A
Preliminary Project Report
On

CROP PREDICTION

SUBMITTED TOWARDS THE
FULFILLMENT OF THE REQUIREMENTS OF

Bachelor Of Engineering (Computer Engineering)

BY

Sohel Tamboli	Exam No: B150194269
Arbaj Shaikh	Exam No: B150194261
Dattatray Pawar	Exam No: B150194250
Shubham Gawade	Exam No:B150194214

Under The Guidance of

Prof. A. A. Chavan

Department Of Computer Engineering
SVPM's College Of Engineering,Malegaon(Bk.),Baramati,Pune-413115.

SAVITRIBAI PHULE PUNE UNIVERSITY
2020-21

SVPM's COLLEGE OF ENGINEERING, DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Entitled

CROP PREDICTION

Submitted by

Sohel Tamboli	Exam No: B150194269
Arbaj Shaikh	Exam No: B150194261
Dattatray Pawar	Exam No: B150194250
Shubham Gawade	Exam No:B150194214

is a bonafide work carried out by Students under the supervision of Prof. A. A. Chavan and it is submitted towards the fulfillment of the requirement of Bachelor of Engineering (Computer Engineering) Project.

Prof. A. A. Chavan
Internal Guide

Prof. H. R. Kumbhar
H.O.D

External Examiner

Dr. S.M.Mukane
Principal

Place : SVPM's COE Malegaon(Bk.)
Date :

PROJECT APPROVAL SHEET

A

Project Stage-I

on

(CROP PREDICTION)

Is successfully completed by

Sohel Tamboli (Exam NO: B150194269)
Arbaj Shaikh (Exam NO: B150194261)
Dattatray Pawar (Exam NO: B150194250)
Shubham Gawade (Exam NO: B150194214)

at

Department Of Computer Engineering
SVPM's College Of Engineering,Malegaon(Bk.), Baramati,Pune-413115.

SAVITRIBAI PHULE PUNE UNIVERSITY
2019-20

Prof. A. A. Chavan
Department of Computer Engg.

Prof. H. R. Kumbhar
Head

Abstract

Agricultural data is being produced constantly and enourmosly. As a result, agricultural data has come in the era of big data. Smart technologies contribute in data collection using electronic devices. In our project we are going to analyse and mine this agricultural data to get useful results using technologies like data analytics and machine learning and this result will be given to farmers for better crop yield in terms of efficiency and productivity.

Keywords—Big data, K-Means clustering, Apriori, Naive Bayes (key words), Agriculture

Acknowledgments

It gives us great pleasure in presenting the project report on **'CROP PREDICTION'**. We would like to express our special thanks of gratitude to our project guide **Prof. A. A. Chavan**.

I am also grateful to Prof. **H. R. Kumbhar** Head of Computer Engineering Department,SVPM COE for his indispensable support, suggestions and which also helped us to know about so many new things we are really thankful to them.

We would like to thanks our principle **Dr. S. M. Mukane**, College of Engineering, Malegaon(Bk.) who gave us the golden opportunity to work on this wonderful project.

Sohel Tamboli
Arbaj Shaikh
Dattatray Pawar
Shubham Gawade
(B.E. Computer Engg.)

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction

- As we know farmer is backbone of our country and agriculture is backbone of our Indian economy.

- Agriculture in India is popular for its vast diversity. It depends on climatic and weather conditions, soil components, manures used, resources available, and political and socio-economical factors

- Crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil.

- Parameter like pH , Nitrogen , Phosphate , Potassium, and also parameter related to atmosphere such as sunshine hours , rainfall , temperature.

- Agriculture field contains many data such as soil data, harvest data, and meteorological data, etc

1.2 Objectives

- The main objective of our system is to give idea about achieve maximum crop yield.
- Another main objective is to predict the yield of the crops using different parameters like rainfall, temperature, fertilizers, pesticides, ph level, and other atmospheric conditions and parameters.
- Maximum yielding at minimum cost and identifying and resolving the problems facing by the farmers and food security are major objective of this research work

1.3 Scope Of The Project

- The scope of the project is to determine the crop yield of an area by considering dataset with some features which are important or related to crop production such as temperature, moisture, rainfall, and production of the crop in previous years. To predict a continuous value, regression models are used. It is a supervised technique. The coefficients are preprocessed and fit into the trained data during training and construction the regression model. The main focus here is to reduce the cost function . The output function facilitates in error measurement. During training period, error between the predicted and actual values is reduced in order to minimize error function.

1.4 Motivation of The project

- In India many of peoples mainly dependent on the agriculture and agriculture related industries like post harvesting related industries, food transport industries, fertilisers production industries and so on.
- Agriculture plays a major role in increasing economy of the country. But to achieve this, farmers needs to resolve so many difficulties like crop yield prediction problems, crop loss due to floods, drought, sudden change in temperature, crop diseases infections and disease detection etc.
- The whole agricultural land of our country is divided into fifteen Agro-climatic regions based on type of land and type of crops to be cultivated in those respective regions, so based on these Agro-climatic region's agriculture data we can make the prediction and suggest it to farmers.

