A Project Report on

RECOMMENDATION OF CROP AND YIELD PREDICTION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY

IN

COMPUTER ENGINEERING

OF

VISHWAKARMA INSTITUTE OF TECHNOLOGY

Savitribai Phule Pune University

BY

Mohit Heda (Gr No. 151027)
Apeksha Chaudhari (Gr No. 151065)
Vinayak Mohitkar (Gr No. 151365)
Esha Chakradeo (Gr No. 151393)

UNDER THE GUIDANCE OF
Prof. DR. S. T. PATIL

DEPARTMENT OF COMPUTER ENGINEERING

BANSILAL RAMNATH AGARWAL CHARITABLE TRUST'S
VISHWAKARMA INSTITUTE OF TECHNOLOGY
(An Autonomous Institute affiliated to Savitribai Phule Pune University)
PUNE - 411037

2018 - 2019

BANSILAL RAMNATH AGARWAL CHARITABLE TRUST'S

VISHWAKARMA INSTITUTE OF TECHNOLOGY

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

PUNE – 411037

CERTIFICATE

This is to certify that the project titled **Recommendation of crop and yield production** submitted by **Apeksha Chaudhari (GR Number: 151065), Mohit Heda (GR Number: 151027), Vinayak Mohitkar (GR Number: 151365) and Esha Chakradeo (GR Number: 151393)** is in partial fulfillment for the award of Degree of **Bachelor of Technology in Computer Engineering** of Vishwakarma Institute of Technology. This project is a record of bonafide work carried out by them under my guidance during the academic year 2018-2019.

Guide

[Prof. Dr. S. T. Patil]

Dept. of Computer Engg.

Head

Vishwakarma Institute of Technology, Pune.

DSPuntambekar

Sign of External Examiner

Devdatta Puntambekar

B.R.A.C. Trust's Vishwakarma Institute of Tech. Computer Engg. Dept.

Date: 16/5/2019

ACKNOWLEDGMENT

It gives us an immense pleasure and satisfaction in presenting this project on **"Recommendation of crop and yield production"**.

This project work has opened up new vistas of knowledge for us. We can now justifiably claim that this experience will stand me in good stead in the years to come. There is large number of people without whose help this Unique learning experience would be a nonstarter.

I wish to express my deep sense of gratitude to our Internal Guide, **Prof. Dr. S. T. Patil**, Computer Engineering, for their valuable guidance and useful suggestions, which helped us in completing this project work, in time. Without their cooperation, it would have been extremely difficult for us to complete this project work.

I would like to take the opportunity to thank **Director Prof. (Dr.) R. M. Jalnekar, Prof. (Dr.) V. S. Deshpande** (Head of the Department, Computer Engineering) for extending their kind co-operation and for providing valuable educational environment during my project work.

Needless to mention, I thank the management and Lab assistants of the department for their encouragement and cooperation in carrying out the project work.

Apeksha Chaudhari	**Mohit Heda**	**Vinayak Mohitkar**	**Esha Chakradeo**
(151065)	**(151027)**	**(151365)**	**(151393)**

CONTENTS

List of Tables

List of Figures

List of Graphs

ABSTRACT

India is agrarian country. Agriculture is the main backbone of the Indian economy. Total 60% of India's land comes under the agriculture. 70 % of population are directly or indirectly depend on agriculture in India. Agriculture not only play the important role in large industries but also each and every person of the society are depend on it. So crop monitoring is crucial task to know the crop health. As we know agriculture is backbone of economy so productivity and quality is key points in agriculture sector. To grow the high quality product and maximize the yield production proper and in time "Recommendation of crops" is very much needed. This project presents the recommendation of crops and prediction of yield using the machine learning algorithm which uses rainfall, temperature, pH and area of crop yield. Classification and regression algorithms are used for crop disease analysis. Proposed systems application would be useful to the farmer to detect the yield of crop and the system will recommend which crop to be grown in which area. Achieving maximum crop yield at minimum cost is one of the goals of agricultural production. Early detection and management of problems associated with crop yield indicators can help increase yield and subsequent profit. Predictions could be used by crop managers to minimize losses when unfavorable conditions may occur.

Keywords: Feature Extraction, Machine Learning (ML), Linear Regression, Crop Yield, Agriculture, Crop, Accuracy, Classification, Crop Recommendation, Rainfall, Temperature, pH value.

CHAPTER 1

INTRODUCTION

1.1 PROBLEM DEFINITION

To develop a system which recommends crop and yield prediction and to suggest the crops suitable for farming.

1.2 OVERVIEW

Agriculture plays a vital role in India's economy. Over 58 per cent of the rural households depend on agriculture as their principal means of livelihood. The Indian food and grocery market is the world's sixth largest, with retail contributing 70 per cent of the sales. The Indian food processing industry accounts for 32 per cent of the country's total food market, one of the largest industries in India and is ranked fifth in terms of production, consumption, export and expected growth. India is agrarian country. Agriculture is the main backbone of the Indian economy. Total 60% of India's land comes under the agriculture. 70 % of population are directly or indirectly depend on agriculture in India. Agriculture not only play the important role in large industries but also each and every person of the society are depend on it. So crop monitoring is crucial task to know the crop health. As we know agriculture is backbone of economy so productivity and quality is key points in agriculture sector. To grow the high quality product and maximize the yield production proper and in time "Crop Recommendation" is very much needed.

This chapter gives a description about machine learning, tasks involved, models, advantages, disadvantages, objective of the proposed work and thesis overview. The classified crop field areas along with the historical weather and yield data are modelled to obtain the predicted crop yield and recommend suitable crops for a particular field. This in turn involves machine learning for classification and prediction.

1.3 MACHINE LEARNING

Machine Learning is a field of Computer Science, where new developments evolve at recent times, and also helps in automating the evaluation and processing done by the mankind, thus reducing the burden on the manual human power. According to techtarget, Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

Machine learning focuses on the development of computer programs that can change when exposed to new data. Finding out the suitable crops based on the soil's appearance becomes tedious for novice farmers. There also exists a need to prevent the agricultural decay.

1.3.1 TYPES OF PROBLEMS AND TASKS

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal. Another example is learning to play a game by playing against an opponent

1.3.2 LINEAR REGRESSION

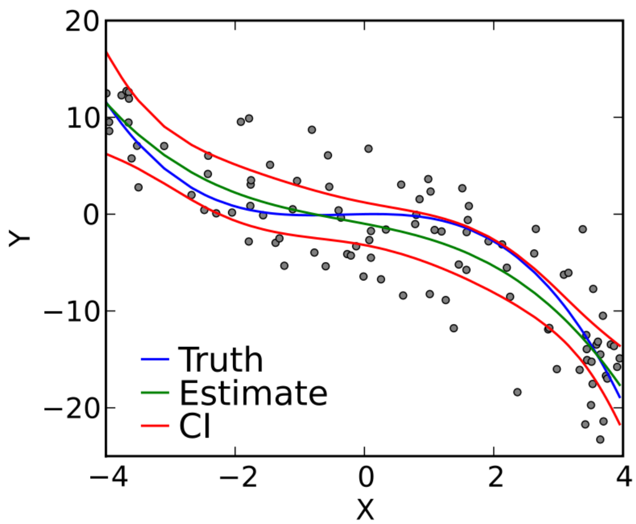
Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.



A linear regression model predicts the target as a weighted sum of the feature inputs. The linearity of the learned relationship makes the interpretation easy. Linear regression models have long been used by statisticians, computer scientists and other people who tackle quantitative problems.

Linear models can be used to model the dependence of a regression target y on some features x. The learned relationships are linear and can be written for a single instance i as follows:

$$y=\beta0+\beta1x1+\ldots+\beta pxp+\epsilon y=\beta0+\beta1x1+\ldots+\beta pxp+\epsilon$$

The predicted outcome of an instance is a weighted sum of its p features. The betas (βjβj) represent the learned feature weights or coefficients. The first weight in the sum (β0β0) is called the intercept and is not multiplied with a feature. The epsilon (ϵϵ) is the error we still make, i.e. the difference between the prediction and the actual outcome. These errors are assumed to follow a Gaussian distribution, which means that we make errors in both negative and positive directions and make many small errors and few large errors.

1.3.2.1 ADVANTAGES OF LINEAR MODEL

It makes the estimation procedure simple and, most importantly, these linear equations have an easy to understand interpretation on a modular level (i.e. the weights). This is one of the main reasons why the linear model and all similar models are so widespread in academic fields such as medicine, sociology, psychology, and many other quantitative research fields. For example, in the medical field, it is not only important to predict the clinical outcome of a patient, but also to quantify the influence of the drug and at the same time take sex, age, and other features into account in an interpretable way.

Estimated weights come with confidence intervals. A confidence interval is a range for the weight estimate that covers the “true” weight with a certain confidence. For example, a 95% confidence interval for a weight of 2 could range from 1 to 3. The interpretation of this interval would be: If we repeated the estimation 100 times with newly sampled data, the confidence interval would include the true weight in 95 out of 100 cases, given that the linear regression model is the correct model for the data.

Whether the model is the “correct” model depends on whether the relationships in the data meet certain assumptions, which are linearity, normality, homoscedasticity, independence, fixed features, and absence of multicollinearity.

1.3.2.2 DISADVANTAGES OF LINEAR MODEL

Linear regression looks at a relationship between the mean of the dependent variable and the independent variables. For example, if you look at the relationship between the birth weight of infants and maternal characteristics such as age, linear regression will look at the average weight of babies born to mothers of different ages. However, sometimes you need to look at the extremes of the dependent variable.

Linear regression assumes that the data are independent. That means that the scores of one subject (such as a person) have nothing to do with those of another. This is often, but not always, sensible. Two common cases where it does not make sense are clustering in space and time.

A classic example of clustering in space is student test scores, when you have students from various classes, grades, schools and school districts. Students in the same class tend to be similar in many ways, i.e., they often come from the same neighbourhoods, they have the same teachers, etc. Thus, they are not independent.

Examples of clustering in time are any studies where you measure the same subjects multiple times. For example, in a study of diet and weight, you might measure each person multiple times. These data are not independent because what a person weighs on one occasion is related to what he or she weighs on other occasions.

1.3.3 SUPPORT VECTOR MACHINE

SVM constructs a hyperplane or set of hyperplanes in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

The computational load should be sensible, the mappings are used by the SVM scheme to ensure the dot products will be computed in terms of the variable in the original scope, for that a kernel function **k(x,y)** selected to get the optimal computational time.

The higher-dimensional space in the hyper planes is distinct as the set of points whose dot product with a vector in that space is constant. These vectors in the hyper planes defining the hyper planes can be chosen to be linear combinations with parameters of images of feature vectors that occur in the data base. With this choice of a hyperplane, the points in the feature space that are mapped into the hyperplane are defined by the relation: The equation of the output from a linear SVM is

$$\mathbf{u = w \cdot x - b}$$

Where w is the normal vector of the hyperplane, and x is the input vector.

1.3.3.1 ADVANTAGES OF SUPPORT VECTOR MACHINE

Support vector machine is one of the most widely used classification algorithms due to the advantages it enjoys which are as follows:

SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings.

Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback. This is also true of image segmentation systems, including those using a modified version SVM.

1.4 MOTIVATION

Agriculture is the backbone for a developing economy like India and there is an enormous need to maintain the agricultural sustainability. Hence it is a significant contribution towards the economic and agricultural welfare of the countries across the world.

