A Project Report

on

CROP QUALITY AND YIELD PREDICTION BY USING DEEP LEARNING

Submitted in partial fulfillment of the
requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

Under The Supervision of
Dr.Basetty Malligarjuna
Assistant Professor
Department of Computer Science and Engineering

Submitted By

18SCSE1010331-PAWAN KUMAR
18SCSE1010463-PASHPATI NATH SRIVASTAVA

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA, INDIA DECEMBER - 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the project, entitled "**CROP QUALITY AND YIELD PREDICTION BY USING DEEP LEARNING**" in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of **JULY-2021 to DECEMBER-2021**, under the supervision of **Dr.Basetty Malligarjuna, Assistant Professor, Department of Computer Science and Engineering** of School of Computing Science and Engineering , Galgotias University, Greater Noida.

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

18SCSE1010331-PAWAN KUMAR

18SCSE1010463-PASHPATI NATH SRIVASTAVA

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor

(Dr.Basetty Malligarjuna, Assistant Professor)

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **PAWAN KUMAR-18SCSE1010331,PASHPATI NATH SRIVASTAVA-18SCSE1010463** has been held on _____ and his/her work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING.**

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December,2021

Place: Greater Noida

Abstract

The impact of climate change in India, most of the agricultural crops are being badly affected in terms of their performance over a period of the last two decades. Predicting the crop yield in advance of its harvest would help the policy makers and farmers for taking appropriate measures for marketing and storage. This project will help the farmers to know the yield of their crop before cultivating onto the agricultural field and thus help them to make the appropriate decisions. It attempts to solve the issue by building a prototype of an interactive prediction system. Implementation of such a system with an easy-to-use web based graphic user interface and the machine learning algorithm will be carried out. The results of the prediction will be made available to the farmer. Thus, for such kind of data analytics in crop prediction, there are different techniques or algorithms, and with the help of those algorithms we can predict crop yield. Random forest algorithm is used. By analysing all these issues and problems like weather, temperature, humidity, rainfall, moisture, there is no proper solution and technologies to overcome the situation faced by us. In India, there are many ways to increase the economic growth in the field of agriculture. Data mining is also useful for predicting crop yield production. Generally, data mining is the process of analysing data from various viewpoint and summarizing it into important information. Random forest is the most popular and powerful supervised machine learning algorithm capable of performing both classification and regression tasks, that operate by constructing a multitude of decision trees during training time and generating output of the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Table of Contents

Title		Page No.
Candidates Declaration		
Acknowledgement		
Abstract		
Contents		
List of Table		
List of Figures		
Acronyms		
Chapter 1	**Introduction**	**1**
	1.1 Introduction	**2**
	1.2 Required Tools	
Chapter 2	**Literature Survey/Project Design**	**3**
	2.1 Literature Survey	
	2.2 Project Design	**6**
	2.2.1 Review protocol	
	2.2.2 Research Question	
	2.2.3 Research Strategy	
	2.2.4 Exclusion Criteria	
Chapter 3	**Functionality/Working of Project**	**11**
	3.1 Deep learning-based crop yield prediction	
	3.2 Methodology	
Chapter 4	**Results and Discussion**	**22**
	4.1 Result	**25**
Chapter 5	**Conclusion and Future Scope**	**30**
	5.1 Conclusion	**31**
	5.2 Future Scope	**31**
	Reference	**32**

List of Table

S.No.	Caption	Page No.
1	Distribution of papers based on the databases	10
2	Dataset	14
3	Description of the crop dataset	22
4	Performance comparison of feature selection methods based on soil characteristics	24
5	All featured	28
6	Grouped Featured	28

List of Figures

S.No.	Title	Page No.
1	Details of the Plan Review Step	7
2	Details of the Conducting Review Step	8
3	Details of the Reporting Review Step	8
4	Proposed Approach	12
5	Home page	13
6	Feature Diagram	26

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

CHAPTER-1

Introduction

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

The main objectives are

a. To use machine learning techniques to predict crop yield.

b. To provide easy to use User Interface.

c. To increase the accuracy of crop yield prediction.

d. To analyse different climatic parameters (cloud cover, rainfall, temperature

Required Tools

In this Project we are going to develop a model that will predict Crop Quality and yield so we required to use Machine Learning and Data Science Algorithms, Agriculture datasets that are available on government websites as well as Kaggle Platform, Jupyter Notebook or Google Colab Environement , knowledge of Python, and data cleaning algorithms.

CHAPTER-2

Literature Survey/Project Design

2.1 LITERATURE SURVEY

In [1] Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Science Research Technology. This paper focuses on predicting the yield of the crop based on the existing data by using Random Forest algorithm. Real data of Tamil Nadu were used for building the models and the models were tested Pashpati Nath Srivastava Bachelor of technology in Computer Science and Engineering School Of Computer Science And Engineering,Greater Noida ,UP,India . Pashpati007@gmail.com Pawan Kumar Bachelor of technology in Computer Science and Engineering School Of Computer Science And Engineering,Greater Noida ,UP,India . Pawanchahar612@gmail.com BT4409 with samples. Random Forest Algorithm can be used for accurate crop yield prediction.

In [2] Random forests for global and regional crop yield prediction. PLoS ONE Journal. Our generated outputs show that RF is an effective and adaptable machine-learning method for crop yield predictions at regional and global scales for its high accuracy and precision, ease of use, and utility in data analysis. Random Forest is the most efficient strategy and it outperforms multiple linear regression (MLR).

