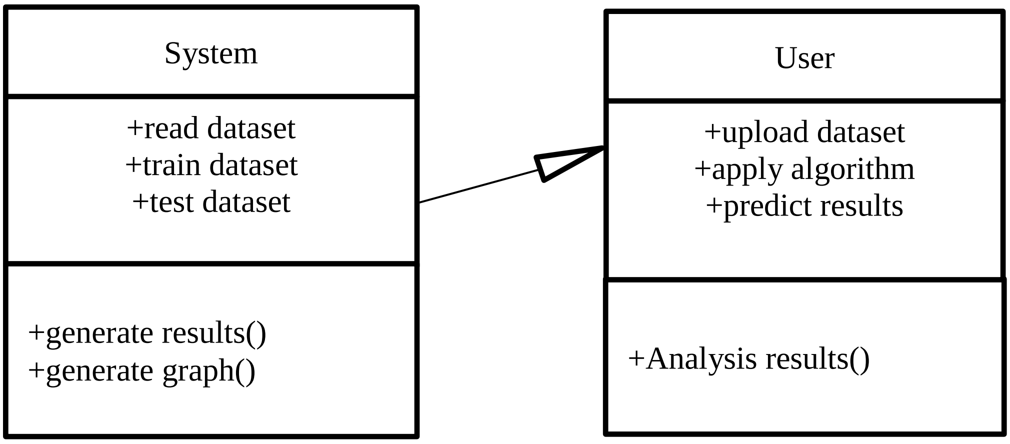


Fig 5.2: Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

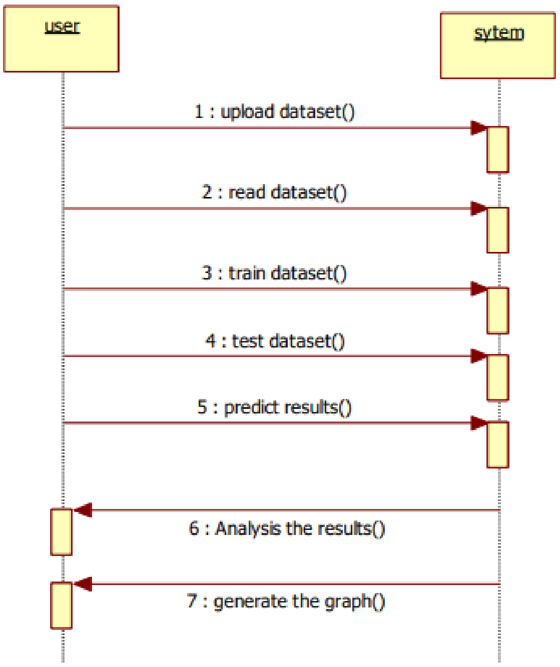


Fig 5.3: Sequence Diagram

COLLABORATION DIAGRAM:

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization where as the collaboration diagram shows the object organization.