Now-a-days, even India is cultivated country and 70% of population are depend directly or indirectly on agriculture, India's farmer are facing lots of problems in their life. Two major reasons are there behind it. First uncertain rain water because of more industrialization and second one is unavailability of good system for crop disease detection. So first issue is not in our hand that's why we are try to cover the second one. Crop Disease is something which causes the unhealthy crop and it tends to the overall outcome from the crop. Crop Disease Restrict the crop to produce the lesser amount of product than their ability and also the low quality yield. Using the proposed system farmer can easily detect the crop disease at early stage and can take the proper action which directly results in yield production and quality of the product.

1.5 OBJECTIVE

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

- Classify the image based on Soil type, moisture content, weather conditions, pH value, organic nitrogenetic.
- Perform satellite image processing with respect to textural and spatial features.
- Analyze crop patterns with the help of past records and map them with calculated data.
- Monitor crop yield and find ways for increasing it.
- Recommend profitable crops for each land type.

CHAPTER 2

LITERATURE REVIEW

Sr no	Author	Paper Title	Methodology	Conclusion
1	Miss.Snehal S.Dahikar Dr.Sandeep V.Rode	Agricultural Crop Yield Prediction Using Artificial Neural Network Approach	Supervised Learning(Back Propagation, Feed Forward)	feed forward back propagation methods used for developing ANN model.
2	S.Veenadhari Dr. Bharat Misra Dr. CD Singh S.Veenadhari	Machine learning approach for forecasting crop yield based on climatic parameters	Decision Tree	The decision rules developed based on the model for soybean crop in Dewas district
3	Anup K. Prasad , Lim Chai , Ramesh P. Singh, Menas Kafatos Anup K. Prasad , Lim Chai , Ramesh P. Singh, Menas Kafatos	Crop yield estimation model for Iowa using remote sensing and surface parameters	Linear regression	The model discussed in the present paper reasonably minimizes inconsistency and errors in yield prediction giving high R 2 -values with maximum accounting of variability in model
4	Askar Choudhury James Jones	Crop yield prediction using time series models	Simple Exponential Smoothing, Double Exponential Smoothing, Damped-Trend Linear Exponential	Among all different time series models estimated, ARMA models performed best with higher coefficient of determinations for all five districts considered in this paper.

			Smoothing, and ARMA(autoregres sive moving average)	
5	X.E. Pantazi , D. Moshou , T. Alexandridis R.L. Whetton , A.M. Mouazen	Wheat yield prediction using machine learning and advanced sensing techniques	Simple Exponential Smoothing, Double Exponential Smoothing, Damped-Trend Linear Exponential Smoothing, and ARMA(autoregres sive moving average)	Among all different time series models estimated, ARMA models performed best with higher coefficient of determinations for all five districts considered in this paper.
6	X.E. Pantazi , D. Moshou	Wheat yield prediction using machine learning and advanced sensing techniques	Counter-propagation Artificial Neural Network (CPANN), Supervised Kohonen Network (SKN) and XY-fusion network (XYF)	The average overall accuracy of cross-validation for SKN was 81.65%, for CP-ANN 78.3% and for XY-F 80.92%, showing that the SKN model had the best overall per-formance.
7	Alberto Gonzalez-Sanchez , Juan Frausto-Solis Waldo Ojeda-Bustamante	Predictive ability of machine learning methods for massive crop yield prediction	Multiple linear regression, Regression trees	Average RMSE shows that kNN has the lowest mean error, followed closely by SVR and M5-Prime.

Table 2.1. List of References

After a thorough background work, some of the most valuable recent documents and papers are, Miss Snehal S. Dahikar, Dr. Sandeep V. Rodhe et al [2] worked on agriculture crop yield prediction using artificial neural networks. Supervised learning (back propagation, feed forward) was used in this regard. In this paper, artificial neural networks is shown to give accurate predictions for agriculture yield production.

B.A. Smith et al [3] discuss year-round air temperature prediction models were developed for prediction horizons of 1 to 12 h using Ward-style ANNs. These models were intended for use in general decision support. The ANN design modifications described herein provided increased accuracy over previously developed, winter specific models during the winter period. It was shown that models that included rainfall terms in the input vector were more accurate than those that did not.

D.L. Ehret et al [5] introduce all crop attributes responded in much the same way to individual climatic factors. Radiation and temperature generally induced strong positive responses while RH produced a negative response. In the NN models, radiation and temperature were still prominent, but the importance of CO_2 in predicting a crop response increased. One advantage of these automated systems is that they offer continuous information across a range of timescales. Furthermore, these systems can readily be used in commercial greenhouses so the derived NN models are relatively easy to deploy to a commercial setting where they can subsequently be improved over time. In this paper crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. For that purpose they used artificial neural network.

Zhang et al. (2010) researchers consider the linear regression model taking into account that the ordinary least square estimation is a generally utilized strategy for prediction of crop yield. Here, autoregressive model performed better than OLS with higher R. The research concluded that NDVI and precipitation contributed more to the corn yield in Iowa, excluding temperature.

Sanchez et al. (2014), researchers show a correlation among a few strategies (linear and nonlinear) for prediction of crop yield. The comparison is made utilizing the best property subset found in the preparation dataset for every strategy which was distinguished utilizing the percentage split validation and a complete algorithm. To search the optimal attribute subset in the training datasets the algorithm uses the oldest samples to build the models. The test datasets is composed of unseen samples where the performance is measured. The most widely recognized information driven method for prediction of crop yield like stepwise regression, multiple linear regression, regression trees and neural systems were assessed. The experimentation demonstrates that quality determination utilizing a complete system generously enhances the execution of all the assessed strategies.

Zaw and Naing (2009), the researchers consider the Polynomial Regression Model (MPR) in order to predict the rainfall in the region of Myanmar. The authors have created prediction forecast model in view of 15 predictors utilizing second-arrange MPR. As a consequence of a few examinations, the anticipated precipitation sum is near to the genuine qualities. SMR expectation model was created with four indicators. The model results were used in the territories, for example in the harvest planning and yield prediction, water administration and repository control. The fundamental point of the improvement of the forecast model is to help in water management. All possible subsets of predictors have been examined regression which utilizes only 2006 test data. Authors demonstrate that MPR performs better than MLR.

2.1 LIMITATION IN EXISTING SYSTEM

- The existing work in this field is done using various models like Artificial Neural Networks, Regression like polynomial regression, multiple linear regression, decision trees, fuzzy algorithm and genetic algorithms.

- These existing systems consider only some of the parameters, multiple parameters are generally not considered.

- The accuracy and precision has scope of improvement.

- Farming is dependent on various factors therefore it is very important to train our model on the basis of multiple parameters for accurate results.

CHAPTER 3

RECOMMENDATION OF CROPS AND YIELD PREDICTION

3.1 INTRODUCTION

Prediction of crop yield mainly strategic plants such as wheat, corn, rice has always been an interesting research area to agro meteorologists, as it is important in national and international economic programming. Dry farming crop production, apart from relationship to the genetic of cultivator, adaphic terms, effect of pests and pathology and weeds, the management and control quality during the growing season and etc. is severely depend to climatic events. Therefore it is not beyond the possibility to acquire relations or systems which can predict the more accuracy using meteorological data. Nowadays, there are a lot of yield prediction models, that more of them have been generally classified in two group a) Statistical Models, b)Crop Simulation Models (e.g. CERES). Recently, application of machine learning models, Fuzzy Systems and Genetic Algorithm has shown more efficiency in dissolving the problem. Application of them can make models easier and more accuracy from complex natural systems with many inputs. In this research it has been tried need simple and accurate estimation techniques to predict rice yields in the planning process. The necessity of the present study were to: (1) identify whether machine learning models could effectively result out in predicting the accurate crop yields (2) identify what are the present models available and how this particular model can prove to be more efficient.

Crop production is a complex phenomenon that is influenced by agro-climatic input parameters. Agriculture input parameters varies from field to field and farmer to farmer. Collecting such information on a larger area is a daunting task. However, the climatic information collected in India at every 1sq.m area in different parts of the district are tabulated by Indian Meteorological Department. The huge such data sets can be used for predicting their influence on major crops of that particular district or place. There are different forecasting methodologies developed and evaluated by the researchers all over the world in the field of agriculture or associated sciences.

3.2 PROPOSED MODEL

- Our project aims at predicting yield and crop based on historical data and suggesting most suitable crop for farming.
- We will give the approximate prediction of yield
- Suggest the crop according to parameters like area, temperature and soil.