In [3]. Crop production Ensemble Machine Learning model for prediction. International Journal of Computer Science and Software Engineering (IJCSSE). In this paper, AdaNaive and AdaSVM are the proposed ensemble model used to project the crop production over a time period.

Implementation done using AdaSVM and AdaNaive. AdaBoost increases efficiency of SVM and Naive Bayes algorithm.

In [4]. Machine learning approach for forecasting crop yield based on parameters of climate. The paper provided in International Conference on Computer Communication and Informatics (ICCCI). In the current research a software tool named Crop Advisor has been developed asa user friendly web page for predicting the influence of climatic parameters on the crop yields.C4.5 algorithm is used to produce the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh. The paper is implemented using Decision Tree.

In[5]. Prediction On Crop Cultivation. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 10, October 2016. Presently, soil analysis and interpretation of soil test results is paper based. This in one way or another has contributed to poor interpretation of soil test results which has resulted into poor recommendation of crops, soil amendments and fertilizers to farmers thus leading to poor crop yields, micro-nutrient deficiencies in soil and excessive or less application of fertilizers. Formulae to Match Crops with Soil, Fertilizer Recommendation.

In [6]. Analysis of Crop Yield Prediction by making Use DataMining Methods. IJRET: The paper provided in International Journal of Research in Engineering and Technology. In this paper the main aim is to create a user-friendly interface for farmers, which gives the analysis of rice production based on the available data. For maximizing the crop productivity various Data mining techniques were used to predict the crop yield. Such as K-Means algorithm to forecast the pollution factor in the atmosphere.

In [7]. Applications of Machine Learning Techniques in Agricultural Crop Production. Indian Journal of Science and Technology, Vol 9(38), DOI:10.17485/ijst/2016/v9i38/95032, October 2016. From GPS based colour images is provided as an intensified indistinct cluster analysis for classifying plants, soil and residue regions of interest. The paper includes various parameters which can help the crop yield for better enhancement and ratio of the yield can be increased during cultivation.

In [8] In this paper, we present a comprehensive review of research dedicated to the application of machine learning in agricultural production systems. Machine learning (ML) has emerged together with big data technologies, techniques, methods and high-performance computing to generate new opportunities to unravel, quantify, and analyse data intensive processes in agricultural operational sectors. By using Support Vector Machines (SVP) the Paper isImplemented.

In [9]. A Study to Determine Yield for Crop Insurance using Precision Agriculture on an Aerial Platform. Symbiosis Institute of Geoinformatics Symbiosis International University 5th & 6th Floor, Artur Centre, Gokhale Cross Road, Model Colony, Pune – 411016. Precision agriculture (PA) is the application of geospatial methodologies and remote sensors to identify variations in the field and to deal with them using different strategies. The causes of variability of crop growth in an agricultural field might be due tocrop stress, irrigation practices, incidence of pest and disease etc. The Paper is Implemented using Ensemble Learning (EL).

In [10]. Random Forests for Global and Regional Crop Yield Predictions. institute on the Environment, University of Minnesota, St. Paul, MN 55108, United States of America. The generated outputs show that RF is an effective and different machine-learning method for crop

yield predictions at regional and global scales for its high accuracy. The Paper is Implemented using k-nearest neighbour, Support Vector Regression (SVG).

2.2 PROJECT DESIGN

2.2.1 Review protocol

Before conducting the systematic review, a review protocol is defined. The review has been done using the well-known review guidelines provided by Kitchenham et al. (2007). Firstly, the research questions are defined. When research questions are ready, databases are used to select the relevant studies. The databases that were used in this study are Science Direct, Scopus, Web of Science, Springer Link, Wiley, and Google Scholar. After the selection of relevant studies, they were filtered and assessed using a set of exclusion and quality criteria. All the relevant data from the selected studies are extracted, and eventually, the extracted data were synthesized in response to the research questions. The approach we followed can be split up into three parts: plan review, conduct review, and report review. The first stage is planning the review. In this stage, research questions are identified, a protocol is developed, and eventually, the protocol is validated to see if the approach is feasible. In addition to the research quetions, publication venues, initial search strings, and publication selection criteria are also defined. When all of this information is defined, the protocol is revised one more time to see if it represents a proper review protocol

The second stage is conducting the review. When conducting the review, the publications were selected by going through all the databases. The data was extracted, which means that their

information regarding authors, year of publication, type of publication, and more information regarding the research questions were stored. After all the necessary data was extracted correctly, the data was synthesized in order to provide an overview of the relevant papers published so far. In the final stage, a.k.a., Reporting the Review, the review was concluded by documenting the results and addressing the research questions.

2.2.2 Research questions

This SLR aims to get insight into what studies have been published in the domain of ML and crop yield prediction. To get insight, studies have been analyzed from several dimensions.

For this SLR study, the following four research questions(RQs) have been defined.

- RQ1- Which machine learning algorithms have been used in the literature for crop yield prediction?
- RQ2- Which features have been used in literature for crop yield prediction using machine learning?
- RQ3- Which evaluation parameters and evaluation approaches have been used in literature for crop yield prediction?
- RQ4- What are challenges in the field of crop yield prediction using machine learning?