Chapter 2

LITERATURE SURVEY

2.1 Literature Survey

Ananthara, M. G. et al. (2013, February) proposed a prediction model for datasets pertaining to agriculture which is called as CRY algorithm for crop yield using beehive clustering techniques. They considered parameters namely crop type, soil type, soil pH value, humidity and crop sensitivity. Their analysis was mainly in paddy, rice and sugarcane yields in India. Their proposed algorithm was then compared with CR tree algorithm and it outperformed well with an accuracy of 90 percent [2]. Awan, A. M. et al. (2006, April) built a new, smart framework focused on farm yield prediction clustering kernel methodology and they considered parameters like plantation, latitude, temperature and precipitation of rainfall in that latitude. They had experimented weighted k-means kernel method with spatial constraints for the analysis of oil palm fields [3]. Chawla, I. et al. (2019, August) used fuzzy logic for crop yield prediction through statistical time series models. They considered parameters like rainfall and temperature for prediction. Their prediction was classification with levels 'good yield' , 'very good yield' [4]. Chaudhari, A. N. et al. (2018, August) used three algorithms namely clustering kmeans, Apriori and Bayes algorithm, then they hybridized the algorithm for better efficiency of yield prediction and they considered parameters like Area, Rainfall, Soil type and also their system was able to tell which crop is suitable for cultivation based on the mentioned features [5]. Gandge, Y. (2017, December) used many machine learning algorithms for different crops. They studied and analyzed which algorithm would be suitable for which crop. They have used K-means, Support vector Regression, Neural Networks, C4.5 Decision tree, Bee-Hive Clustering, etc. The factors implying were soil nutrients like N, K, P and soil ph. [6]. Armstrong, L. J. et al. (2016, July) used ANNs for the prediction of rice yield in the districts of Maharashtra, India. They considered climatic factors namely (considering range) temperature, precipitation and reference crop evapotranspiration. The records were collected from Indian Government repository from 1998 to 2002 [7]. Tripathy, A. K. et al. (2016, July) were same authors who used support vector machines to predict the rice crop yield with same features as the previous paper mentioned [8]. Petkar, O. (2016, July) were also the same authors who applied for SVM and neural networks for rice crop yield prediction proposed a new decision system which is an interface to give the input and get the output [9]. Chakrabarty, A. et al. (2018, December) analyzed crop prediction in the country of Bangladesh where they majorly cultivate three kinds of rice, Jute, Wheat, and Potato. Their research used a deep neural network where the data had around 46 parameters into their consideration. Few of them were soil composition, type of fertilizer, type of soil and its structure, soil consistency, reaction and texture [10]. Jintrawet, A. et al. (2008, May) used SVR model for crops like rice to predict the yield where the model was divided into three steps- predicting the soil nitrogen weight followed by prediction of rice stem weight and rice grain weight respectively. Their factors were solar radiation, temperature and precipitation along with those three steps [11]. Miniappan, N. et al. (2014, August) used artificial neural network in modelling multi-layer perceptron model with 20 hidden layers for prediction wheat yield which considered factors like sunlight, rain, frost and temperature [12]. Manjula, A et al. built a crop selection and to predict the yield which considered various indexes like vegetation, temperature and normalized difference vegetation as factors. They distinguished between climate factors and agronomic factors and other disturbances caused in the prediction for better understanding [13]. Mariappan, A. K. et al. analyzed the data regarding rice crop in the state of Tamil Nadu, India. They have considered factors like soil, temperature, sunshine, rainfall, fertilizer, paddy, and type of pest used and other factors like pollution and season [14]. Verma, A. et al. (2015, December) used classification techniques like Naïve Bayes, K-NN algorithm for crop prediction on soil datasets which constituted nutrients of soil like zinc, copper, manganese, pH, iron, Sulphur, Phosphorous, Potassium, nitrogen, and Organic Carbon [15]. Kalbande, D. R. et al. (2018) used support vector regression, multi polynomial regression and random forest regression for prediction of corn yield and evaluated the models using metrics like errors namely MAE, RMSE and R-square values [16]. Rahman, R. M. et al. (2015, June) used mainly clustering techniques for crop yield prediction. The paper explained the analysis of major crops in Bangladesh and divided the variables into environmental and biotic variables. The algorithms applied were linear regression, ANN, and KNN approach for classification [17].

Hegde, M. et al. (2015, June) used multiple linear regression and neuro fuzzy systems for predicting crop yield by taking biomass, soil water, radiation and rainfall as input parameters for the research and their majorly concentrated crop was wheat [18]. Sujatha, R., Isakki, P. (2016, January) used classification techniques like ANN, j48, Naïve Bayes, Random Forest and Support vector Machines. They have also included both climatic parameters and soil parameters as features in their modelling [19]. Ramalatha, M. et al. (2018, October) used a hybrid approach of combining Kmeans clustering and classification based on modified K-NN approach. The data was collected from Tamil Nadu, India where the majorly concentrated crops were rice, maize, Ragi, Sugarcane, and Tapioca [20]. Singh, C. D. et al. (2014, January) developed an application to advise crops which works on selected districts of Madhya Pradesh, India. The user would give input cloud cover, rainfall, temperature, observed yield in the past and the system would predict the yield and Depending on the trigger values set, the crop will be labeled and obtain the results .