3.2 Architecture

3.2.1 Block Diagram

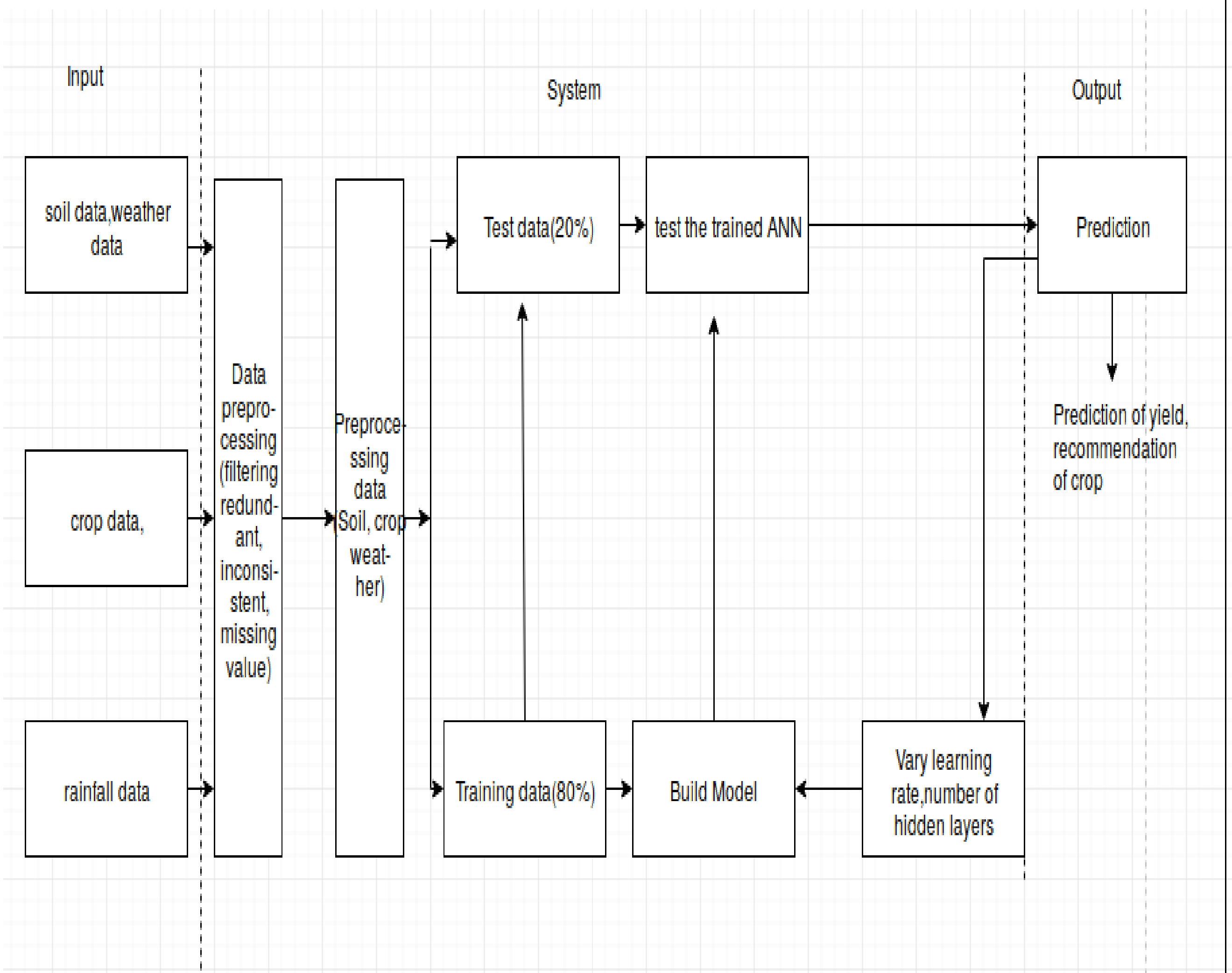


Figure3.2.1. Block diagram

3.2.2 UseCase Diagram

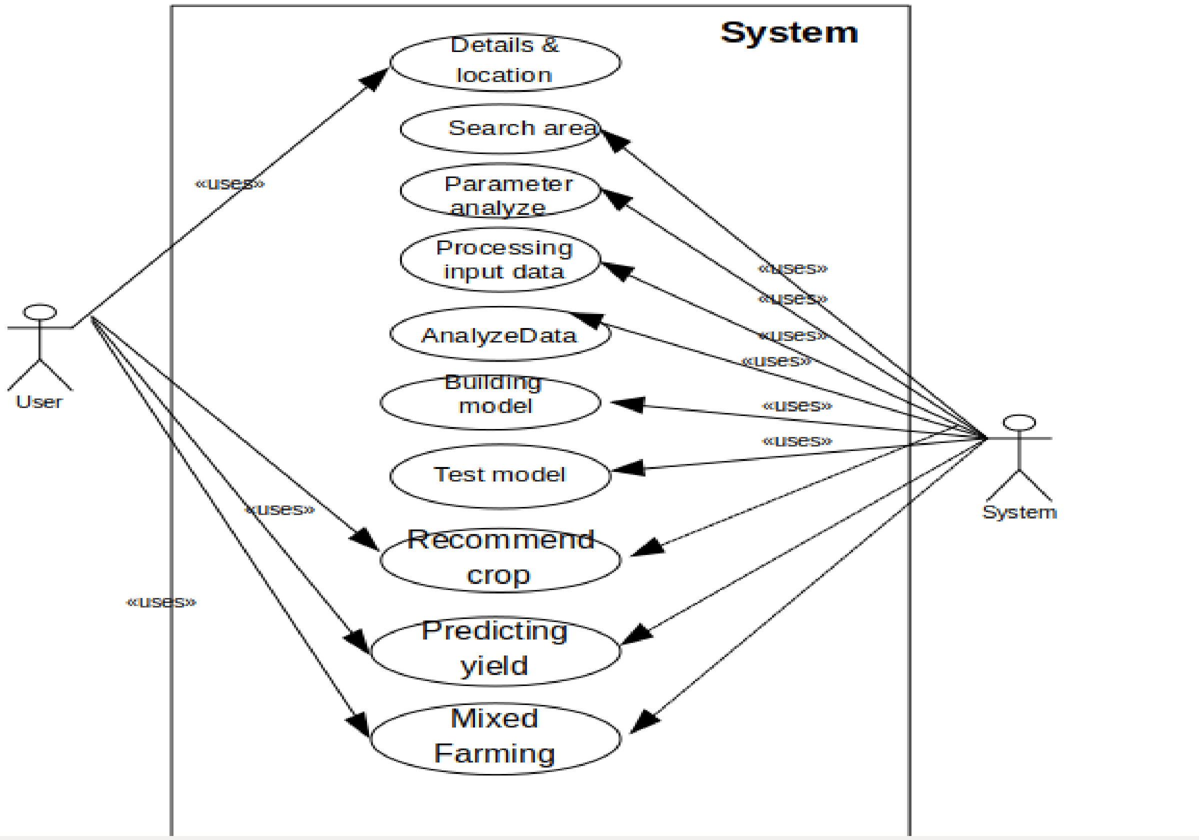


Figure3.2.2. UseCase diagram

3.3 FLOW DIAGRAM

3.3.1 Level 0

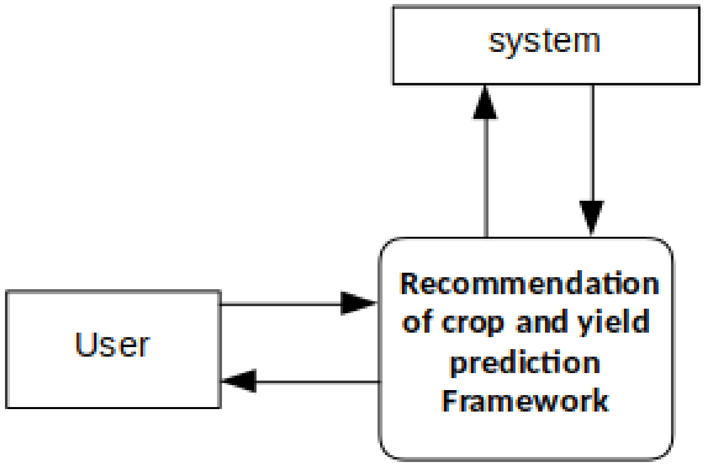


Figure3.3.1. Level 0

3.3.2 Level 1

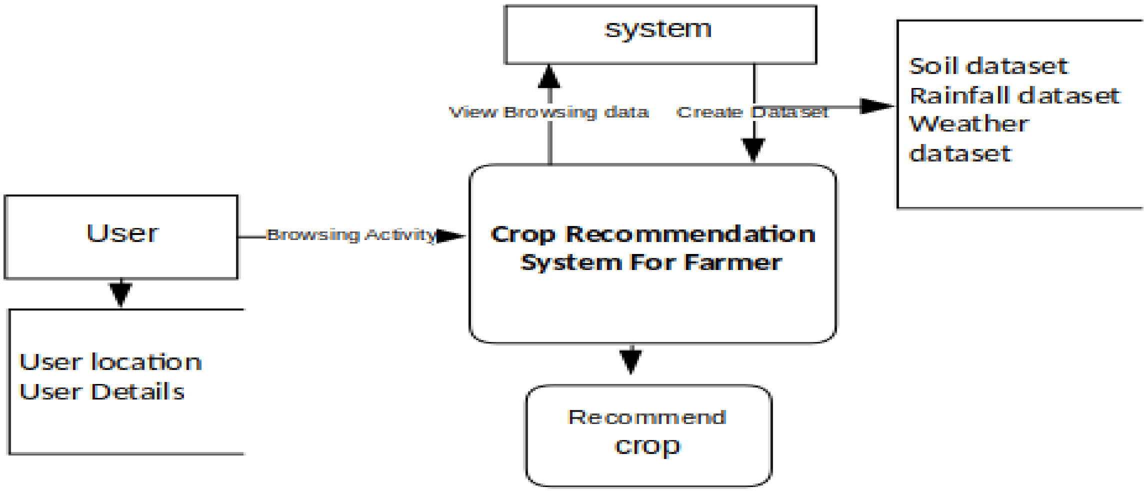


Figure3.3.2. Level 1

3.3.3 Sequence Diagram

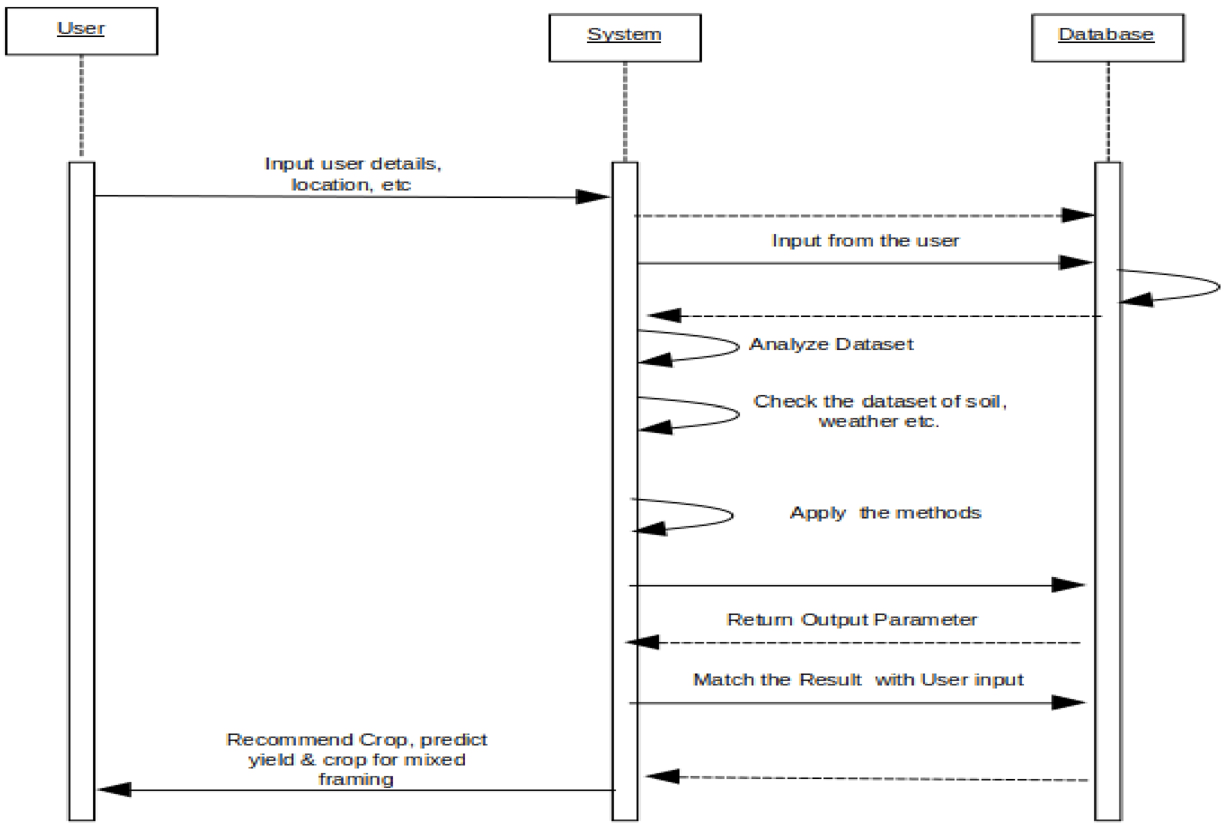


Figure3.3.3. Sequence diagram

3.3.4 Level 2

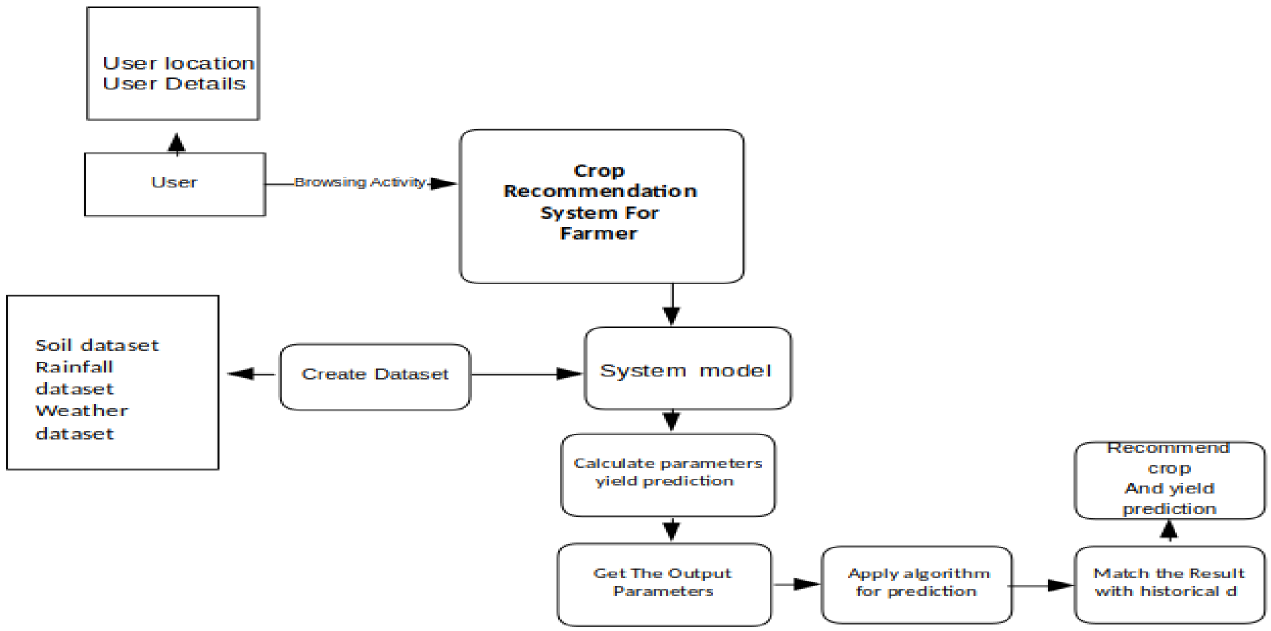


Figure3.3.4. Level 2

3.3.5 Class Diagram

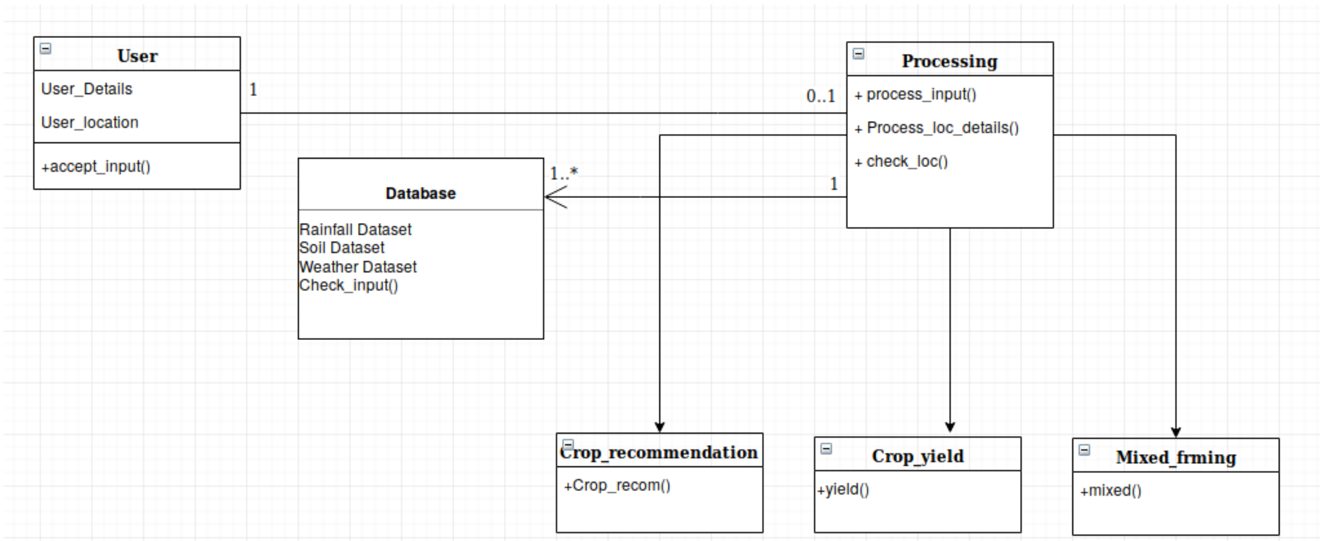


Figure3.3.5. Class Diagram

CHAPTER 4

DATA PREPROCESSING

4.1 INTRODUCTION

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

Data goes through a series of steps during pre-processing:

- Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- Data Integration: Data with different representations are put together and conflicts within the data are resolved.
- Data Transformation: Data is normalized, aggregated and generalized.
- Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
- Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

1. Data Cleaning:
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
(a). Missing Data:
This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