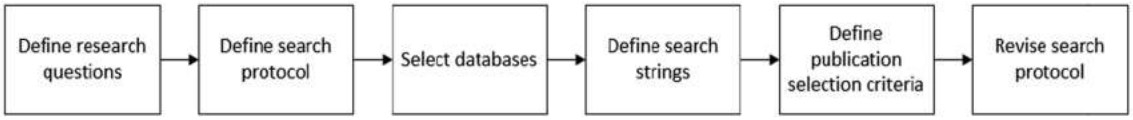

Fig 1. Details of the Plan Review Step

2.2.3. Search strategy

The searching is done by narrowing down to the basic concepts that are relevant for the scope of this review. Machine learning has many

Fig 2. Details of the Conducting Review Step

application fields, which means that there are a lot of published studies that are probably not in the scope of this review article. The basic searching is done by an automated search. The starting input for the search was “machine learning” AND “yield prediction”. Articles were retrieved, and abstracts were read to find the synonyms of the keywords. The search was performed in six databases. The search input “machine learning” AND “yield prediction” was used to get a broad view of the studies. After the exclusion criteria were applied, and all the results were processed, and a more complex search string was built in order to avoid missing relevant studies. This final search string is as follows: ((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”)). After executing this search string, 567 studies were retrieved.

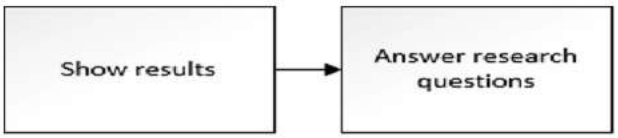

Fig 3. Details of the Reporting Review Step

A specific description of the search strings per database are provided as follows: Science direct: The search string is [“machine learning” AND “yield prediction”] (Title, abstract, keywords) and [((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction”

OR “yield forecasting” OR “yield estimation”))](Title, abstract, keywords). Scopus: The search string is [“machine learning” AND “yield prediction”](Title, abstract, keywords) and [((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”))] (Title, abstract, keywords). Web of Science: The search string is [“machine learning” AND “yield prediction”] (title, abstract, author keywords, and Keywords Plus). Springer Link: The search string is [“machine learning” AND “yield prediction”](anywhere) and [((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”))] (anywhere) Wiley: The search string is [“machine learning” AND “yield prediction”] (anywhere). Google Scholar: The search string is [“machine learning” AND “yield prediction”] (anywhere) and [((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”))] (anywhere). For Web of Science and Wiley, the search string [((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”))] did not result in any publications.

2.2.4 Exclusion criteria

To exclude irrelevant studies, the studies were analyzed and graded based on exclusion criteria to set the boundaries for the systematic review. The exclusion criteria (EC) are shown as follows:

Exclusion criteria 1 - Publication is not related to the agricultural sector and yield prediction combined with machine learning

Exclusion criteria 2 – Publication is not written in English

Exclusion criteria 3 – Publication that is a duplicate or already retrieved from another database

Exclusion criteria 4 – Full text of the publication is not available

Exclusion criteria 5 – Publication is a review/survey paper

Exclusion criteria 6 – Publication has been published before 2008

Database	# of initially retrieved papers	# of papers after exclusion criteria	Percentage of Papers (%)
Science Direct	17	4	8
Scopus	68	11	22
Web of Science	32	0	0
Springer Link	132	10	20
Wiley	20	1	2
Google Scholar	298	24	48
Total	567	50	100

Table 1. Distribution of papers based on the databases

After the first three exclusion criteria were applied, only 77 studies remained for further analysis. After applying all the six exclusion criteria, 50 studies were selected for further analysis. In Table 1, we show the number of initially retrieved papers and the number of papers after selection criteria were applied. Fig. 4 shows the distribution of selected publications based on the databases we searched. As shown in Table 1, most of the papers were retrieved from Google Scholar, Scopus, and Springer databases. To answer the four research questions, data from the selected studies have been extracted and synthesized. The information retrieved was focused on checking whether or not the studies meet the requirements stated in the exclusion criteria and on responding to the research questions. The selected studies that passed the exclusion criteria are presented in Appendix A. During the data synthesis, all the extracted data have been combined and synthesized, and the research questions were answered accordingly.

CHAPTER-3

Functionality/Working of Project

Data is a very important part of any Machine Learning System. To implement the system, we decided to focus on Maharashtra State in India. As the climate changes from place to place, it was necessary to get data at district level. Historical data about the crop and the climate of a particular region was needed to implement the system. This data was gathered from different government websites. The data about the crops of each district of Maharashtra was gathered from www.data.gov.in and the data about the climate was gathered from www.imd.gov.in. The climatic parameters which affect the crop the most are precipitation, temperature, cloud cover, vapour pressure, wet day frequency. So, the data about these climatic parameters was gathered at a monthly level. Dataset Collection: In this phase, we collect data from various sources and prepare datasets. And the provided dataset is in the use of analytics (descriptive and diagnostic). There are several online abstracts sources such as Data.gov.in and indiastat.org. For at least ten years the yearly abstracts of a crop will be used. These datasets usually accept behaviour of anarchic time series. Combined the primary and necessary abstracts. Random Forests for Global and Regional Crop Yield Predictions. Data Partitioning: The Entire dataset is partitioned into 2 parts: for example, say, 75% of the dataset is used for training the model and 25% of the data is set aside to test the model. To predict future events Machine Learning Algorithms: Supervised learning: Supervised machine learning algorithms can apply what has been learned in the past to new data using labelled examples. After Sufficient training the system can provide targets for

any new input. IN order to change the model accordingly the learning algorithm can also differentiate its results with the correct, intended output and find errors. Unsupervised learning: IN comparison, unsupervised machine learning algorithms are used when the information used to train is neither labelled nor classified. Unsupervised learning does analysis of how systems can infer a function to describe a hidden structure from unlabelled data. In order to describe hidden structures from unlabelled data the system doesn't figure out the right output, but it examinesthe data and can draw inferences from datasets. Random Forest Classifier: Random forest is the most popular and powerful supervised machine learning algorithm capable of performing both classification and regression tasks, that operate by constructing a multitude of decision trees at the time of training and generating outputs of the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The more trees in a forest the more robust the prediction