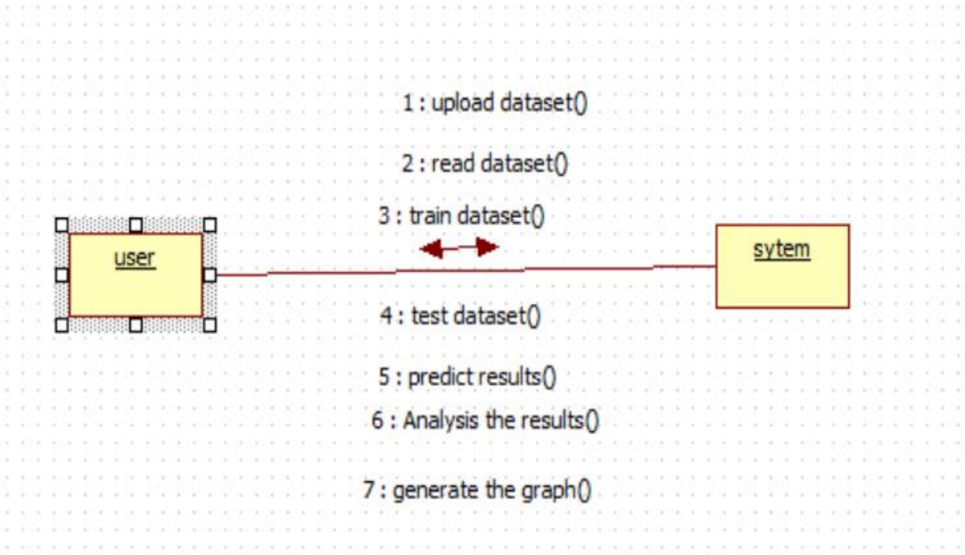


Fig 5.4: Collaboration Diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

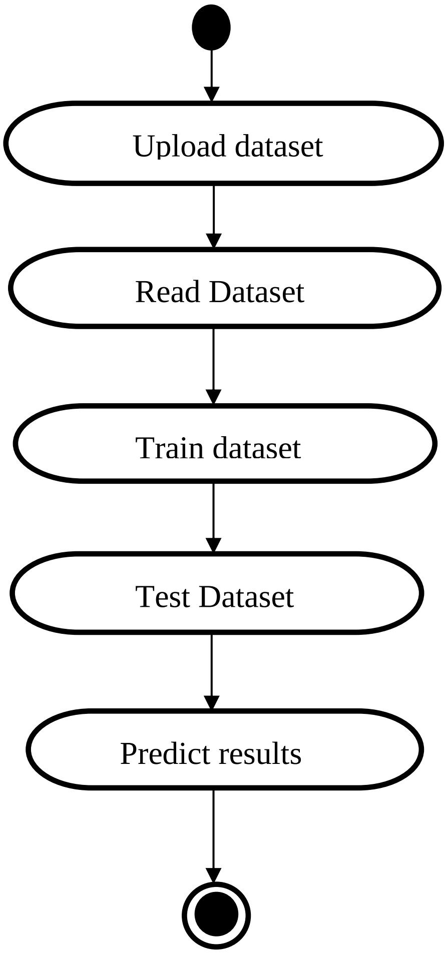


Fig. 5.5: Activity Diagram

5.1.1 Usage of UML in Project

As the strategic value of software increases for many companies, the industry looks for techniques to automate the production of software and to improve quality and reduce cost and time to the market. These techniques include component technology, visual programming, patterns and frameworks. Additionally, the development for the World Wide Web, while making somethings simpler, has exacerbated these architectural problems. The UML was designed to respond to these needs. Simply, systems design refers to the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements which can bed one easily through UML diagrams

5.2 System Architecture

Architecture diagrams can help system designers and developers visualize the high-level, overall structure of their system or application for the purpose of ensuring the system meets their users' needs. They can also be used to describe patterns that are used throughout the design. It's somewhat like a blueprint that can be used as a guide for the convenience of discussing, improving, and following among a team.

5.3 Flowchart

A flowchart is simply a graphical representation of steps. It shows steps in sequential order and is widely used in presenting the flow of algorithms, workflow or processes. Typically, a flowchart shows the steps as boxes of various kinds, and their order by connecting them with arrows. It originated from computer science as a tool for representing algorithms and programming logic but had extended to use in all other kinds of processes. Nowadays, flowcharts play an extremely important role in displaying information and assisting reasoning. They help us visualize complex processes, or make explicit the structure of problems and tasks. A flowchart can also be used to define a process or project to be implemented.