1. Ignore the tuples:
   This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
2. Fill the Missing values:
   There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:
Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each

segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

1. Regression:
   Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
2. Clustering:
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. Normalization:

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. Discretization:

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. Concept Hierarchy Generation:

Here attributes are converted from level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

2. Attribute Subset Selection:

The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

3. Numerosity Reduction:

This enable to store the model of data instead of whole data, for example: Regression Models.

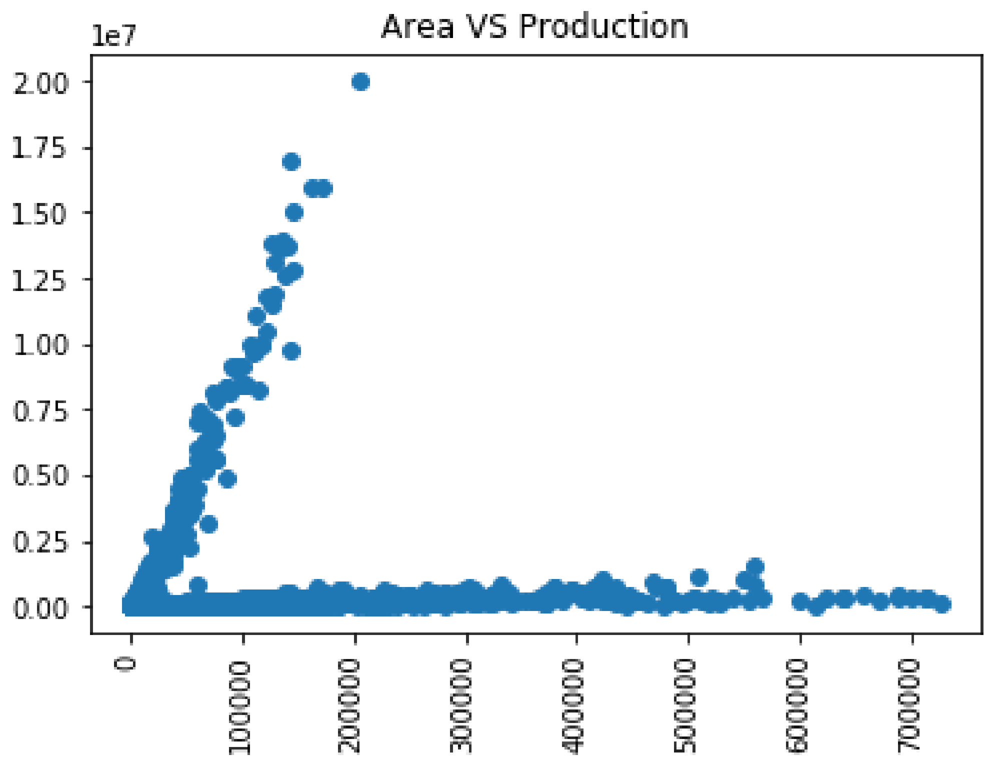
4. Dimensionality Reduction:
This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called

lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Componenet Analysis).

We considered the region of crop production, temperature, rainfall and pH value of soil as parameters for this study. Data is preprocessed and data visulaisation is carried out and following results are obtained:

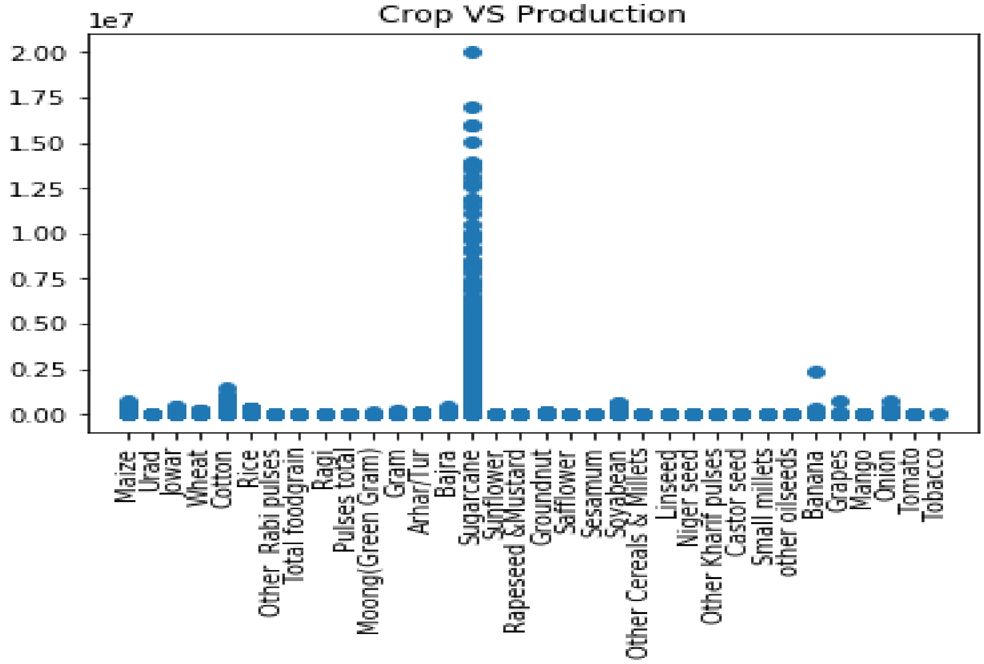
4.2 CROP DATA

4.2.1 AREA VS PRODUCTION

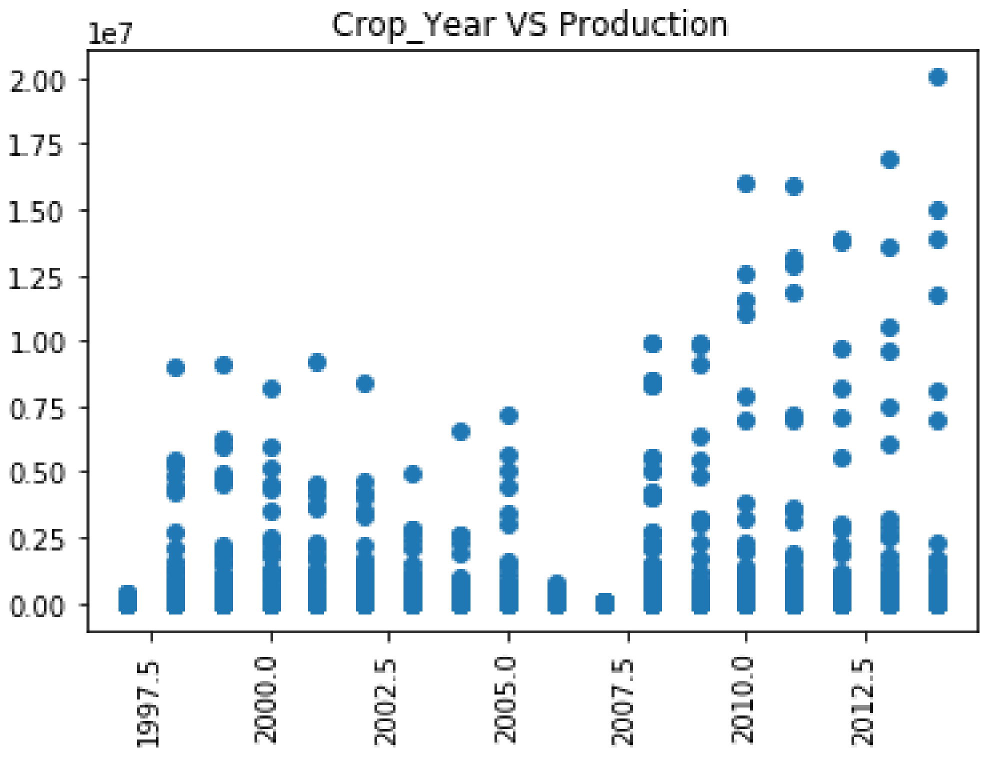


Graph 4.2.1. Area vs Production

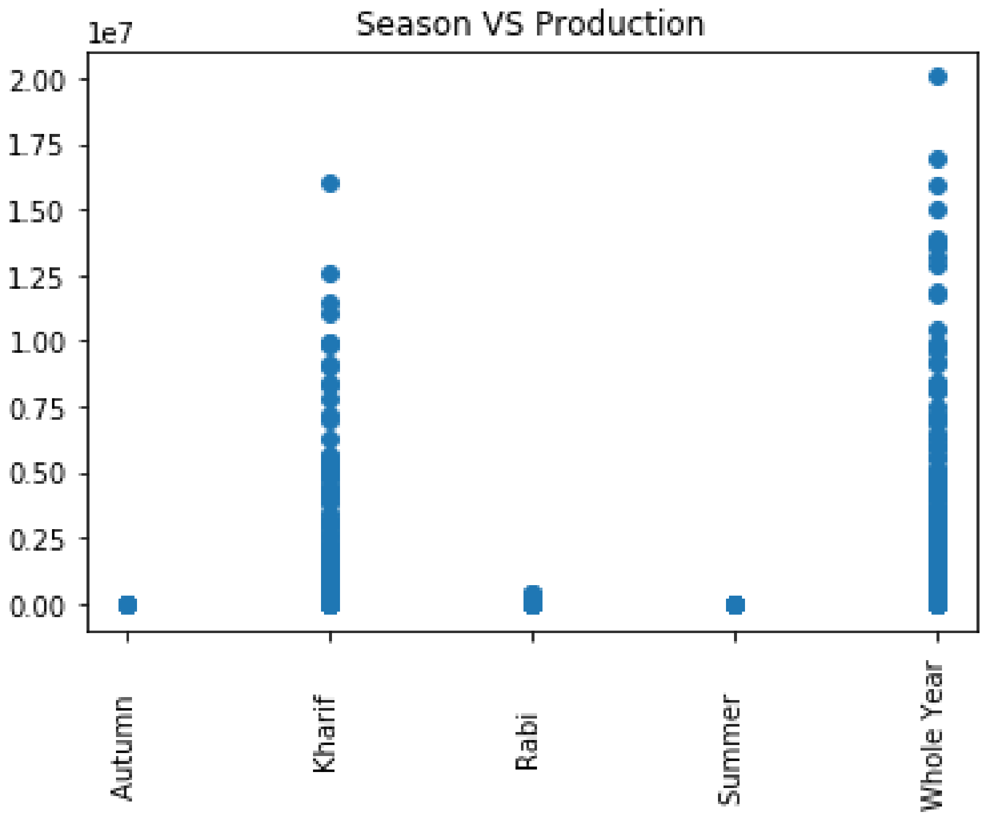
4.2.2 CROP VS PRODUCTION



4.2.3 CROP_YEAR VS PRODUCTION

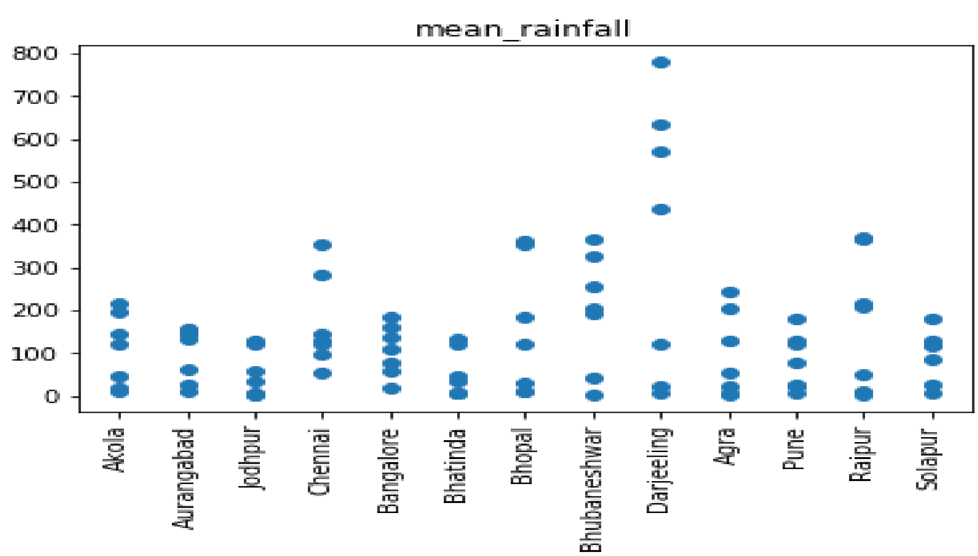


4.2.4 SEASON VS PRODUCTION

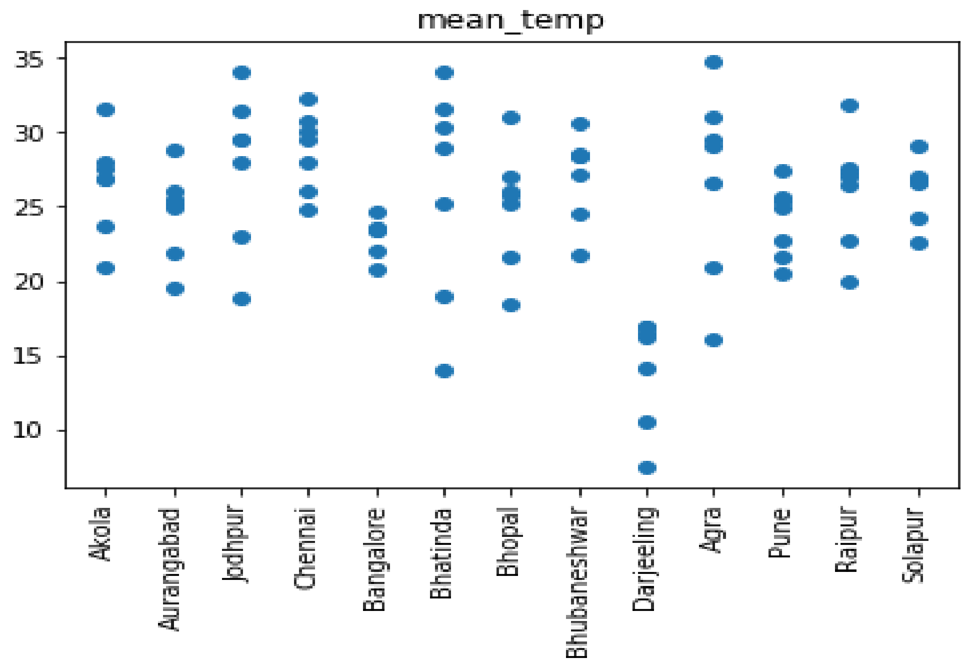


4.3 TEMPERATURE RAINFALL DATA

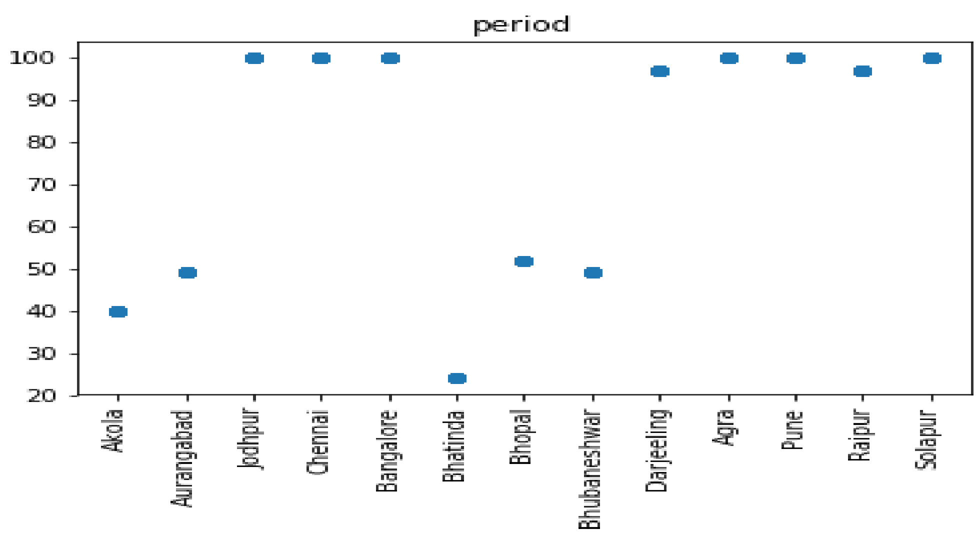
4.3.1 MEAN RAINFALL



4.3.2 MEAN TEMPERATURE



4.3.3 PERIOD

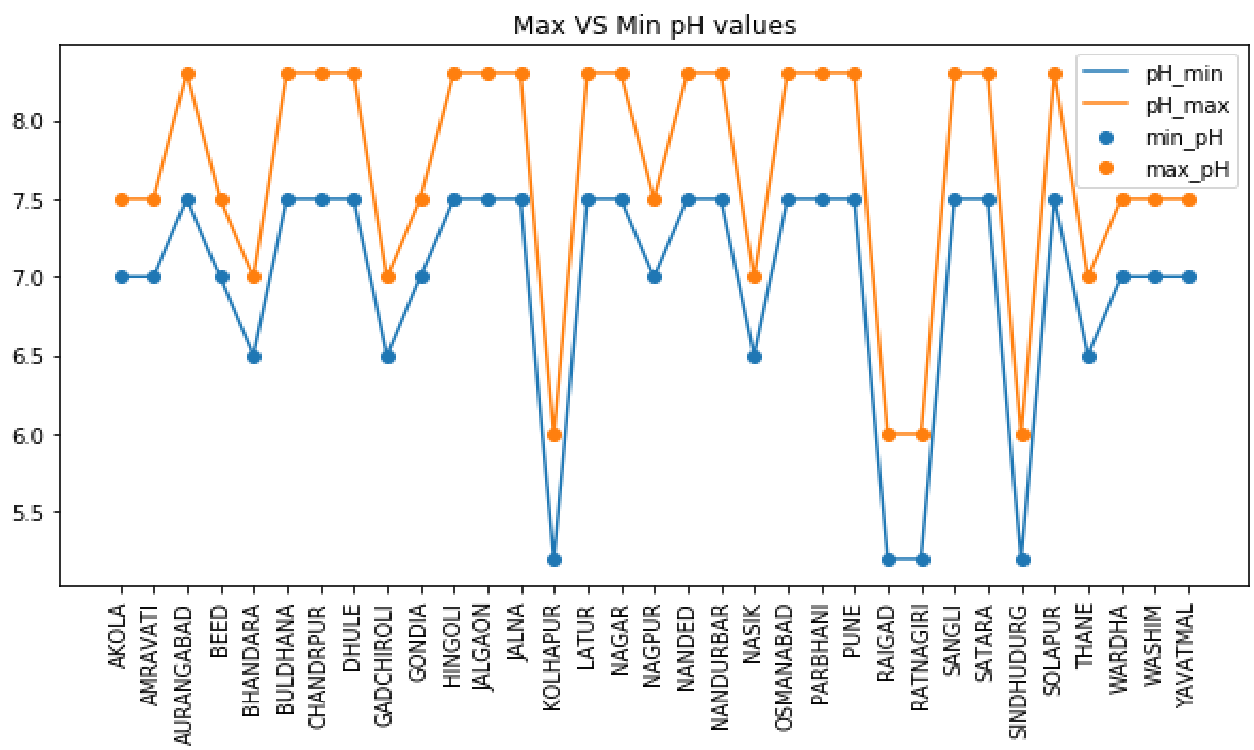


4.4 pH AND DISTRICTS

4.4.1 COUNTS OF DISTRICTS

```
In [120]: countofDistricts
Out[120]:
(array(['AHMEDNAGAR', 'AKOLA', 'AMRAVATI', 'AURANGABAD', 'BEED',
        'BHANDARA', 'BULDHANA', 'CHANDRAPUR', 'DHULE', 'GADCHIROLI',
        'GONDIA', 'HINGOLI', 'JALGAON', 'JALNA', 'KOLHAPUR', 'LATUR',
        'MUMBAI', 'NAGPUR', 'NANDED', 'NANDURBAR', 'NASHIK', 'OSMANABAD',
        'PALGHAR', 'PARBHANI', 'PUNE', 'RAIGAD', 'RATNAGIRI', 'SANGLI',
        'SATARA', 'SINDHUDURG', 'SOLAPUR', 'THANE', 'WARDHA', 'WASHIM',
        'YAVATMAL'], dtype=object),
 array([495, 355, 400, 432, 464, 288, 403, 325, 418, 306, 229, 377, 402,
        435, 409, 466,   1, 388, 461, 377, 459, 461,  11, 467, 495, 236,
        213, 452, 478, 207, 458, 268, 328, 307, 357], dtype=int64))
```

4.4.2 pH CITY WISE



CHAPTER 5

MODELLING FOR CROP RECOMMENDATION AND YIELD PREDICTION

We have divided our project in three parts:

5.1. AREA PRODUCTION:

We used Cropdata.csv dataset here. It consists of

- State name
- District name
- Crop year
- Season
- Crop
- Area
- Production

It has 12629 entries of such data of all the districts in the state of Maharashtra.