2.2 Proposed System

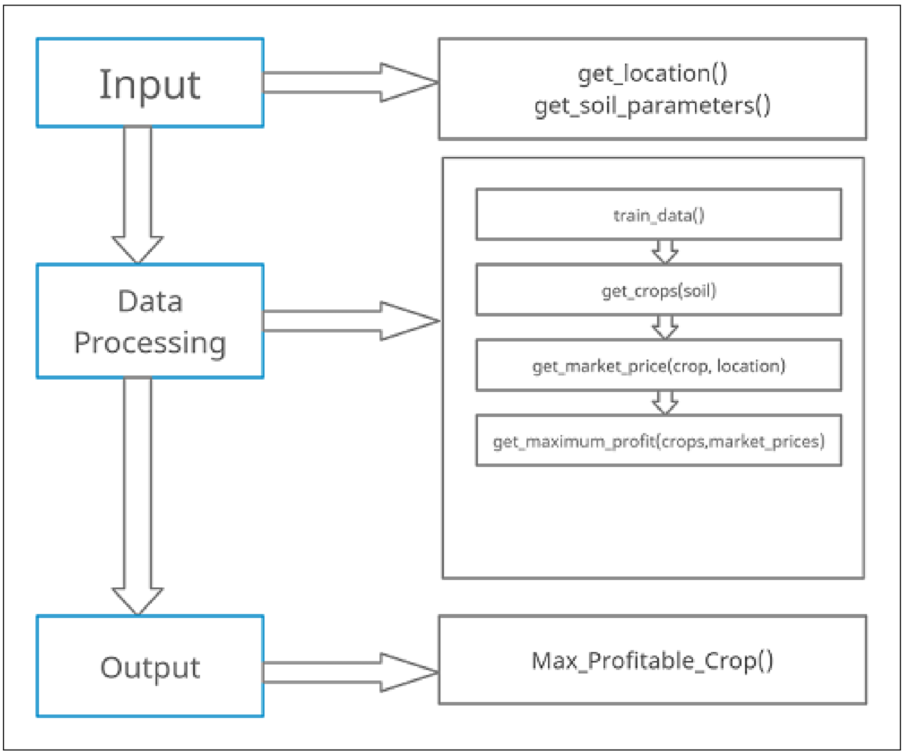


Figure 2.1: Praposed System

Input: The prediction of crop is dependent on numerous factors such as temperature, humidity, pH, rainfall in order to predict the crop accurately. All these factors are location reliant and thus the location of user is taken as an input to the system.

Data Processing: A crop can be cultivable only if apropos conditions are met. These include extensive parameters allied to soil and weather. These constraints are compared and the apt crops are ascertained.Random Forest Classifier is used to Predict the Crop and Decision Tree Regressor is used for Predict the Market Price of the Crop.

Output: The most profitable crop is predicted by the system.

Chapter 3

REQUIREMENT ANALYSIS

Software Requirement Specification

3.1 Introduction

3.1.1 Purpose

The aim of the system is to reduce the losses due to drastic climatic changes and increase the yield rates of crops. The system integrates the data obtained from the past prediction, current weather and soil condition due to this farmers gets the idea and list of crops that can be cultivated.

3.1.2 Scope

The scope of the project is to determine the crop yield of an area by considering dataset with some features which are important or related to crop production such as temperature, moisture, rainfall, and production of the crop in previous years. To predict a continuous value, regression models are used. It is a supervised technique. The coefficients are preprocessed and fit into the trained data during training and construction the regression model. The main focus here is to reduce the cost function . The output function facilitates in error measurement. During training period, error between the predicted and actual values is reduced in order to minimize error function.

3.1.3 Definition

The Proposed system will predict the most suitable crop for particular land based on soil contents and weather parameters such as Temperature, Humidity, soil PH and Rainfall.

3.2 Overview

The system is depends on five sub point as discuss below:

3.2.1 Project management tool

Project management is the practice of initiating, planning, executing, controlling, and closing the work of a team to achieve specific goals and meet specific success criteria at the specified time. The primary challenge of project management is to achieve all of the project goals within the given constraints. **Project management consist have below point:**

3.2.1.1 Crop Management

Crop Management is a software system to communicate between Farmers and system.Farmers can get the Most Profitable Crop After Entering Soil Status and Location.

3.3 Overall Description

3.3.1 Product Perspective

3.3.1.1 System Interfaces

The system provide a GUI environment so that one can easily enter the information with the help of forms into the database. Application should be web based. Supported by any browser (mobile browser,pc browser).

3.3.1.2 User Interfaces

Home page where user can enter their crop and location details and get crop prediction.

3.3.1.3 Memory Constraints

The system requires a minimum of 2GB of primary memory and 500GB of secondary memory for installation and execution.

3.4 Specific Requirements

3.4.1 External Interface Requirements:

The system takes input from scanner, keyboard, and files in the memory. The system generates printable output on the screen and peripherals.

3.4.2 Performance requirements:

The system is required to support multiple terminals simultaneously. The system should handle reasonable number of users without break or inconsistency.

3.4.3 Design constraints:

Design constraints are those constraints that are imposed on the design solution, which in this example refers to the ESS design. These constraints are typically imposed by the customer, by the development organization, or by external regulations.

3.4.4 Software System attribute:

3.4.4.1 Reliability:

Designs are usually based on specifications. Reliability requirements are typically part of a technical specifications document. They can be requirements that a company sets for its product and its own engineers or what it reports as its reliability to its customers. They can also be requirements set for supplier.

3.4.4.2 Availability:

For the purposes of this project the person who needs a meal has provide a meal on time. The software is easily available to user and easy to use.

3.4.4.3 Maintainability:

Maintainability is the ease with which faults in a software system can be found and fixed. Maintainability requirements address the user concern for how easy it is to upkeep and repair the system.

Chapter 4

ALGORITHM ANALYSIS AND MATHEMATICAL MODELING

4.1 Random Forest

4.1.1 What is Random Forest Algorithm?

Random forest is a supervised learning algorithm. The ”forest” it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

4.1.2 How does Random Forest Algorithm work?

Random forest is a supervised learning algorithm. The ”forest” it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Put simply: 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. Let’s look at random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:

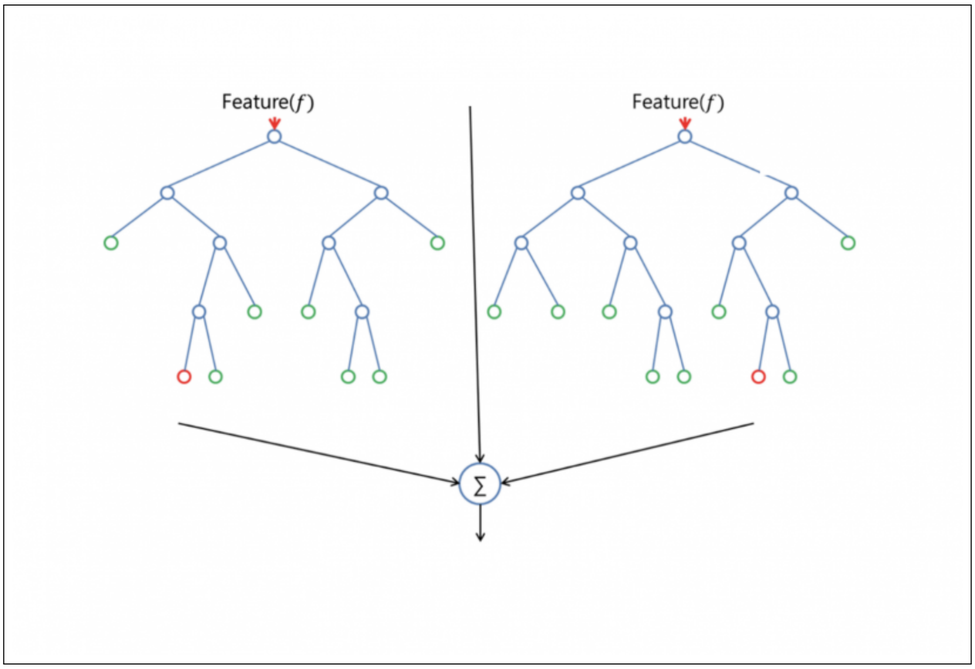


Figure 4.1: Random Forest

Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

4.1.3 Pseudocode of Random Forest ClassifierS

- Randomly select “k” features from total “m” features.(Where k ¡¡ m)
- Among the “k” features, calculate the node “d” using the best split point.
- Split the node into daughter nodes using the best split.
- Repeat 1 to 3 steps until “l” number of nodes has been reached.
- Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