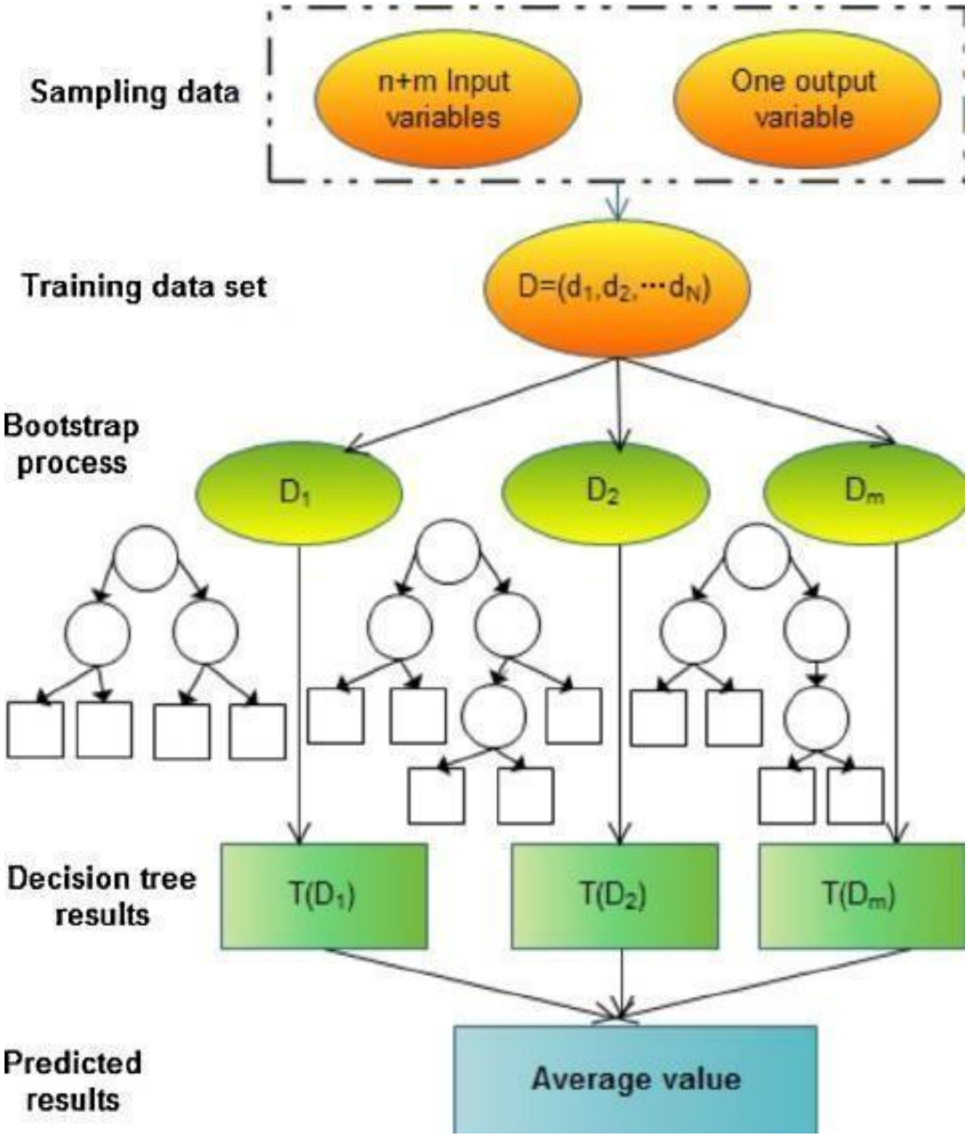


Fig 5.6: Working process of data

CHAPTER 6

IMPLEMENTATION

The implementation of the project was divided into two .i.e. crop yield prediction and rainfall prediction (for fertilizers module).

Crop Yield Prediction

This module returns the predicted production of crops based on the user's input. If the user wants to know the production of a particular crop, the system takes the crop as the input as well.

Else, it returns a list of crops along with their production as output.

These are the following steps of the algorithm implemented:

● **Step 1** : Choose the functionality i.e., crop prediction or yield prediction.

● **Step 2** : If the user chooses crop prediction:-

o Take soil type and area as inputs.

o These values are given as input to the random forest implementation in the backend and the corresponding predictions are returned.

o The algorithm returns a list of crops along with their production predicted.

● **Step 3** : If the user chooses yield prediction:-

o Take crop, soil type and area as inputs.

o These values are given as input to the random forest implementation in the backend and the corresponding crop yield prediction is returned.

o The algorithm returns the predicted production of the given crop.

Fertilizers Module

This module is used to suggest the farmer on usage of fertilizer based on the rainfall in next few days. To predict the rainfall for the next 15 days we are using an API service provided by OpenWeather. If it is likely to rain we suggest the farmer not to use the fertilizer.

These are the following steps of the algorithm implemented:

● **Step 1**: On selection of this module, API call is made to the Open Weather Services.

● **Step 2**: The user is required to give inputs of climatic conditions.

● **Step 3**: The system will suggest the suitable fertilizer.

6.1 Datasets

Machine Learning depends heavily on data. It‘s the most crucial aspect that makes algorithm training possible. It uses historical data and information to gain experiences. The better the collection of the dataset, the better will be the accuracy.

The first step is Data Collection. For this project, we require two datasets. One for modelling the yield prediction algorithm and other for predicting weather .i.e. Average Rainfall and Average Temperature. These two parameters are predicted so as to be used as inputs for predicting the crop yield. The sources of our datasets are: https://en.tutiempo.net/‘for weather data and https://www.kaggle.com/srinivas1/agricuture-crops-production-in-india‘ for crop yield data.

The yield prediction module dataset requires the following columns: State, District, Crop, Season, Average Temperature, Average Rainfall, Soil Type, Area and Production as these are the major factors that crops depend on. Production‘ is the dependent variable or the class variable. There are eight independent variables and 1 dependent variable. We achieved this by merging the datasets. The datasets were merged taking the location as the common attribute in both. We are considering only two states here, Maharashtra & Karnataka as the suicide rates in farmers in these two States were found to be very high.

CSV file:

The dataset used in this project is a .csv file. In computing, a comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. A CSV file stores tabular data (numbers and text) in plaintext. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. CSV is a simple file format used to store tabular data, such as a spreadsheet or database. Files in the CSV format can be imported to and exported from programs that store data in tables, such as Microsoft Excel or Open Office Calc. Its data fields are most often separated, or delimited, by a comma Computer Science and Engineering, SRIT Bitcoin Price Prediction Page 55 of 66 A CSV is a comma-separated values file, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a .csv extension. CSV files can be used with most any spreadsheet program, such as Microsoft Excel or Google Spreadsheets. The difference between CSV and XLS file formats is that CSV format is a plain text format in which values are separated by commas (Comma Separated Values), while XLS file format is an Excel Sheets binary file format which holds information about all the worksheets in a file, including both content and formatting.