Execution step by step process:

A list of 34 prominent cities are shown to choose from all the districts.

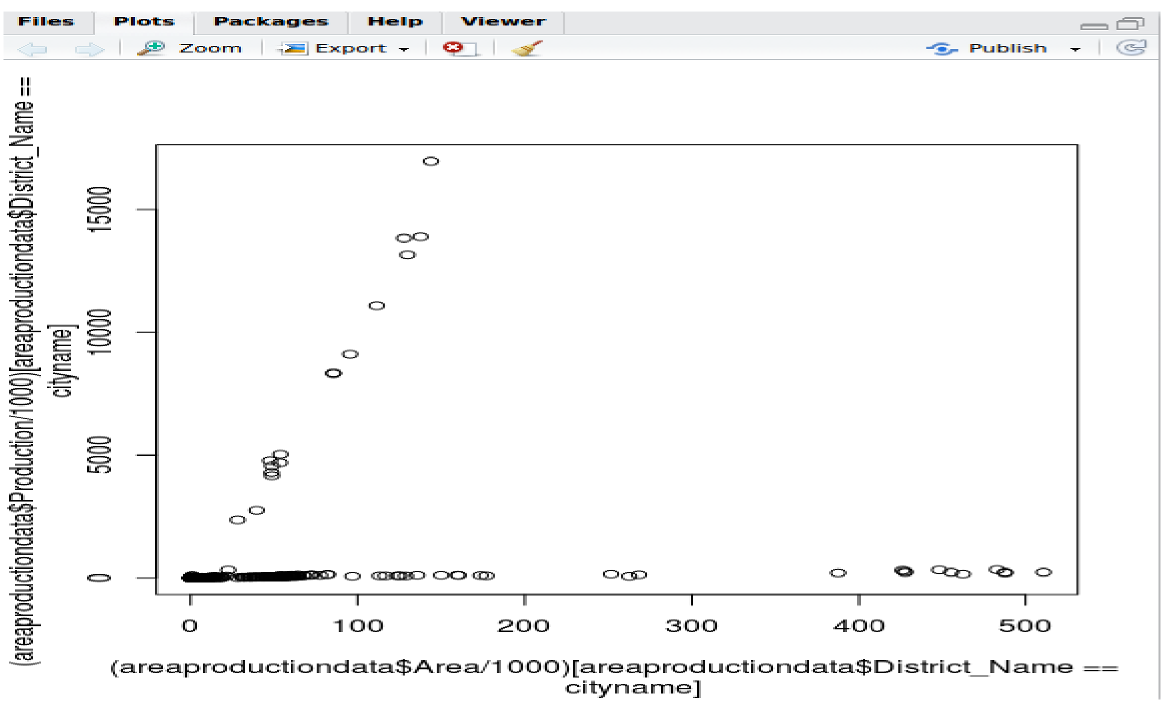
```
Console ~/prgm/ds/
2 AKOLA
3 AMRAVATI
4 AURANGABAD
5 BEED
6 BHANDARA
7 BULDHANA
8 CHANDRAPUR
9 DHULE
10 GADCHIROLI
11 GONDIA
12 HINGOLI
13 JALGAON
14 JALNA
15 KOLHAPUR
16 LATUR
17 NAGPUR
18 NANDED
19 NANDURBAR
20 NASHIK
21 OSMANABAD
22 PARBHANI
23 PUNE
24 RAIGAD
25 RATNAGIRI
26 SANGLI
27 SATARA
28 SINDHUDURG
29 SOLAPUR
30 THANE
31 WARDHA
32 WASHIM
33 YAVATMAL
34 MUMBAI

Choose Corresponding Number to Your City from the Given List:
> 23
[1] 23
>
```

- User is asked to enter the number corresponding to the city.
- The data-set CropData.csv is loaded in ‘areaproductiondata’ variable.
- Few of the starting entries are shown to confirm that he data is loaded.

```
> head(areaproductiondata)
   State_Name District_Name Crop_Year       Season       Crop   Area Production
1 Maharashtra    AHMEDNAGAR      1997 Autumn            Maize      1       1113
2 Maharashtra    AHMEDNAGAR      1997 Kharif        Arhar/Tur  17600       6300
3 Maharashtra    AHMEDNAGAR      1997 Kharif            Bajra 274100     152800
4 Maharashtra    AHMEDNAGAR      1997 Kharif             Gram  40800      18600
5 Maharashtra    AHMEDNAGAR      1997 Kharif            Jowar    900       1100
6 Maharashtra    AHMEDNAGAR      1997 Kharif            Maize   4400       4700
```

- Then the entries corresponding to choosen city is loaded.
- This is then printed for confirmation of loading process.
- Then a plot of area vs production is plotted.(Bth the values are divided by 1000 for better visualization purposes



```
> max((subval$Production)/(subval$Area),na.rm=TRUE)
[1] 1142.5
> expectedcropname <- subval[which.max(subval$Production/subval$Area),]
> print(expectedcropname)
     State_Name District_Name Crop_Year      Season  Crop Area Production
8830 Maharashtra          PUNE      1997 Autumn      Maize    4       4570
```


```
> max((subval$Production)/(subval$Area),na.rm=TRUE)
[1] 1142.5
> expectedcropname <- subval[which.max(subval$Production/subval$Area),]
> print(expectedcropname)
     State_Name District_Name Crop_Year      Season  Crop Area Production
8830 Maharashtra          PUNE      1997 Autumn      Maize    4       4570
```


Average area/productioin is calculated and printed.

- Then all above average crops (area/production) are segregatted from the areaproductiondata, and a list of useful crops are printed.
- Similarly all the crops below average are then printed as unseful crops.

```
> print(unusefulcrop)
     State_Name District_Name Crop_Year      Season              Crop Area Production
9029 Maharashtra          PUNE      2003 Rabi        Rapeseed &Mustard  100          0
```

5.2. TEMPERATURE RAIN CORELATION

We used 3 datasets here.

MaharastraRainfall.csv consists of

- State name
- District name
- Year
- All the 12 months from Jan-December
- Annual Total rainfall.

It has 232 entries of such data of all the districts in the state of Maharashtra.

TempRainfalldata.csv dataset consists of

- City
- Month
- Year
- Period
- Mean Temperature
- Mean Rainfall

It has 92 entries of such data.

CropRequiredTemparature.csv dataset consists of.

- Crop name
- Min temperature
- Max temperature
- Min rainfall
- Max rainfall

It has 12 such entries.

```
Console ~/prgm/ds/
2 AKOLA
3 AMRAVATI
4 AURANGABAD
5 BEED
6 BHANDARA
7 BULDHANA
8 CHANDRAPUR
9 DHULE
10 GADCHIROLI
11 GONDIA
12 HINGOLI
13 JALGAON
14 JALNA
15 KOLHAPUR
16 LATUR
17 NAGPUR
18 NANDED
19 NANDURBAR
20 NASHIK
21 OSMANABAD
22 PARBHANI
23 PUNE
24 RAIGAD
25 RATNAGIRI
26 SANGLI
27 SATARA
28 SINDHUDURG
29 SOLAPUR
30 THANE
31 WARDHA
32 WASHIM
33 YAVATMAL
34 MUMBAI

Choose Corresponding Number to Your City from the Given List:
> 23
[1] 23
>
```

- User is asked to enter the number corresponding to the city.
- The data-set is loaded in ‘rainfallavailbaledata’ variable.
- Few of the starting entries are shown to confirm that he data is loaded.

```
        State   District Year January February March April  May  June  July August September
1 MAHARASHTRA AHMEDNAGAR 2004       0        0     0     0 17.1 123.6 103.5   99.7     262.0
2 MAHARASHTRA AHMEDNAGAR 2005     0.8        0   0.1   1.8  0.7  82.7 146.9   79.6     214.0
3 MAHARASHTRA AHMEDNAGAR 2006       0        0     0     0  8.1 166.1 124.3  176.2     225.9
4 MAHARASHTRA AHMEDNAGAR 2007       0        0     0     0    0 205.2 101.8  126.4     157.8
5 MAHARASHTRA AHMEDNAGAR 2008       0        0  51.9   4.5    0  37.7  65.2  108.4     327.3
6 MAHARASHTRA AHMEDNAGAR 2009       0        0     0     0    0  75.6 122.0  127.3     123.0
  October November December Annual_Total
1    60.2     12.2        0        678.3
2   114.9        0        0        641.5
3    71.2        0        0        771.8
4     0.0        0        0        591.2
5    53.0      0.9        0        648.9
6    73.4    127.6      2.9        651.8
```

- The entries corresponding to the slectes city are loaded as a vector and mean is calculated of the annul_total.> 81.50238
- The temprainfall.csv dataset is then loaded.
- A linearmodel is created between mean temp and mean rainfall.

```
> summary(linearMod)

Call:
lm(formula = mean_temp ~ mean_rainfall, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.6381  -2.7697   0.8757   3.3661   9.7497 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.504e+01  7.505e-01  33.361   <2e-16 ***
mean_rainfall 2.489e-04  3.797e-03   0.066    0.948    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.2 on 89 degrees of freedom
Multiple R-squared:  4.829e-05,	Adjusted R-squared:  -0.01119 
F-statistic: 0.004298 on 1 and 89 DF,  p-value: 0.9479

> linearMod

Call:
lm(formula = mean_temp ~ mean_rainfall, data = dataset)

Coefficients:
  (Intercept)  mean_rainfall  
    2.504e+01      2.489e-04  
```

Then croprequiredtemparature dataset is loaded and printed for confirmation.
Then crops having required minimum-maximum rainfall falling inside average rainfall are found out.

```
> predictedcroprain <- subset(croprainfalltempdata, (croprainfalltempdata$min_rainfall<=avgrainf
all & croprainfalltempdata$max_rainfall>=avgrainfall), select = c("crop_name"))
> print(predictedcroprain)
   crop_name
2     cotton
5      jowar
7      moong
8      maize
9      wheat
10 groundnut
```

5.3. pH CONTRIBUTION

We used 2 datasets.