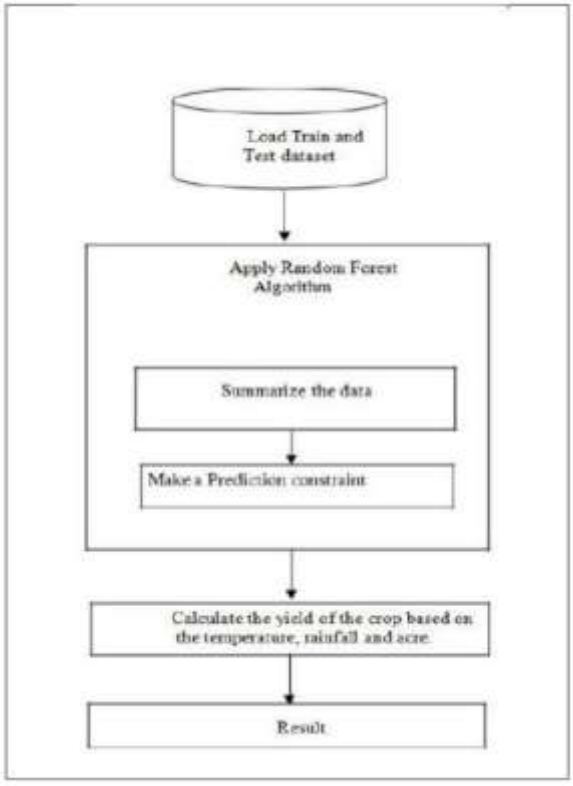

Fig. 4. Proposed Approach

Fig. 1. Shows the proposed approach and how the data is summarized, and Random Forest algorithm is applied, and the result is calculated

IMPLEMENTATION

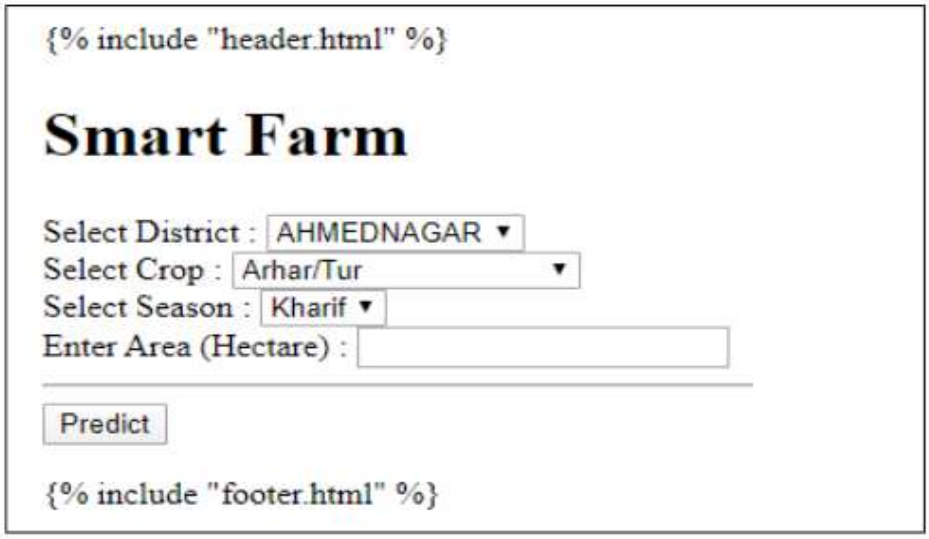

Fig. 5. Home page

Fig. 2 shows the home page of the website where the person accessing the website enters the details such as the district, crop, season and the area in Hectare and by clicking on predict the result is printed

District_Name	Season	Crop	PJan	PFeb	PMar	PApr	PMay	PJ
AHMEDNAGAR	Kharif	Arhar/Tur	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Bajra	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Gram	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Jowar	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Maize	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Moong(Green (	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Pulses total	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Ragi	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Rice	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Sugarcane	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Total foodgrain	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Kharif	Urad	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Rabi	Jowar	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Rabi	Maize	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Rabi	Other Rabi puls	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Rabi	Wheat	3.099	0	1.671	23.129	4.646	
AHMEDNAGAR	Summer	Maize	3.099	0	1.671	23.129	4.646	

Table 2. Data set

Fig. 3. It is the snapshot of the final processed data set that is being used for this project