4.1.4 Flowchart of Random Forest Classifier

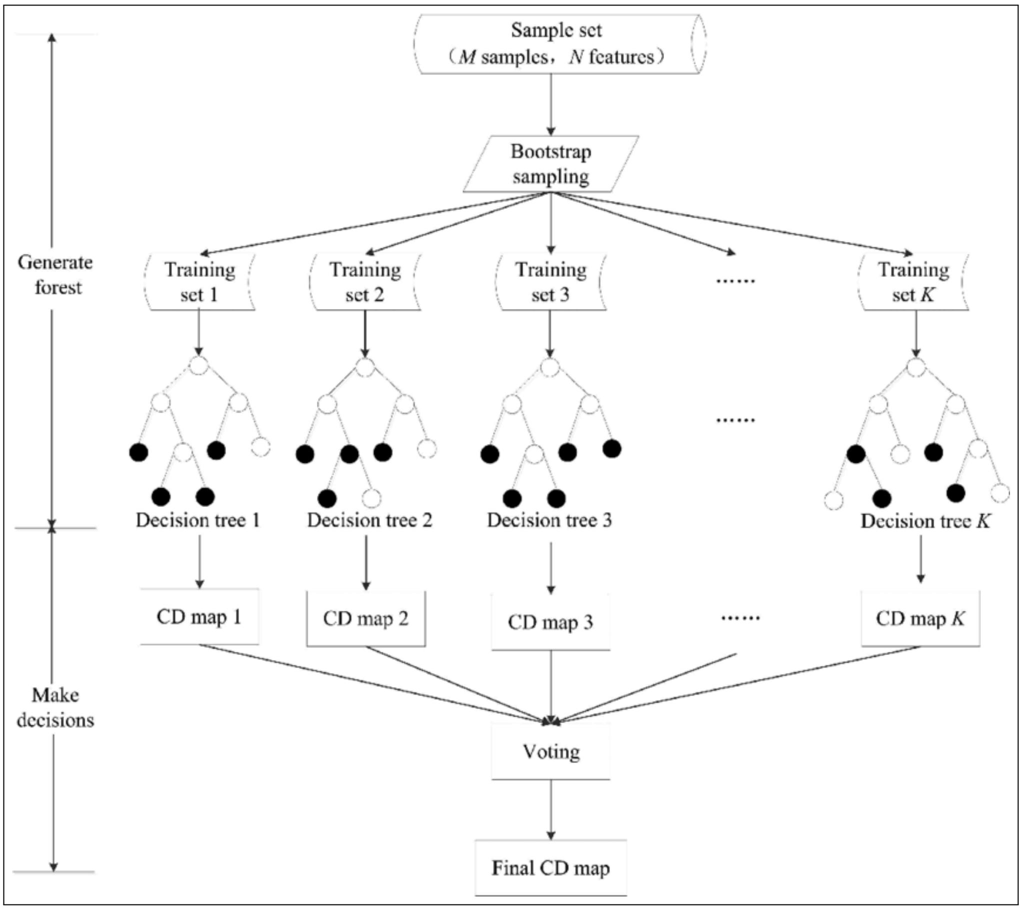


Figure 4.2: Random Forest Flowchart

4.2 Decision Tree

4.2.1 What is Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

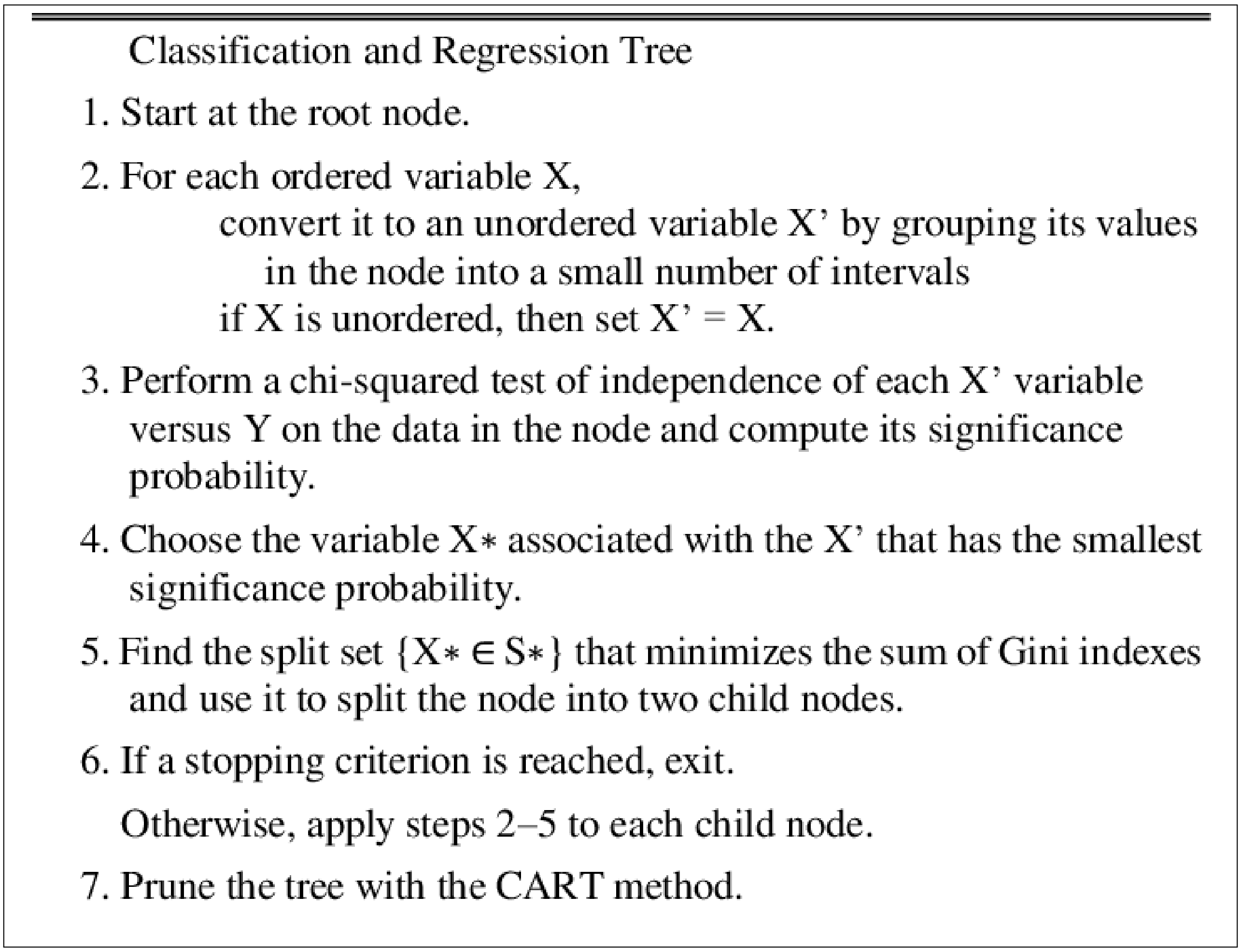
4.2.1.1 Construction of Decision Tree :

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

4.2.1.2 Decision Tree Representation:

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree,testing the attribute specified by this node,then moving down the tree branch corresponding to the value of the attribute as shown in the above figure.This process is then repeated for the subtree rooted at the new node.

4.2.2 pseudocode of Decision Tree Regression:



```
Classification and Regression Tree
1. Start at the root node.
2. For each ordered variable X,
        convert it to an unordered variable X' by grouping its values
            in the node into a small number of intervals
        if X is unordered, then set X' = X.
3. Perform a chi-squared test of independence of each X' variable
   versus Y on the data in the node and compute its significance
   probability.
4. Choose the variable X* associated with the X' that has the smallest
   significance probability.
5. Find the split set {X* ∈ S*} that minimizes the sum of Gini indexes
   and use it to split the node into two child nodes.
6. If a stopping criterion is reached, exit.
   Otherwise, apply steps 2–5 to each child node.
7. Prune the tree with the CART method.
```

Figure 4.3: Pseudocode

4.2.3 Flowchart of Decision Tree Regression:

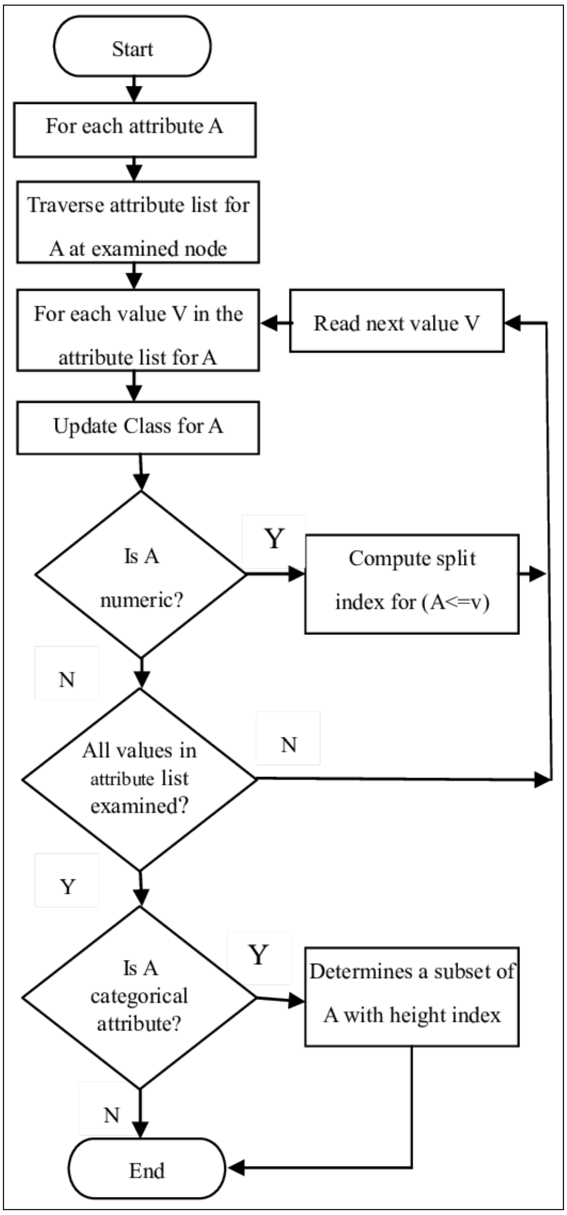

Start
For each attribute A
Traverse attribute list for A at examined node
For each value V in the attribute list for A
Read next value V
Update Class for A
Is A numeric?
Y
Compute split index for (A<=v)
N
All values in attribute list examined?
N
Y
Is A categorical attribute?
Y
Determines a subset of A with height index
N
End

Figure 4.4: Flowchart

Chapter 5

DETAILED DESIGN

5.1 Architectural Design

For system developers, they need system architecture diagrams to understand, clarify, and communicate ideas about the system structure and the user requirements that the system must support. It's a basic framework can be used at the system planning phase helping partners understand the architecture, discuss changes, and communicate intentions clearly.