Exploratory Data Analysis

It is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. It is also about knowing your data, gaining a certain amount of familiarity with the data, before one starts to extract insights from it. The idea is to spend less time coding and focus more on the analysis of data itself. After the data has been collected, it undergoes some processing before being cleaned and EDA is then performed. After EDA, go back to processing and cleaning of data, i.e., this can be an iterative process. Subsequently, use the cleaned dataset and knowledge from EDA to perform modelling and reporting. Exploratory data analysis is generally cross- classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate). It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand. EDA can give us the following:

- Preview data
- Check total number of entries and column types using in built functions. It is a good practice to know the columns and their corresponding data types.
- Check any null values.
- Check duplicate entries.
- Plot count distribution of categorical data.

Using various built in functions, we can get an insight of the number of values in each column which can give us information about the null values or the duplicate data. We can also find the mean, standard deviation, minimum value and the maximum value. This is the basic procedure of EDA. To get a better insight of the data that is being used, we can plot graphs like the correlation matrix which is one of the most important concept which gives us a lot of information about how variables (columns) are related to each other and the impact each of them have on the other.

A few other graphs like box plot and distribution graphs can be plotted too.

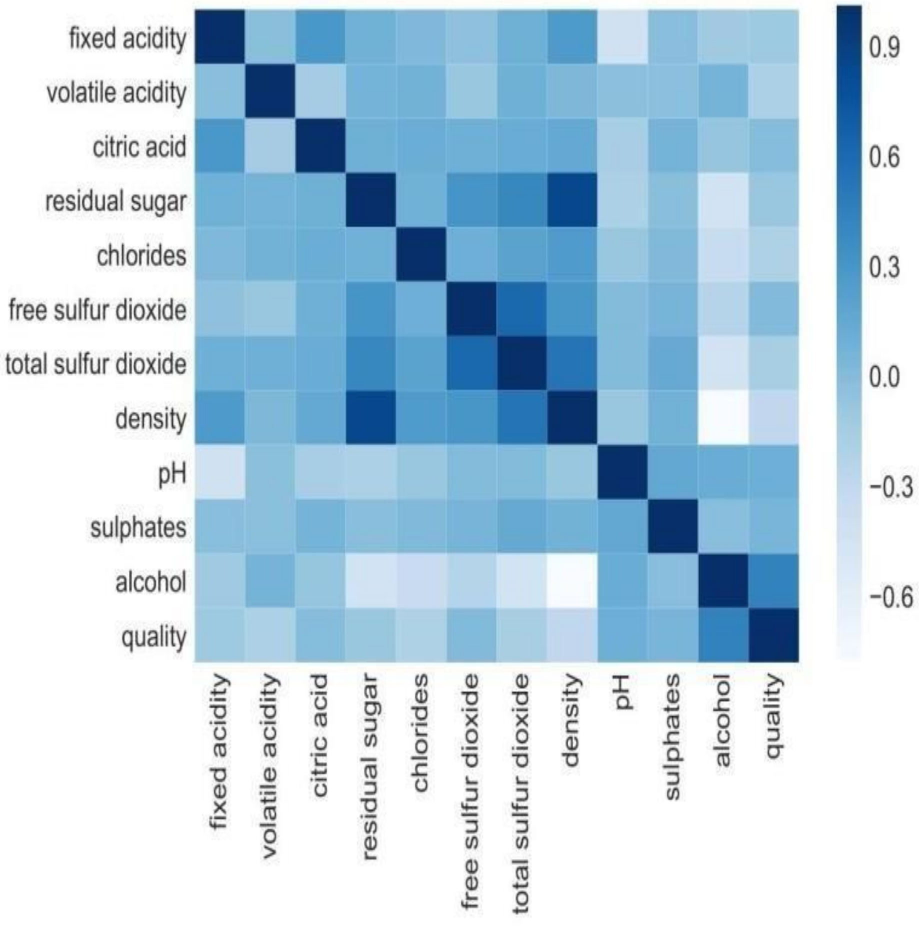


Fig 6.1: Example of Correlation matrix

Dark shades represent positive correlation while lighter shades represent negative correlation.

Hence, we can make the following inferences from the above example:

- Here we can infer that density has strong positive correlation with residual sugar‖ whereas it has strong negative correlation with alcohol.
- Free sulphur dioxide and citric acid‖ have almost no correlation with quality.

- Since correlation is zero we can infer there is no linear relationship between these two predictors.

6.2 Data Pre-Processing

Data Pre-Processing is a Data Mining method that entails converting raw data into a format that can be understood. Real-world data is frequently inadequate, inconsistent, and/or lacking in specific activities or trends, as well as including numerous inaccuracies. This might result in low-quality data collection and, as a result, low-quality models based on that data. Preprocessing data is a method of resolving such problems. Machines do not comprehend free text, image, or video data; instead, they comprehend 1s and 0s. So putting on a slideshow of all our photographs and expecting our machine learning model to learn from it is probably not going to be adequate. Data Pre-processing is the step in any Machine Learning process in which the data is changed, or encoded, to make it easier for the machine to parse it. In other words, the algorithm can now easily interpret the data's features. Data Pre-processing can be done in four different ways. Data cleaning/cleaning, data integration, data transformation, and data reduction are the four categories.