Deep learning-based crop yield prediction

In the first part of our research (i.e., Systematic Literature Review), we observed that Artificial Neural Networks (ANN) is the most used algorithm for crop yield prediction. Recently, deep learning, which is a sub-branch of machine learning, has provided state-of-the-art results in many different domains, such as face recognition and image classification. These Deep Neural Networks (DNN) algorithms use similar concepts of ANN algorithms; however, they include different hidden layer types such as convolutional layer and pooling layer and consist of many hidden layers instead of a single hidden layer. As such, in the second part of our research, we aimed to investigate to what extent deep learning algorithms have been applied in crop yield prediction. To broaden our analysis and reach recent applications of deep learning algorithms in yield prediction, we designed a new search criterion (i.e., "deep learning" AND "yield prediction") and performed a new search in the same electronic databases that were used during the SLR study. We reached the following 30 papers. We investigated these articles in detail, extracted, and synthesized the deep learning algorithms applied by researchers. The yearly distribution of deep learning-based papers. Although we are in the half of the year 2020, the number of papers that belong to the year 2020 is now equal to the number of papers published in 2019. This shows that the number of papers is increasing every year. In Table 8, we show the distribution of deep learning-based papers per database. Most of the papers were retrieved from Google Scholar, and the second top database was Scopus. Science Direct and Springer Link returned a similar number of deep learning-based paperswe show the distribution of applied deep learning algorithms in the identified papers list. The most applied deep learning algorithm is Convolutional Neural Networks (CNN), and the other widely used algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN) algorithms. Since some papers

applied more than one deep learning algorithm, the total number of usages shown in the second column is larger than the total number of papers. These deep learning algorithms are shortly described as follows:

• Deep Neural Networks (DNN): These DNN algorithms are very similar to the traditional Artificial Neural Networks (ANN) algorithms except the number of hidden layers. In DNN networks, there are many hidden layers that are mostly fully connected, as in the case of ANN algorithms. However, for other kinds of deep learning algorithms such as CNN, there are also different types of layers, such as the convolutional layer and the pooling layer.

• Convolutional Neural Networks (CNN): Compared to a fully connected network, CNN has fewer parameters to learn. There are three types of layers in a CNN model, namely convolutional layers, pooling layers, and fully-connected layers. Convolutional layers consist of filters and feature maps. Filters are the neurons of the layer, have weighted inputs, and create an output value (Brownlee, 2016). A feature map can be considered as the output of one filter. Pooling layers are applied to down-sample the feature map of the previous layers, generalize feature representations, and reduce the

METHODOLOGY

1 .Data Pre-Processing

Data Preprocessing is a method that is used to convert the raw data into a clean data set. The data are gathered from different sources, it is collected in raw format which is not feasible for the analysis. By applying different techniques like replacing missing values and null values, we can transform data into an understandable format. The final step on data preprocessing is the splitting of training and testing data. The data usually tend to be split unequally because training the

model usually requires as much data- points as possible. The training dataset is the initial dataset used to train ML algorithms to learn and produce right predictions (Here 80% of dataset is taken as training dataset)..

2 .Factors affecting Crop Yield and Production

There are a lot of factors that affects the yield of any crop and its production. These are basically the features that help in predicting the production of any crop over the year. In this paper we include factors like Temperature, Rainfall, Area, Humidity and Windspeed .

3 .Comparison and Selection of Machine Learning Algorithm

Before deciding on an algorithm to use, first we need to evaluate and compare, then choose the best one that fits this specific dataset. Machine Learning is the best technique which gives a better practical solution to crop yield problem. There are a lot of machine learning algorithms used for predicting the crop yield. In this paper we include the following machine learning algorithms for selection and accuracy comparison :

- .Logistic Regression:- Logistic regression is a supervised learning classification algorithm used to predict the probability of target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. When logistic regression algorithm applied on our dataset it provides an accuracy of 87.8%.

- Naive Bayes:- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity,

  Naive Bayes is known to outperform even highly sophisticated classification methods. It provides an accuracy of 91.50%.

- Random Forest:- Random Forest has the ability to analyze crop growth related to the current climatic conditions and biophysical change. Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better solution for the system. Random Forest uses the bagging method to train the data which increases the accuracy of the result. For our data, RF provides an accuracy of 92.81%.

  It is clear that among all the three algorithms, Random forest gives the better accuracy as compared to other algorithms.

4 .Random Forest Model for Crop Prediction

Random forests are the aggregation of tree predictors in such a way that each tree depends on the values of a random subset sampled independently and with the same distribution for all trees in

the forest. Random Forest used the bagging method to trained the data which increases the accuracy of the result. For getting high accuracy we used the Random Forest algorithm which gives accuracy which predicate by model and actual outcome of predication in the dataset. The predicted accuracy of the model is analyzed 91.34%.

5 .System Architecture

System architecture represented in the Fig.3 mainly consists of weather API where we fetch the data such as temperature, humidity, rainfall etc. The data fetched from the API are sent to the server module. The data gets stored on to the database on the server. Using the mobile application, the user can provide details like location, area, etc. The user can create an account on the mobile app by one-time registration and all these entered data are sent to server. The trained Random forest model deployed on the server uses all the fetched and input data for crop yield prediction, finds the yield of predicted crop with its name in the particular area.

6 .Proposed System

Our proposed system system is a mobile application which predicts name of the crop as well as calculate its corresponding yield. Name of the crop is determined by several features like temperature, humidity, wind-speed, rainfall etc. and yield is determined by the area and production. In this paper, Random Forest classifier is used for prediction. It will attain the crop prediction with best accurate values.