Pro1.csv consists of

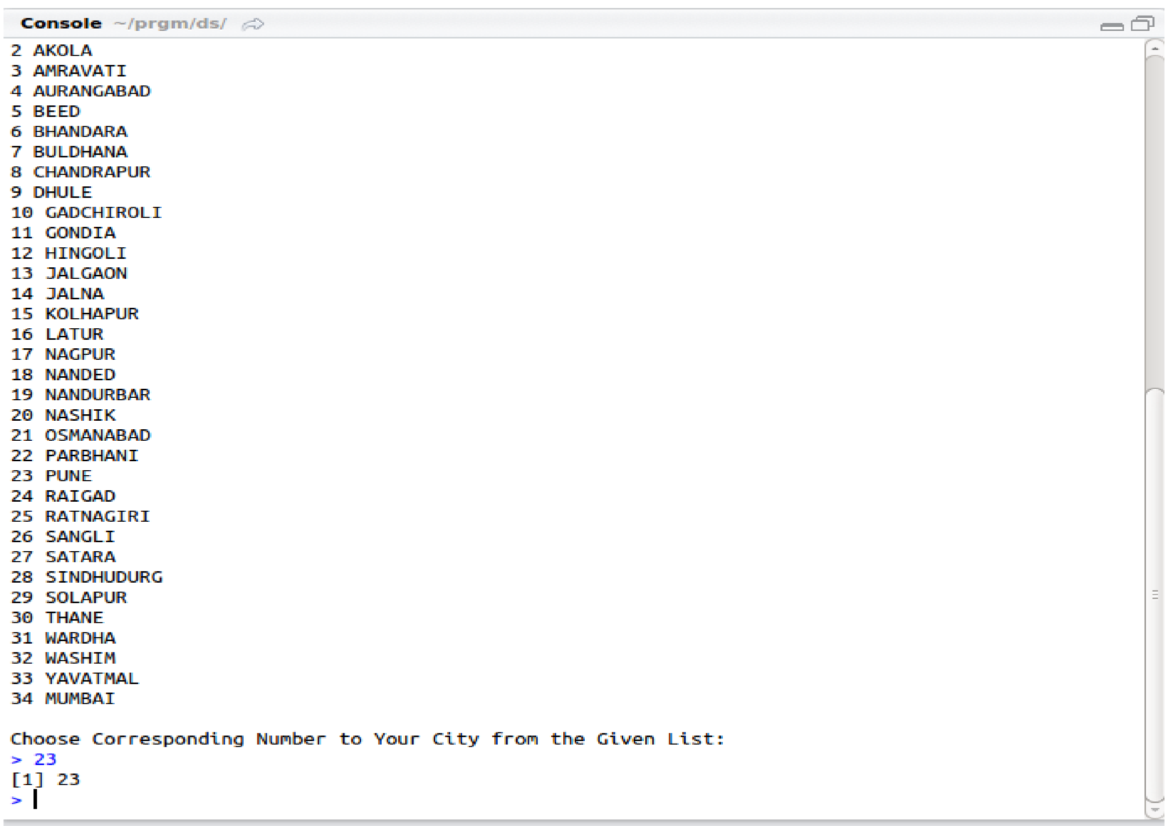
- City name
- pH min
- pH max

Pro2.csv consists of

- Crop name
- pH required
- pH min range
- pH max range
- pH min

Execution step by step process:

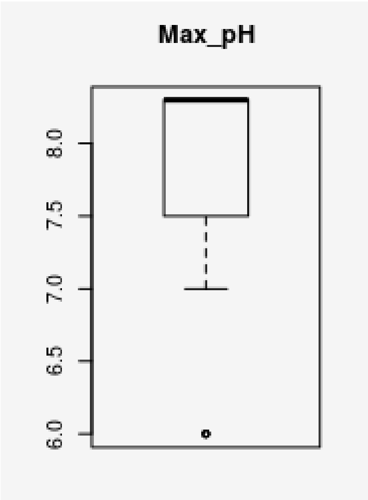
A list of 34 prominant cities are shown to choose from all the districts.



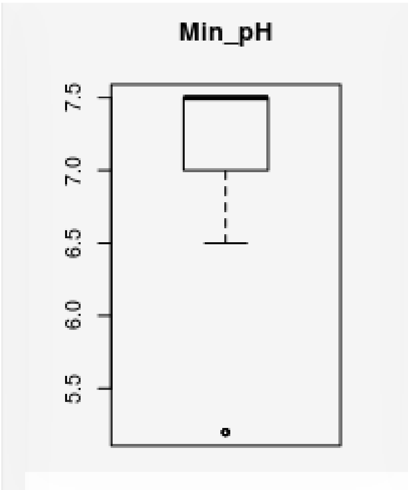
User is asked to enter the number corresponding to the city.
The data-set pro1.csv and pro2.csv is loaded in 'data1' and 'data2' variable respectively.
Entry corresponding to selected city is loaded and printed.

```
> print(subval1)
   City pH_min pH_max
23 PUNE    7.5    8.3
```

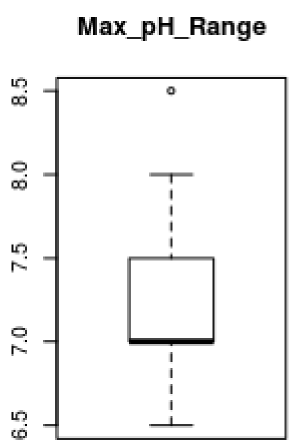
subval2 <- subset(data2,data2$pH_Req <= subval1$pH_min & data2$pH_max_Range >= subval1$pH_max, select = c("Crop"))
Crops ph_req and ph_max_range falling under the range of city's ph_min and ph_max are selected.
Max_pH outliers are plotted.



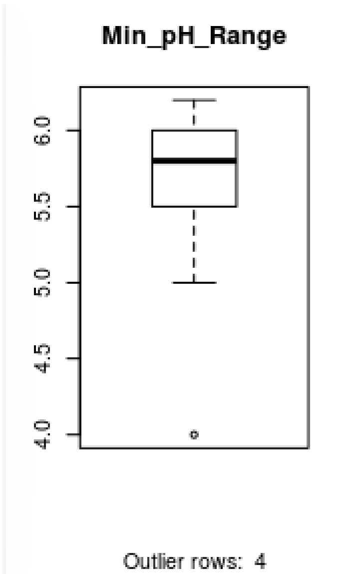
Outlier rows: 6



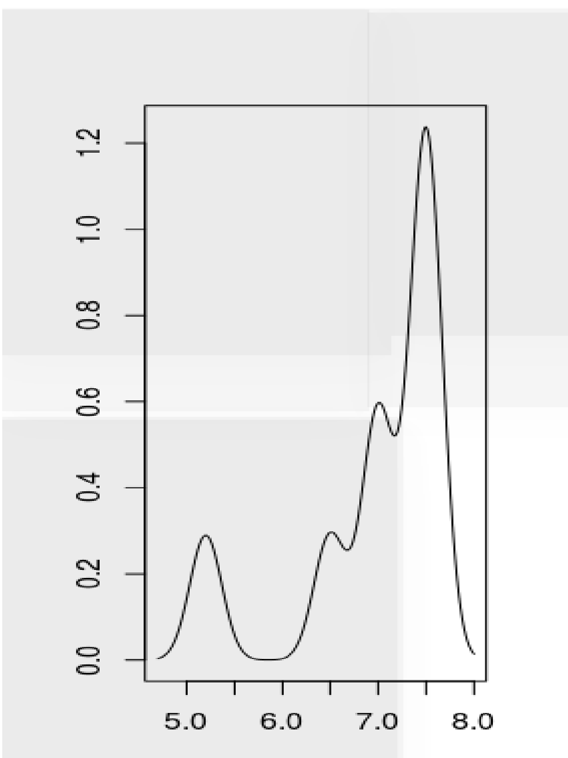
Outlier rows: 5.2

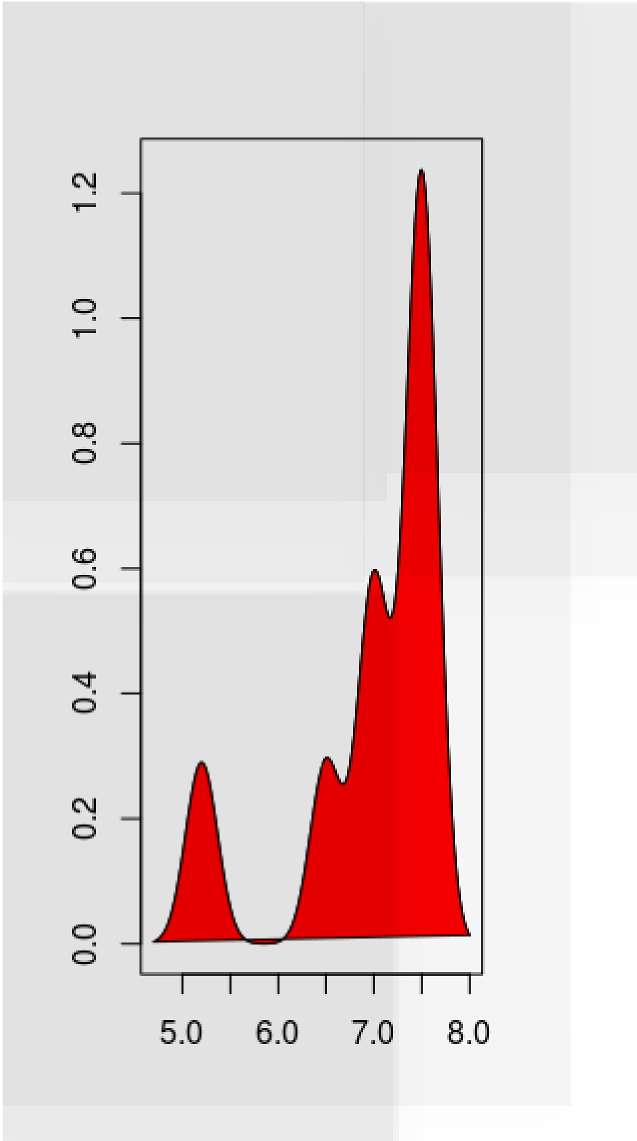


Outlier rows: 8.5



Skewness of city's pH_min:





0.0
0.2
0.4
0.6
0.8
1.0
1.2
5.0
6.0
7.0
8.0

CHAPTER 6
SOFTWARE REQUIREMENT SPECIFICATION

6.1 INTRODUCTION

This Software Requirements Specification (SRS) specifies the requirements of the Crop Yield Prediciton Using Machine Learning, which will be used by the user. This document will be used by the user to ensure all specifications are correct and verified by the software engineer to design the system.

6.2 PURPOSE

This Software Requirements Specification (SRS) specifies the requirements of the Crop Yield Prediction and analysis which will be used by farmers. This document will be used by the user to ensure all specifications are correct and verified by the software engineer to design the system.

6.3 SCOPE

As farmers are now becoming more and more techno sevy and modern and have demand for accurate and quick response system. India's most of the population directly or indirectly depend on the agriculture. They used modern equipment for the farming to grow the productivity of the farm. But still there are many problem present to them which stop the crop to grow its maximum capacity. One of the major is crop disease which restrict the crop to produce its maximum outcomes. Because of lack of proper crop detection system farmer facing the lots of problem in their life. Most important it restrict the maximal production and quality yield.

To overcome this problem of low production and quality yield we are designing this system for farmers "Crop Yield Prediction and Recommendation Using Machine Learning". We have collected three different datasets which are different regions in Maharashtra for crop production, rainfall and temperature dataset and soil pH value dataset.

On successful collection of three datasets, we performed data preprocessing and trained our values in the specific machine learning model. The test data was used for testing and calculating accuracy and precision of model. The results were plotted on the graph and recommender system was developed.