6.2.1 Data Cleaning:

Data in the real world is frequently incomplete, noisy, and inconsistent. Many bits of the data may be irrelevant or missing. Data cleaning is carried out to handle this aspect. Data cleaning methods aim to fill in missing values, smooth out noise while identifying outliers, and fix data discrepancies. Unclean data can confuse data and the model. Therefore, running the data through various Data Cleaning/Cleansing methods is an important Data Pre-processing step.

6.2.2 Data Integration:

It is involved in a data analysis task that combines data from multiple sources into a coherent data store. These sources may include multiple databases. Do you think how data can be matched up?? For a data analyst in one database, he finds Customer_ID and in another he finds cust_id, How can he sure about them and say these two belong to the same entity.

Databases and Data warehouses have Metadata (It is the data about data) it helps in avoiding errors.

6.2.3 Data Reduction :

Because data mining is a methodology for dealing with large amounts of data. When dealing with large amounts of data, analysis becomes more difficult. We employ a data reduction technique to get rid of this. Its goal is to improve storage efficiency while lowering data storage and analysis expenses.

Dimensionality Reduction

A huge number of features may be found in most real-world datasets. Consider an image processing problem: there could be hundreds of features, also known as dimensions, to deal with. As the name suggests, dimensionality reduction seeks to minimize the number of features but not just by selecting a sample of features from the feature set, which is something else entirely Feature Subset Selection or feature selection.

Numerosity Reduction

Data is replaced or estimated using alternative and smaller data representations such as parametric models (which store only the model parameters rather than the actual data, such as Regression and Log-Linear Models) or non- parametric approaches.

CHAPTER 7
TESTING

Software testing is an investigation conducted to provide stakeholders with information about the quality of the software product or service under test. Software testing can also provide an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation. Test techniques include the process of executing a program or application with the intent of finding software bugs (errors or other defects), and verifying that the software product is fit for use.

Software testing involves the execution of a software component or system component to evaluate one or more properties of interest. In general, these properties indicate the extent to which the component or system under test:

- Meets the requirements that guided its design and development,
- Responds correctly to all kinds of inputs,
- Performs its functions within an acceptable time,
- It is sufficiently usable,
- Can be installed and run in its intended environments, and
- Achieves the general result its stakeholder‘s desire.

7.1 Functionality Testing

- Database connection is successfully established.
- The flow of the application from one page to another is correct, accurate and quick.
- All the forms included in the application are working as expected.
- Proper alert messages are displayed in case of wrong inputs.
- After every action on the application the appropriate data is fetched from the backend.

7.2 Usability Testing

- The application enables smooth navigation, hence gives a user-friendly experience.
- The inputs taken from the user are via dropdown hence correct inputs

are provided to the system.

- Wrong inputs given by the system are handled effectively.
- The content provided by the application is verified and is taken by the trusted sources.
- The datasets trained for prediction of the crop yield are accurate and balanced.

7.3 Interface Testing

- The application connects correctly with the server. In case of failure an appropriate message is displayed.
- Interruptions by the server or by the user are handled efficiently.
- If the user enters wrong credentials or invalid email id, the application handles it efficiently by displaying appropriate messages.
- The interaction with the user is smooth and easy.

7.4 Compatibility Testing

- This application is compatible with all the browsers enabled with javascript.
- It is compatible with all the mobile devices and desktop.

7.5 Performance Testing

- It works fine with moderate internet speed.
- The connection is secured and user details are stored in a secured manner.
- The switch from one screen to another is quick and smooth.
- The inputs from users are taken correctly and response is recorded quickly.

7.6 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.

7.7 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.

7.8 System Testing

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.