7 .System Analysis

a. Python 3.8.5(Jupyter Notebook):Python is the coding language used as the platform for machine learning analysis.

Jupyter Notebooks illustrates the analysis process and gives out the needed result.

b. Weather_API (Open Weather Map): Weather API is an application programming interface used to access the current weather details of a location. The generated API key illustrates current weather forecast needed for crop prediction.

c. Android Studio (Version 3.4.1): Android Studio is the official integrated development environment (IDE) for Android application development. This paper uses java as the framework for frontend designing. USB debugging method is used for the connection of IDE and app.

d. Python Flask Framework (Version 2.0.1): Flask is a micro framework in python. Flask is based on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine. In this paper flask is used as the back-end framework for building the application. It is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc.

e. Heroku: Heroku is the container-based cloud platform that allows developers to build, run & operate applications exclusively in the cloud. In this paper Heroku is used for server part. Once created an account in the Heroku we can connect it with the GitHub repository and then deploy.

CHAPTER-4

Results and Discussion

This work utilized an agricultural dataset that chiefly included soil characteristics and environmental factors, collected from the website: www.tnau.ac.in and the Agricultural Department of Sankarankovil Taluk, Tenkasi District, Tamil Nadu, India. The dataset contains 1000 instances, 16 attributes, where 12 attributes are soil characteristics and the remaining 4 environmental characteristics, respectively. The target class is the multiclass representation with 9 classes. The attributes are collected from villages around Sankarankovil.

S.No.	Attributes	Description
	Soil characteristics	
1.	pH (potential of Hydrogen)	pH is the main factor for farming.
2.	EC (Electrical Conductivity)	EC is the numerical parameter that used to measure the salt level in soil and it affects the crop productivity. If EC is 0.01 then the soil is considered for crop cultivation.
3.	OC (Organic Carbon)	OC enters the soil through the decomposition of plant and animal residues, root exudates, living and dead microorganisms, and soil biota.
4.	N (Nitrogen)	Nitrogen is a key element in plant growth.
5.	P (Phosphorus)	Phosphorus helps transfer energy from sunlight to plants, stimulates early root and plant growth, and hastens maturity.
6.	K(Potassium)	Potassium increases vigour and disease resistance of plants, helps form and move starches, sugars and oils in plants, and can improve fruit quality.
7.	S (Sulphur)	Sulphur is a constituent of amino acids in plant proteins and is involved in energy-producing processes in plants.
8.	Z (Zinc)	Zinc helps in the production of a plant hormone responsible for stem elongation and leaf expansion.
9.	B (Boron)	Boron helps with the formation of cell walls in rapidly growing tissue. Deficiency reduces the uptake of calcium and inhibits the plant's ability to use it.
10.	Fe (Iron)	Iron is a constituent of many compounds that regulate and promote growth.
11.	Mn (Manganese)	Manganese helps with photosynthesis.
12.	Cu (Copper)	Copper is an essential constituent of enzymes in plants.
	Environmental Factors	
13.	Texture	It has major influence on crop growth. It influences aeration, water movement etc ...
14.	Season	Season is the challenging factor for crop growth.
15.	Rainfall	Rainfall has the great impact on crop growth. Excessive and insufficient rainfall affects the yield.
16.	Average Temperature	It is important for growth and development.

Table 3. Description of the crop dataset

Table 1 shows a description of the soil and environmental attributes used for crop prediction.

Attributes are selected, using the feature selection method, to find accurate soil and environmental characteristics for predicting a suitable crop for improved cultivation. Classification methods are used with feature selection techniques to find the most suitable crop for a particular stretch of land. The techniques are evaluated thereafter, using parameters such as Attribute Selection, Accuracy, and Error Rate.

1. Performance comparison of feature selection with classifier based on soil characteristics

Table 2 shows a performance evaluation of feature selection methods with classification techniques, based only on soil characteristics such as the Ph,EC, N, P, and K, among others, as described in Table 1.

Feature Selection Method	Classification Method	Attribute Selected		Accuracy		Error Rate	
		Before Reduction	After Reduction	Before Reduction	After Reduction	Before Reduction	Before Reduction
Boruta	kNN	12	12	0.5625	0.5625	0.4375	0.4375
	Naive Bayes			0.6812	0.6812	0.3188	0.3188
	Decision Tree			0.7125	0.7125	0.2875	0.2875
	SVM			0.7750	0.7750	0.225	0.225
	Random Forest			0.8375	0.8375	0.1625	0.1625
	Bagging			0.8875	0.8875	0.1125	0.1125
Sequential Forward Feature Selection	kNN	12	10	0.5625	0.5689	0.4375	0.4311
	Naive Bayes			0.6812	0.69	0.3188	0.31
	Decision Tree			0.7125	0.715	0.2875	0.285
	SVM			0.7750	0.7803	0.225	0.2197
	Random Forest			0.8375	0.8457	0.1625	0.1543
	Bagging			0.8875	0.889	0.1125	0.111
Recursive Feature Elimination	kNN	12	9	0.5625	0.5751	0.4375	0.4249
	Naive Bayes			0.6812	0.7124	0.3188	0.2876
	Decision Tree			0.7125	0.7188	0.2875	0.2812
	SVM			0.7750	0.7837	0.225	0.2163
	Random Forest			0.8375	0.8499	0.1625	0.1501
	Bagging			0.8875	**0.8938**	0.1125	**0.1062**

Table 4. Performance comparison of feature selection methods based on soil characteristics

From the 12 attributes, the RFE, Boruta and SFFS select 9, 12, and 10 attributes, respectively for the crop prediction. Classifier techniques are applied to find the most suitable crop/s, based on soil conditions. Table 2 shows that the RFE selects the most accurate attributes of all the techniques. Further, the RFE with the bagging classifier offers better crop prediction accuracy of 89%, compared with other soil-characteristic-based methods.