6.4 DEFINITIONS, ACRONYMS AND ABBREVIATIONS

Agriculture : A primary source of Indian economy.
Farmer : The person who does the farming.
User : The person who will be using the system.

6.5 REFERENCES

- www.google.com
- http://ieeexplore.ieee.org/Xplore
- http://www.ijcte.org/
- http://www.ijca.org/

6.6 OVERVIEW

The first section of SRS gives a brief idea on Crop Recommendation System. This section also provides the reference information for further study, design and implementation of the product.

The second section provides an overall description of the application, product functions, operating environment, design and implementation constraints, assumptions and data inputs required.

The third section gives details of non-functional requirements such as performance requirements, security requirements, safety requirements etc.

The problem of	No proper system available to reccomend the crop with higher accuracy and different crop, parameter consideration. so solution to this is our system
Affects	Lower Yield production, Problem in Farmers life, Low quality Yield
The impact of which is	Proper crop recommendation tends to Higher Yield Production
A successful solution would	Easily accessible and More Accurate system.

6.7 PRODUCT POSITION STATEMENT

For	Farmers
Who	Want to analysis the crop disease.
The (product name)	Is Recommendation of crop and yield prediction
That	Recommends accurate crop to the farmers
Unlike	Tradition approach
Our product	Is more accurate

6.8 HARDWARE INTERFACES

- Processor: Intel Pentium or More
- Ram: 512 MB Ram
- Hard Disk: PC with 20GB
- Internet-phone

6.9 SOFTWARE INTERFACES

- Operating System : Windows XP or later, Unix
- Frame Work : Keras, Tensorflow
- Tools: Python IDE, Pycharm, R Studio
- Code Behind: Python, R

6.10 FUNCTIONS

- DATA COLLECTION
- PREPROCESSING
- TRAIN CLASSIFIER
- CLASSIFICATION
- CROP RECOMMENDATION
- YIELD PREDICTION

6.11 SOFTWARE SYSTEM ATTRIBUTES

We are proposing Crop Recommendation system which can maintain, create data in huge amount and does not affect on its scalability, growth facilitation, speed etc. So this system is very easy to use for user.

In this application, there are following specific system attributes that need to be considered. These attributes can be Reliability, availability, Accuracy, etc.

Reliability

The application is reliable solution to the farmer which will be effectively recommendation of crop.

Availability

Application should be available any time to the farmers for crop recommendation.

Accuracy

The application should be provide higher accuracy as we choose the most effective Machine Learning Algorithm.

CHAPTER 7

RESULT AND DISCUSSION

In this proposed work, we have recommended crops to the farmers according to the area of production, temperature, and rainfall and pH value of the soil. We predict the yield of the crop and recommend farmers crop suitable for farming which will give them best results.

```
> print(subval2)
       Crop
1 Sugarcane
> data <- merge(data1,data2,by="pH_min")
> summary(data)
     pH_min              City        pH_max               Crop        pH_Req     
 Min.   :6.500   AURANGABAD: 5   Min.   :7.000   Bajra    :17   Min.   :5.000  
 1st Qu.:7.000   BULDHANA  : 5   1st Qu.:7.500   Coffee   :17   1st Qu.:5.800  
 Median :7.500   CHANDRPUR : 5   Median :8.300   Groundnut:17   Median :6.300  
 Mean   :7.308   DHULE     : 5   Mean   :8.012   Moong    :17   Mean   :6.135  
 3rd Qu.:7.500   HINGOLI   : 5   3rd Qu.:8.300   Wheat    :17   3rd Qu.:6.500  
 Max.   :7.500   JALGAON   : 5   Max.   :8.300   Cotton   : 8   Max.   :6.800  
                 (Other)   :95                   (Other)  :32                  
  pH_min_Range     pH_max_Range  
 Min.   :4.000   Min.   :6.500  
 1st Qu.:5.500   1st Qu.:7.000  
 Median :5.500   Median :7.000  
 Mean   :5.504   Mean   :7.161  
 3rd Qu.:6.000   3rd Qu.:7.200  
 Max.   :6.200   Max.   :8.500  
```

Figure 6.1. Crop suitable for that region in given pH scale

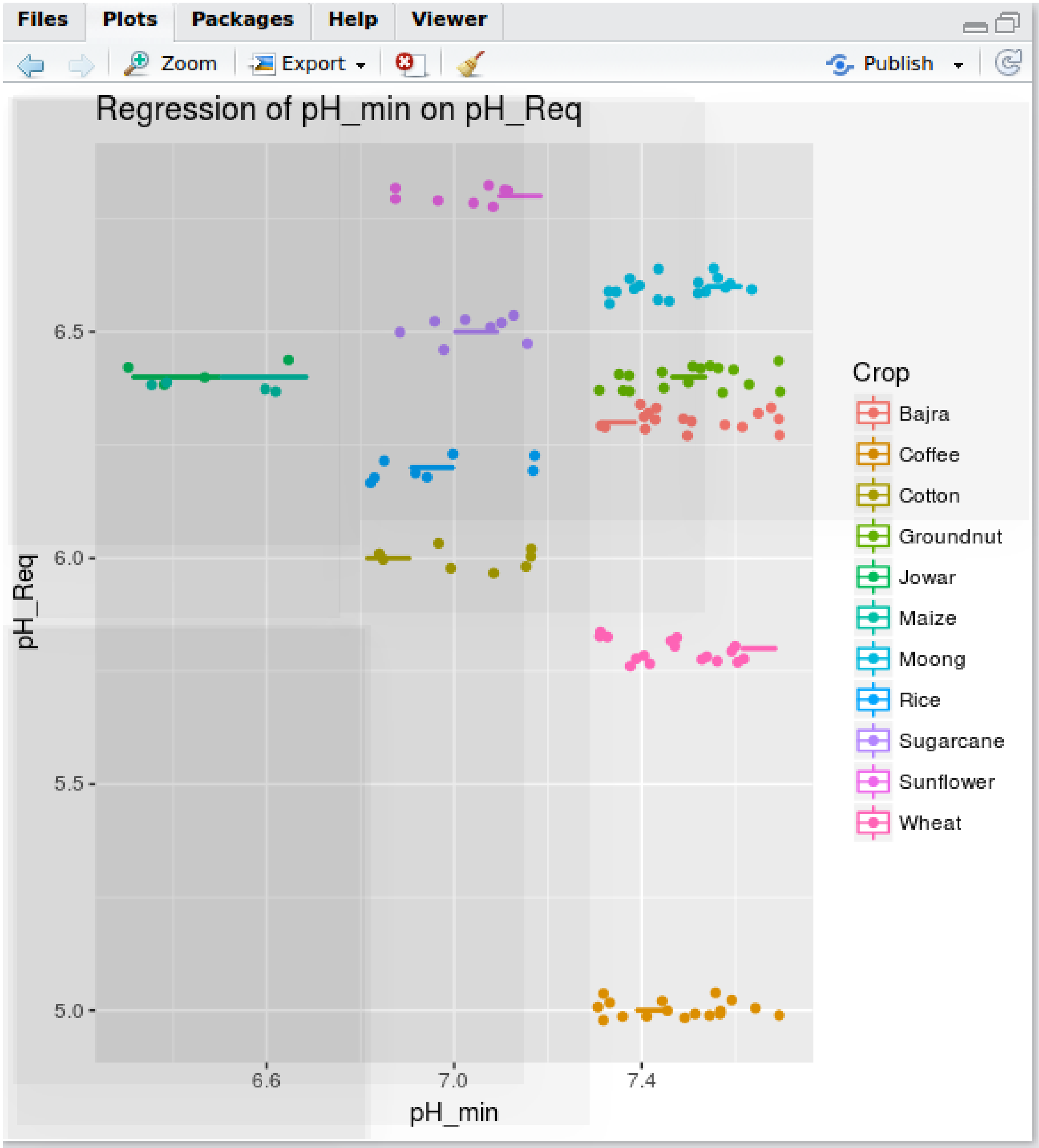


Figure 6.2. Regression of pH min and pH required

The following main results were obtained:

- The analysis of rainfall and temperature areawise helps us determine the most suitable range required for different crops.
- Analyzing pH values of land with the above analysis of rainfall and temperature, a regression model is created to help predict which crop is most suitable in which area.
- Area vs Production helps to reinforce the correctness of this model.

CHAPTER 8

CONCLUSION

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. The proposed machine recommends the suitable crop given type of the soil and the parameters like temperature, rainfall and pH value with classification of the soil class intermediate. The system builds up a Classifier Model, using linear regression, which acts as a classifier builder. The Pre-processor used to remove the noise from the data (all three datasets). This could then ease the work of Classifier Evaluator, to predict the crops with improved accuracy. This model recommends crops to the farmers and predicts the yield of the crops.

CHAPTER 9

FUTURE SCOPE

Machine In future, this project work may be extended by applying different prediction techniques like fuzzy logic, decision trees, Support Vector Regression (SVR) and other machine learning approaches. This work outperforms many models but research can be done to recommend crop with accuracy that is better than what this work proposes.

Further, Correlation between prediction variables may be found which gives importance for variable for prediction and recommendation.

REFERENCES

[1] Aggarwal Sachin (2001). Application of Neural Network to Forecast Air Quality Index. Thesis submitted in partial fulfillment of requirements for a degree in Bachelor of Technology, April 2001.

[2] B. J I ET AL Artificial neural networks for rice yield prediction in mountainous regions. Journal of Agricultural Science (2007), 145, 249– 261.

[3] B.A. Smith et al,.Artificial Neural Networks for Automated Year around Temperature Prediction. Computers and Electronics in Agriculture 68 (2009) 52–6.

[4] Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective Statistical Science, 9: 2-54.

[5] D.L. Ehret et al, Neural network modeling of greenhouse tomato yield, growth and water use from automated crop monitoring data. Computers and Electronics in Agriculture 79 (2011) 82–89.

[6] Enfield, D. B., 1996. Relationships of inter-American rainfall to tropical Atlantic and Pacific SST variability. Geophysical Research

[7] Everingham, Y. L., R. C. Muchow, R. C. Stone, and D. H. Coomans, 2003. Using southern oscillation index phases to forecast sugarcane yields: a case study for Northeastern Australia. International Journal of Climatology 23(10): 1211-1218.

[8] Handler, P, 1990. USA corn yields, the El Niño and agricultural drought: 1867-1988. International Journal of Climatology 10(8): 819-828.

[9] Hansen, J. W., J. W. Jones, C. F. Kiker, A. W. Hodges, 1999. El Niño-Southern Oscillation impacts on winter vegetable production in Florida. Journal of Climate 92-102.

[10] Hansen, J. W., A. W. Hodges, and J. W. Jones, 1998. ENSO Influences on agriculture in the southeastern United States. Journal of Climate 11(3): 404-411.

[11] Haykin, S, 1999. Neural Networks: A Comprehensive Foundation (Second Edition). Upper Saddle River, NJ: Prentice Hall.

[12] Izaurralde, R. C., N. J. Rosenberg, R. A. Brown, D. M. Legler, M. T. Lopez, R. Srinivasan, 1999. Modeled effects of moderate and Los Niños‟ on crop productivity in North America. Agricultural and Forest Meteorology 94(3): 259-268.

[13] Jain, A., R. W. McClendon, G. Hoogenboom, and R. Ramyaa, 2003. Prediction of frost for fruit protection using artificial neural networks. American Society of Agricultural Engineers, St. Joseph, MI, ASAE Paper 03-3075.

[14] Jain, A., R. W. McClendon, and G. Hoogenboom, 2006. Freeze prediction for specific locations using artificial neural networks. Transactions of the ASABE 49(6): 1955-1962.

[15] Jones, J. W., G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. Ritchie, 2003. The DSSAT cropping system model. European Journal of Agronomy 18(3): 235-265.