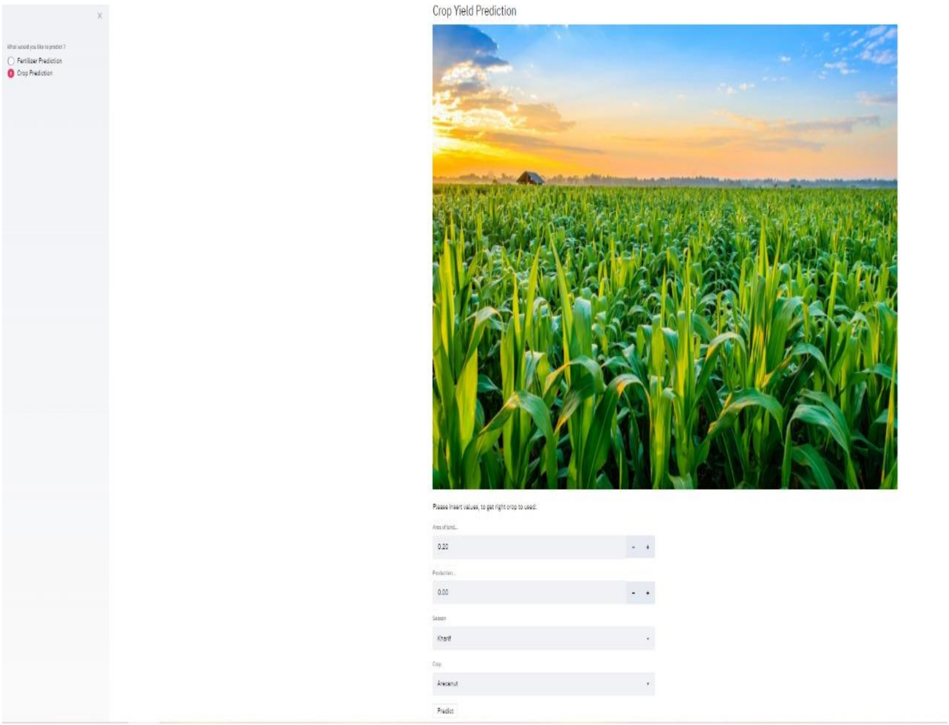
7.9 White Box Testing

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.

CHAPTER 8

RESULTS

In the final implementation of the application the first screen the user can view is the design page that contain two functionalities Fertilizer analysis and Crop yield Prediction.



What would you like to predict ?

○ Fertilizer Prediction

● Crop Prediction

Crop Yield Prediction

Please insert values, to get right crop to used:

Area of land...: 0.20

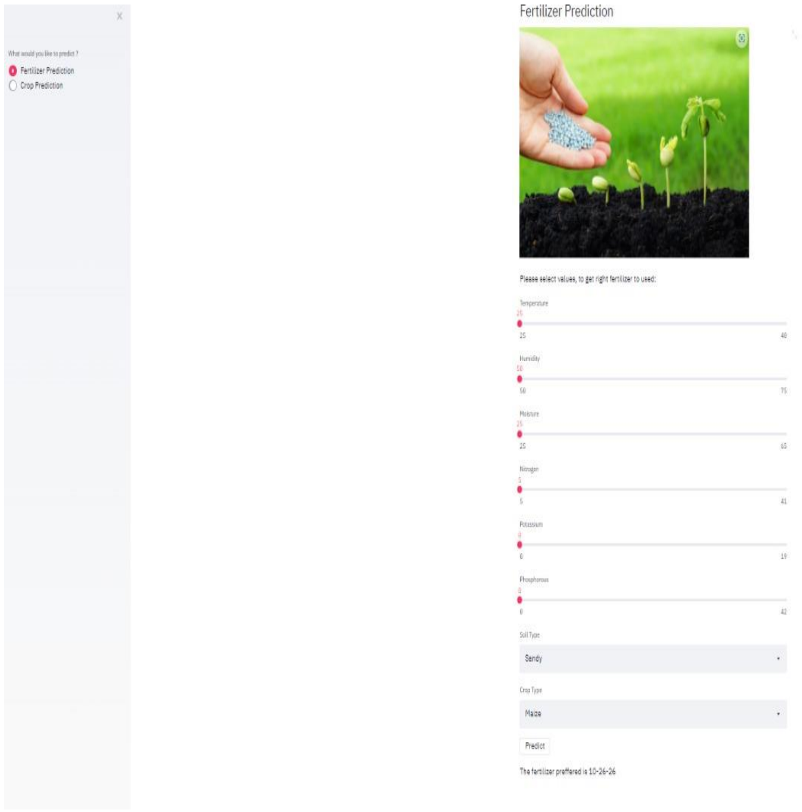
Production...: 0.00

Season: Kharif

Crop: Arecanut

Fig 8.1: Initial output page of crop yield prediction

The above figure represents the Initial web page of the system that has two functionalities that are Crop yield prediction and fertilizer analysis. On selecting the required one we get the new webpage and required to give the inputs.



What would you like to predict ?

● Fertilizer Prediction

○ Crop Prediction

Fertilizer Prediction

Please select values, to get right fertilizer to used:

Temperature: 25 (range 25–40)

Humidity: 50 (range 50–75)

Moisture: 25 (range 25–65)

Nitrogen: 5 (range 5–41)

Potassium: 0 (range 0–19)

Phosphorous: 0 (range 0–42)