It is observed that most of the research works did not concentrate on the various folds or split and have chosen the default parameters. In this research work, the importance is given to data splitting and the experiments are carried out to obtain the optimal data split for training and testing. The performance of feature selection techniques with the classifier based on fold variation and data splitting validation are evaluated using eights metrics which are mentioned above and it helps to find the best fold and data splitting range for predict the suitable crop/s based on soil and environmental characteristics.

4.1 Result

The feature group “soil information” consists of the following variables: soil maps, soil type, pH value, cation exchange capacity, and area of production. Whether or not soil maps were used and the in - formation content of the maps differs among the different publications. In the soil maps, general information about the nutrients in the soil, type of the soil, and location can be found. Crop information refers to information about the crop itself, such as weight, growth during the growth-process, variety of plants, and crop density. Other measure - ments that indicate growth is also included in this group, for example, the leaf area index. Humidity stands for the water in the field. The features that fall under the humidity group include rainfall, humidity, forecasted rainfall, and precipitation. Nutrients can be nutrients that are already in the soil, but the nutrients can also be applied nutrients. These features measure the level of saturation. The measured nutrients are nitrogen, magnesium, potassium, sulphur, zinc, boron, calcium, manganese, and

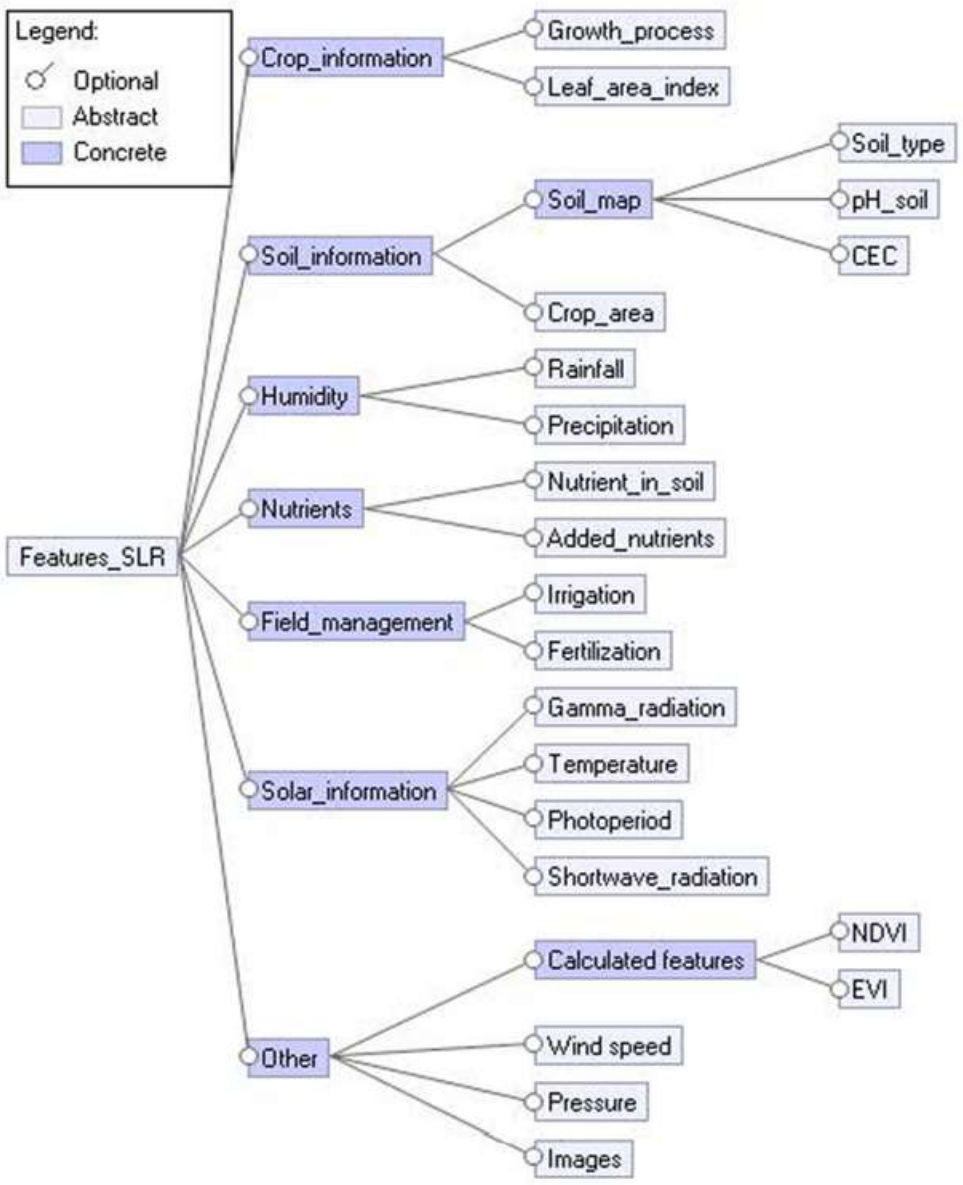

Fig 6. Feature diagram.