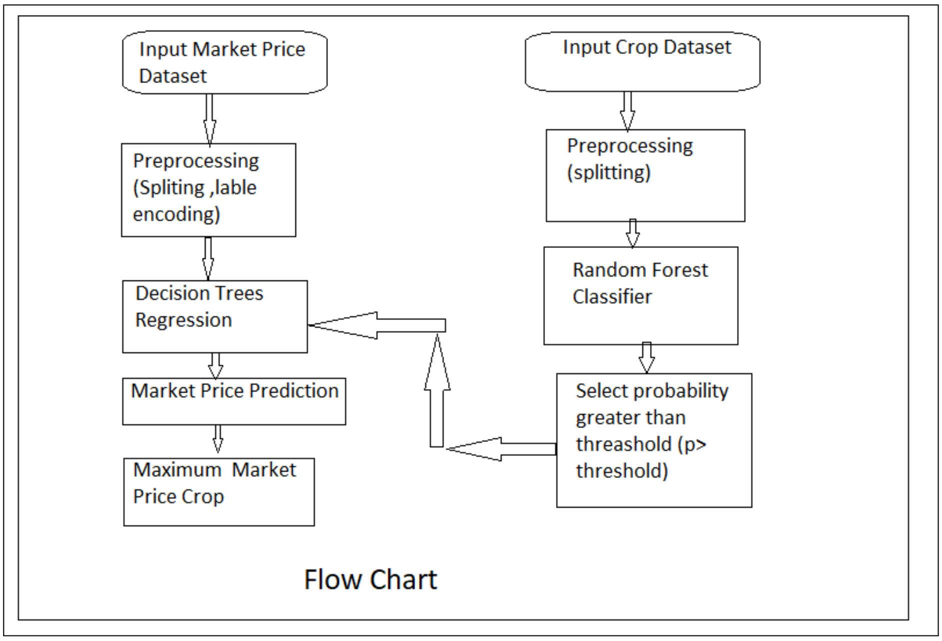

Input Market Price Dataset
Preprocessing (Spliting ,lable encoding)
Decision Trees Regression
Market Price Prediction
Maximum Market Price Crop
Input Crop Dataset
Preprocessing (splitting)
Random Forest Classifier
Select probability greater than threashold (p> threshold)
Flow Chart

Figure 5.1: System Architechture

- Crop Dataset and Market price dataset are given input to the system.
- Preprocessing tasks like splitting and lable encoding is done on the datset
- After preprocessing Random forest classifier is applied on it to get the probabilites of crops.
- Select the probabilities greater than threshold value.
- selected probabilities provided as input to the decision tree regression.
- Decision tree regression predicts the market price of the crops.
- From the predicted crops maximum market price crop is displayed on the screen.

5.2 UML Diagrams/DFD

Use Case diagrams identify the functionality provided by the system (use cases), the users who interact with the system (actors), and the association between the users and the func- tionality. Use Cases are used in the Analysis phase of software development to articulate the high-level requirements of the system.