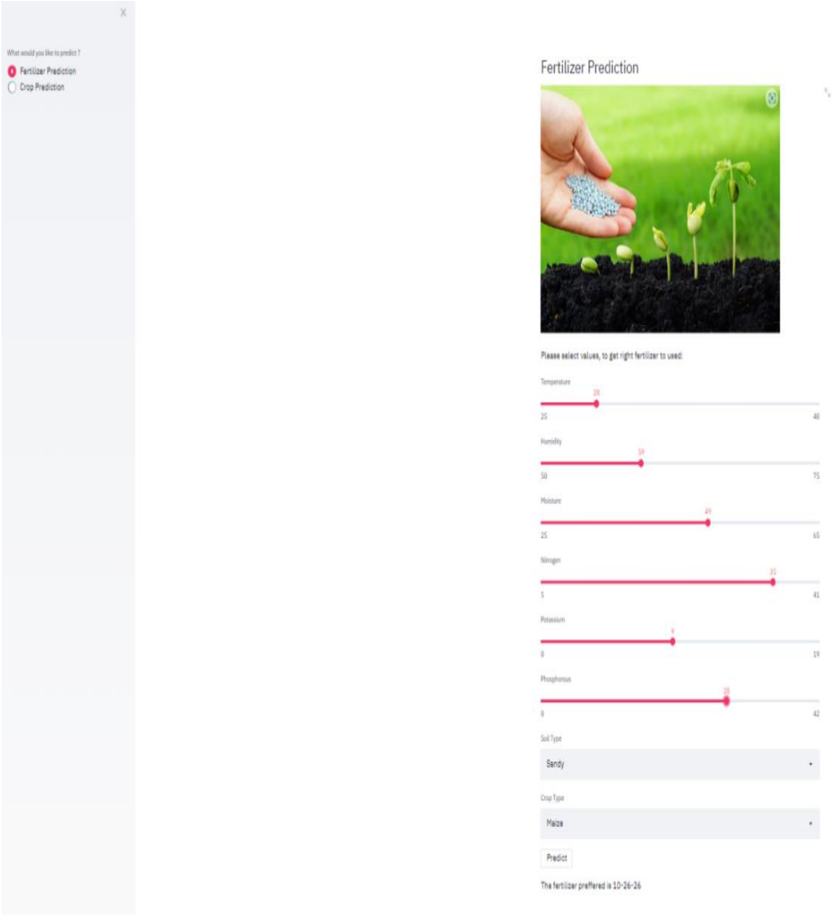
Soil Type: Sandy

Crop Type: Maize

The fertilizer preffered is 10-26-26

Fig 8.2: Initial page of Fertilizer Prediction

The above represents the initial page of the Fertilizer analysis system. Here we are required to give the inputs of temperature, humidity, moisture, nitrogen, phosphorous, potassium, soil-type, crop-type to suggest the suitable fertilizer.



What would you like to predict ?

● Fertilizer Prediction

○ Crop Prediction

Fertilizer Prediction

Please select values, to get right fertilizer to used:

Temperature: 28 (25–40)

Humidity: 59 (50–75)

Moisture: 49 (25–65)

Nitrogen: 35 (5–41)

Potassium: 9 (0–19)

Phosphorous: 28 (0–42)