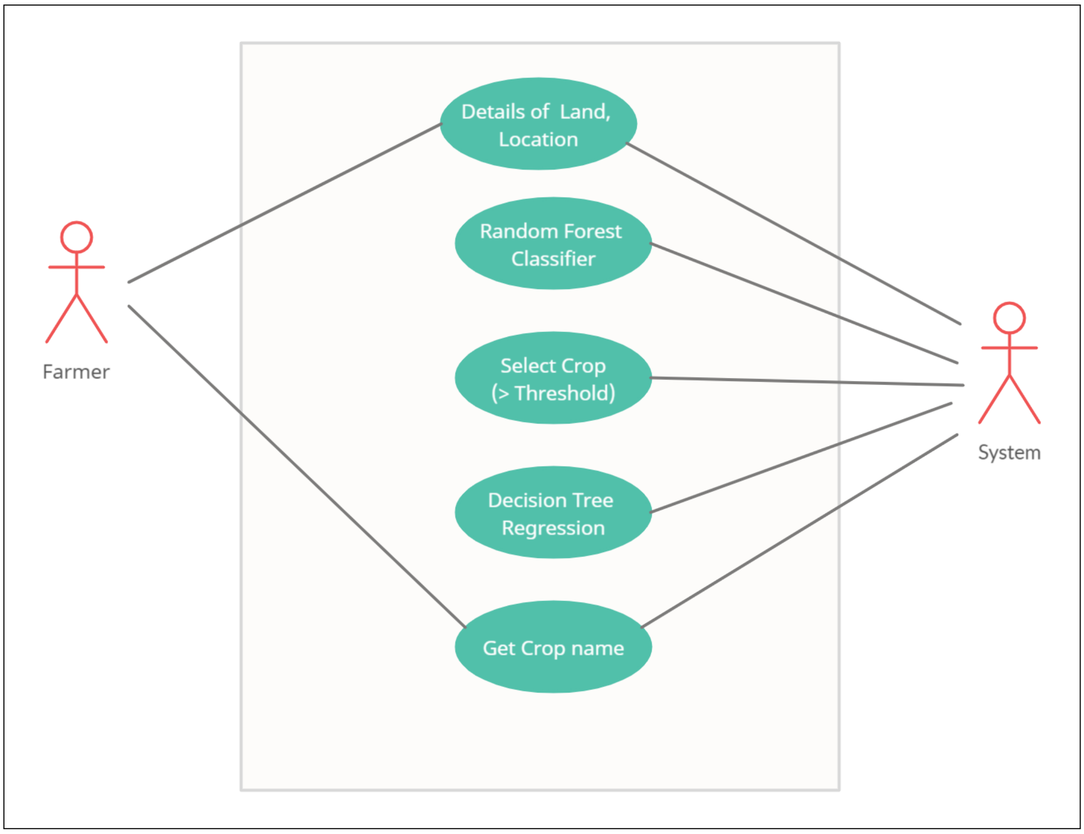
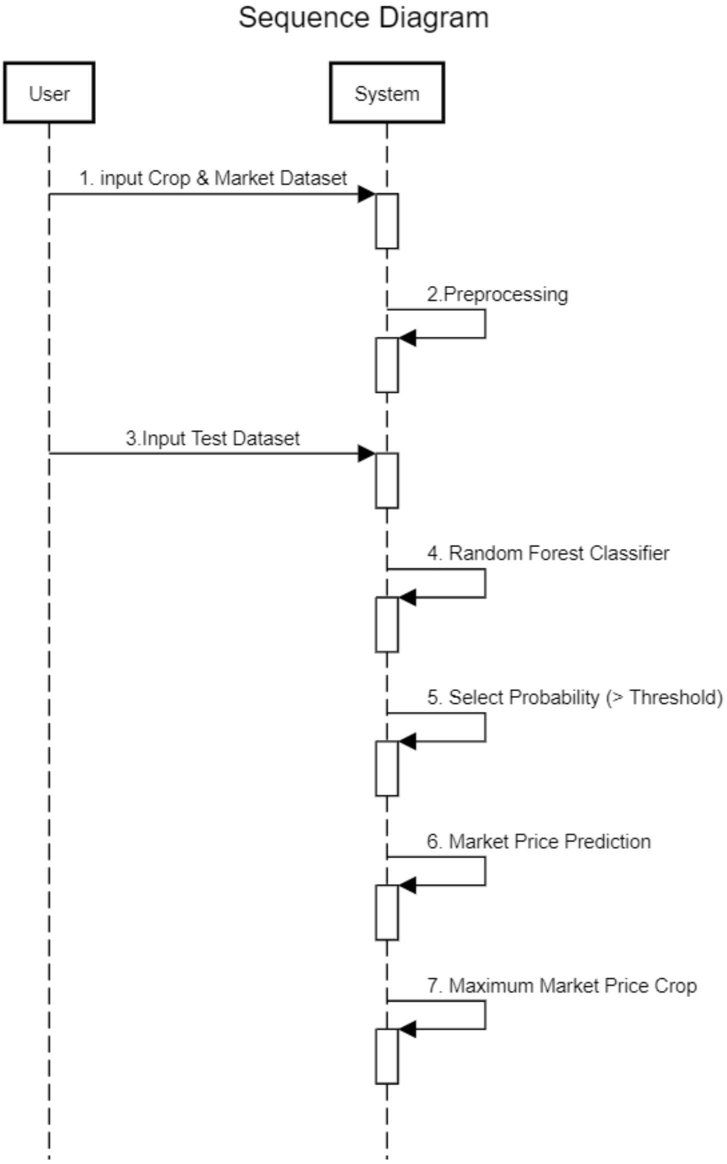


Figure 5.2: Use Case Diagram

5.3 Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.


Sequence Diagram
User
System
1. input Crop & Market Dataset
2.Preprocessing
3.Input Test Dataset
4. Random Forest Classifier
5. Select Probability (> Threshold)
6. Market Price Prediction
7. Maximum Market Price Crop

5.4 Activity Diagram

Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system.The main element of an activity diagram is the activity itself. An activity is a function performed by the system. After identifying the activities, we need to understand how they are associated with constraints and conditions.

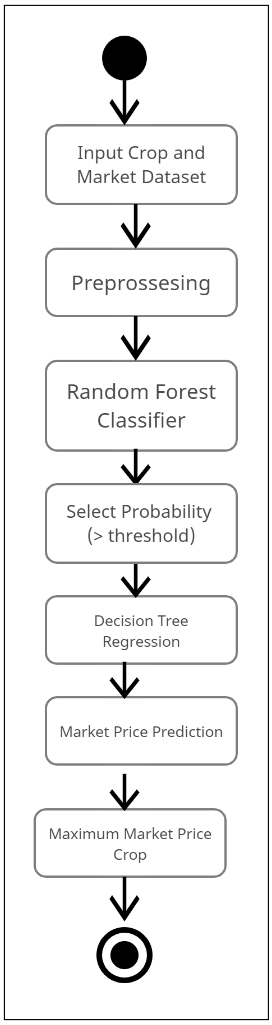

Input Crop and Market Dataset
Preprossesing
Random Forest Classifier
Select Probability (> threshold)
Decision Tree Regression
Market Price Prediction
Maximum Market Price Crop

Figure 5.3: Activity Diagram

5.5 Data Flow Diagram

A data flow diagram (DFD) is a graphical representation of the ”flow” of data through an information system,modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing.