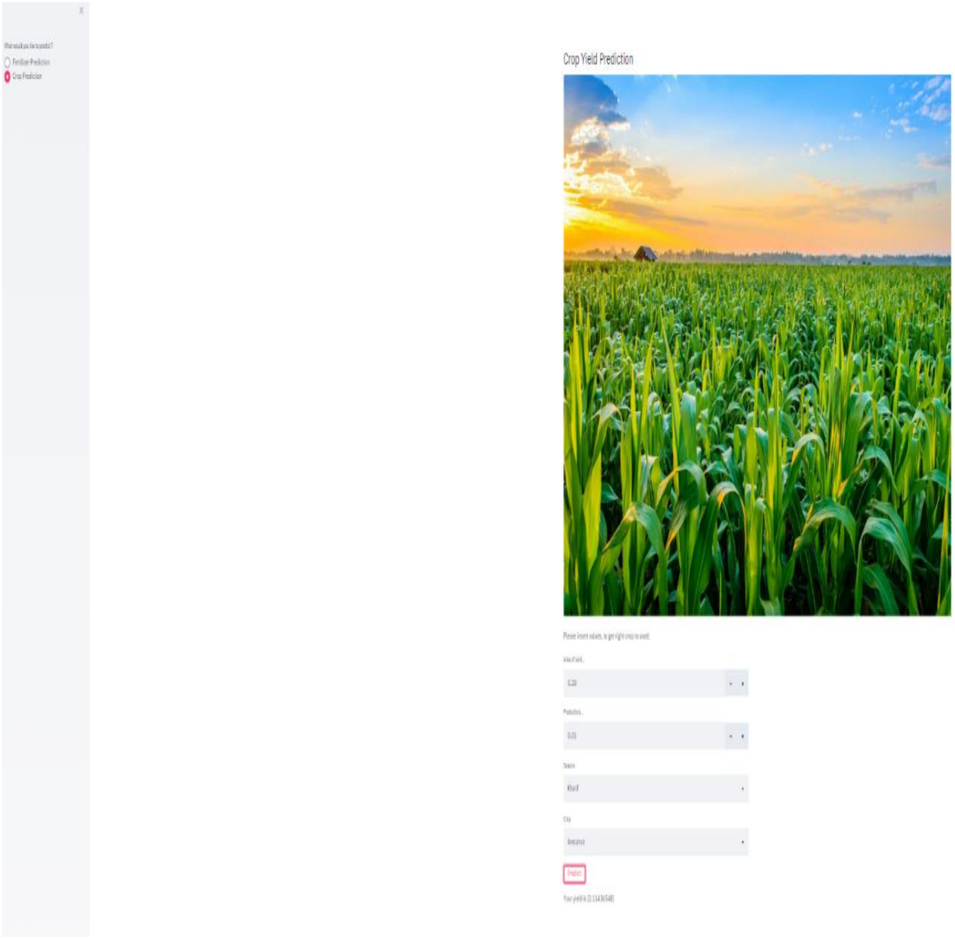
Soil Type: Sandy

Crop Type: Maize

The fertilizer preffered is 10-26-26

Fig 8.3: Final output page of Fertilizer Prediction

The figure represents the inputs using slide bars with initial and final values, and user is required to give inputs and predict the suitable fertilizer for the crop.



What would you like to predict?

○ Fertilizer Prediction

● Crop Prediction

Crop Yield Prediction

Please insert values, to get right crop to used.

Area of land...

0.20

Production...

0.01

Season

Kharif

Crop

Arecanut

Your yield is [2.11436548]

Fig 8.4: Final output page of Crop Yield Prediction

The above figure the represents the required inputs from the user and after giving the inputs it predicts the yield of the crop.

CONCLUSION

By classifying the soil based on the soil type, land type and macro nutrients Nitrogen(N), Phosphorous(P) and Pottasium (K) present in the soil, along with Temperature, humidity, moisture will be considered. The prediction of crop yield based on location and proper implementation of algorithms have proved that the higher crop yield can be achieved. From the above works we conclude that Crop yield prediction and Fertilizer analysis system Random Forest is good with the accuracy of 95.7% and 96.1%.

The Crop yield prediction and suggesting the fertilizer is successfully predicted and also found the efficient algorithm and obtained the most efficient output of the yield and suitable fertilizer. By developing the web application based on this ideology and make the user use this easily and help the user to understand the yield of the crop, he is going to crop in that season and user can easily know the suitable fertilizer that can be used for that crop.

REFERENCES

[1] Rushika G., Juilee K, Pooja M, Sachee N, and Priya R.L.(2018). Prediction of Crop Yield using Machine Learning, Issue 02 IRJET (pg 2337-2339).

[2] Ruchita T, Shreya B, Prasanna D, and Anagha C (2017). Crop Yield Prediction using Big Data Analytics, Volume 6, Issue 11, IJCMS.

[3] Monali P, Santosh K, Vishwakarma, and Ashok V (2015). Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach, ICCICN 2015.

[4] Mrs. N. Hemageetha, and Dr. G.M. Nasira (2016). Analysis of Soil Condition based on pH value using Classification Techniques, IOSR-JCE, Issue 6, (pp.50-54).

[5] E. Manjula, and S. Djodiltachoumy (2017). Data Mining Technique to Analyze Soil Nutrients based on Hybrid Classification. IJARCS.

[6] Rohit Kumar Rajak, Ankit Pawar, Mitalee Pendke, Pooja Shinde, Suresh Rathod, and Avinash Devare (2017). Crop Recommendation System to maximize Crop yield using Machine Learning. Issue 12 IRJET.

[7] S. Veenadhari, Dr. Bharat Misra, Dr. CD singh (2014).Machine Learning Approach for Forecasting Crop Yield based on Climatic Parameters. ICCCI-2014

[8] Sadia A, Abu Talha K, Mahrin Mahia, Wasit A, and Rashedur M.R.(2018). Analysis of Soil Properties and Climatic Data To Predict Crop Yields and Cluster Different Agricultural Regions of Bangladesh, IEEE ICIS 2018 (pp.80-85).

[9] Arun K, Navin K, and Vishal V (2018). Efficient Crop Yield Prediction using Machine Learning, (pp.3151-3159), IRJET.

[10] Subhadra M, Debahuti M, Gour, H. S.(2016). Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper, Vol 9(38), DOI:10.17485/ijst/2016/v9i38/95032 IJST.

[11] Mrs. Prajakta P.B., Yogesh S. P, and Dinesh D.P. (2017). Improved Crop Yield Prediction using Neural Network, IJARIIE, (pg.3094-3101).

[12] Dhivya B H, Manjula R, Siva Bharathi S, and Madhumathi R (2017). A Survey on Crop Yield Prediction based on Agricultural Data,

DOI:10.15680/IJIRSET.2017.0603053, IJIRSET.

[13] Omkar B,Nilesh M,Shubham G,Chandan M, and D.S. Zingade (2017), Crop Prediction System using Machine Learning, Volume 4, Special Issue 5, IJAERD.

[14] S.Bhanumathi, M.Vineeth and N.Rohit (2019). Crop Yield Prediction and Efficient use of Fertilizers, (pg.0769- 0773), ICCSP.