5.5.1 DFD Level 0

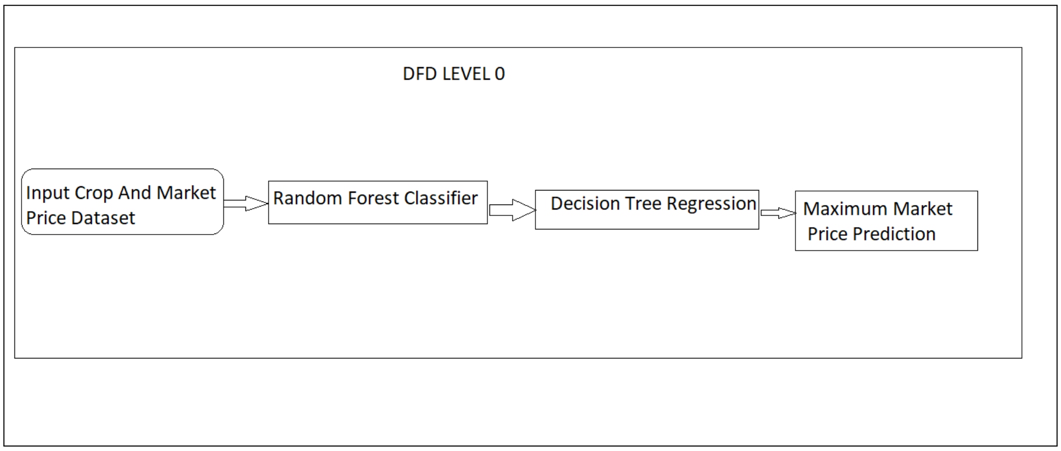

DFD LEVEL 0
Input Crop And Market Price Dataset
Random Forest Classifier
Decision Tree Regression
Maximum Market Price Prediction

Figure 5.4: DFD Level 0

5.5.2 DFD Level 1

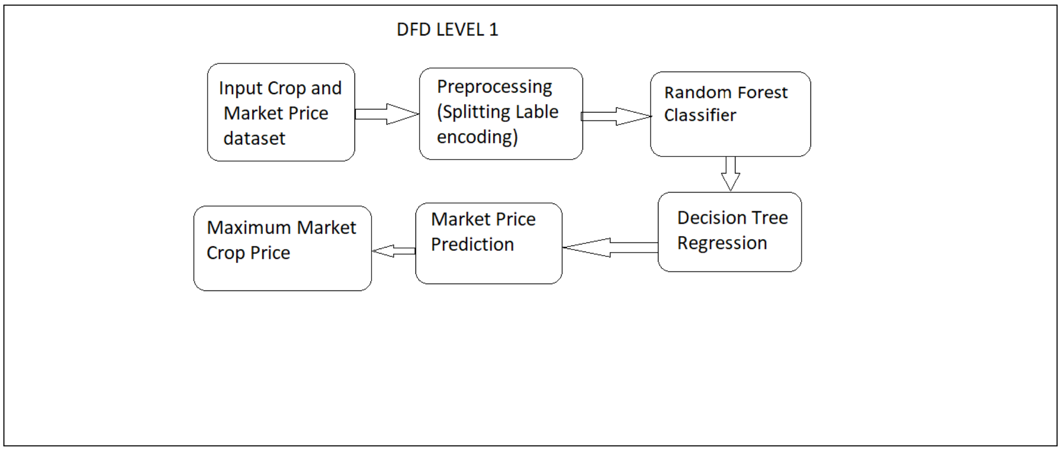

DFD LEVEL 1
Input Crop and Market Price dataset
Preprocessing (Splitting Lable encoding)
Random Forest Classifier
Decision Tree Regression
Market Price Prediction
Maximum Market Crop Price

Figure 5.5: DFD Level 1

5.5.3 DFD Level 2

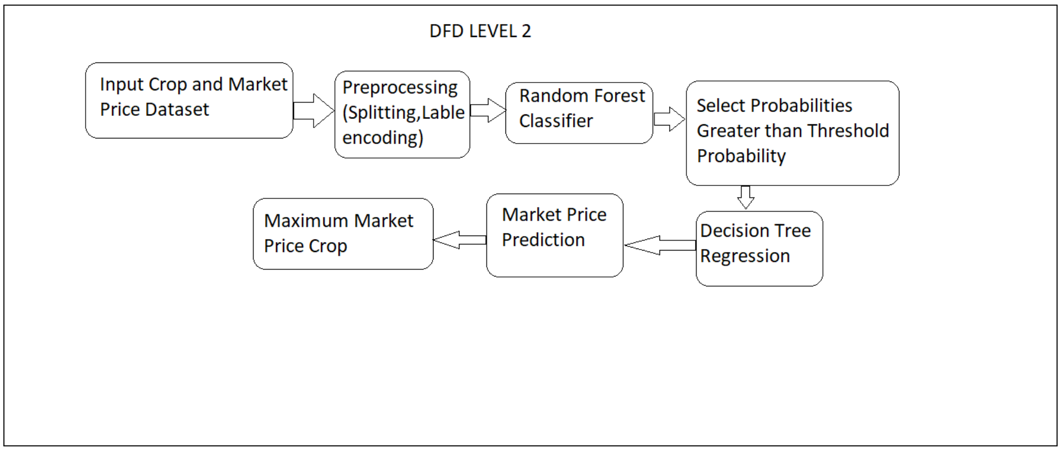

DFD LEVEL 2
Input Crop and Market Price Dataset
Preprocessing (Splitting,Lable encoding)
Random Forest Classifier
Select Probabilities Greater than Threshold Probability
Decision Tree Regression
Market Price Prediction
Maximum Market Price Crop

Figure 5.6: DFD Level 2

Chapter 6

CONCLUSIONS

The proposed system takes into consideration the data related to soil, weather and past year production and suggests which are the best profitable crops which can be cultivated in the apropos environmental condition. As the system lists out all possible crops, it helps the farmer in decision making of which crop to cultivate. Also, this system takes into consideration the past production of data which will help the farmer get insight into the demand and the cost of various crops in market. As maximum types of crops will be covered under this system, farmer may get to know about the crop which may never have been cultivated.

Chapter 7

REFERENCES

- Komal "PREDICTION OF CROP YEILDS BY USING DATA ANALYSIS" irjmets, Volume:02/Issue:06/June -2020
- Potnuru Sai Nishant, Pinapa Sai Venkat, Bollu Lakshmi Avinash, B. Jabber "Crop Yield Prediction based on Indian Agriculture using Machine Learning" IEEE 2020
- S.Bhanumathi worked on "Crop Yield Prediction and Efficient use of Fertilizers" IEEE 2019. data.gov.in." [Online]. Available: https://data.gov.in/