phosphorus. With field management, decisions of farmers to adjust their field are grouped. These features are irrigation and fertilization, and thus field management could also refer to the management of nutrients. The solar information contains features re - lated to radiation or temperature. These are gamma radiometric, tem - perature, photoperiod, shortwave radiation,

degree-days, and solar ra - diation. The feature group labeled as ‘Other’ contains the features that cannot be put in any of the groups mentioned above. Most of these features are used only once or are calculated features (Measuring Vegetation (NDVI & EVI), 2000). These features are used less and in - clude features such as wind speed, pressure, and images. The calculated features are MODIS Enhanced Vegetation Index (MODIS-EVI), Nor - Table 2 malized Vegetation Index (NDVI), and Enhanced Vegetation Index

To represent all the features gathered through this SLR study, we drew a feature map depicted in Fig. 6 shows the significant features and sub-features. To address the first research question (RQ1), machine learning algorithms were investigated and summarized. The algorithms used more than once are listed in Table 5. As shown in the table, Neural Networks (NN) and Linear Regression algorithms are the two algorithms used mostly. Also, Random Forest (RF) and Support Vector Machines (SVM) are widely used, according to Table 5.

To address research question three (RQ3), evaluation parameters were identified. All the evaluation parameters that were used and the number of times they were used are shown in Table 6. As the table shows, Root Mean Square Error (RMSE) is the most used parameter in the studies. Apart from the evaluation parameters, several validation approaches were used as well. Most of the time, cross-validation is used. The most used evaluation method was 10-fold cross-validation.

Feature	# of times used
Temperature	24
Soil type	17
Rainfall	17
Crop information	13
Soil maps	12
Humidity	11
pH-value	11
Solar radiation	10
Precipitation	9
Images	8
Area of production	8
Fertilization	7
NDVI	6
Cation exchange capacity	6
Nitrogen	6
Irrigation	5
Potassium	5
Wind speed	5
Zinc	3
Magnesium	3
Shortwave radiation	2
Sulphur	2
Boron	2
Calcium	2
Organic carbon	2
EVI	2
Phosphorus	2
Gamma radiametrics	1
MODIS-EVI	1
Forecasted rainfall	1
Photoperiod	1
Climate	1
Degree-days	1
Time	1
Pressure	1
Leaf area index	1
Manganese	1

Table 5. All features used.

Group	# of times used
Soil information	54
Solar information	39
Humidity	38
Nutrients	28
Other	24
Crop information	14
Field management	12

Table 6. Grouped features.

To address research question four (RQ4), the publications were read to see if they stated any problems or improvements for future models. In several studies, insufficient availability of data (too few data) was mentioned as a problem. The studies stated that their systems worked for the limited data that they had at hand, and indicated data with more variety should be used for further testing. This means data with different climatic circumstances, different vegetation, and longer timeseries of yield data. Another suggested improvement is that more data sources should be integrated. Finally, the publication indicated that the use of machine learning in farm

management systems should be explored. If the models work as requested, software applications must be created that allow the farmer to make decisions based on the models.

Chapter 5

Conclusion and Future Scope

Based on the climatic input parameters the present study provided the demonstration of the potential use of data mining techniques in predicting the crop yield based. The developed webpage is user friendly and the accuracy of predictions are above 75 per cent in all the crops and districts selected in the study indicating higher accuracy of prediction. By providing climatic data of that place the user-friendly web page developed for predicting crop yield can be used by any user their choice of crop Data mining is a scientific field with applications in the study of crop yields. The prediction of crop cultivation is critical to agriculture, with farmers keen to work out how much they can possibly expect to produce. In the past, cultivar prediction was carried out by taking into account farmers' basic understanding of specific stretches of land and the crops to be grown therein. In agriculture, several data mining strategies are used and analysed to predict crop cultivation in the future. This research has focused on comparing the feature selection methods of different prediction models, and suggested the best method to forecast crop cultivation in the future. Our findings from the results obtained conclusively show that Boruta, SFFS, and RFE feature selection techniques with the bagging classifier performing best with 10 fold and 70% – 30% data splitting range and RFE with the bagging method, outperforms other methods.

1 .CONCLUSION

This paper focuses on the prediction of crop and calculation of its yield with the help of machine learning techniques. Several machine learning methodologies used for the calculation of accuracy. Random Forest classifier was used for the crop prediction for chosen district. Implemented a system to crop prediction from the collection of past data. The proposed technique helps farmers in decision making of which crop to cultivate in the field. This work is employed to search out the gain knowledge about the crop that can be deployed to make an efficient and useful harvesting. The accurate prediction of different specified crops across different districts will help farmers of Kerala. This improves our Indian economy by maximizing the yield rate of crop production.

2 .FUTURE SCOPE

In coming years, can try applying data independent system. That is whatever be the format our system should work with same accuracy. Integrating soil details to the system is an advantage, as for the selection of crops knowledge on soil is also a parameter. Proper irrigation is also a needed feature crop cultivation. In reference to rainfall can depict whether extra water availability is needed or not. This research work can be enhanced to higher level by availing it to whole India.

Reference

[1] P.Priya, U.Muthaiah M.Balamurugan.Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Science Research Technology.

[2]. J.Jeong, J.Resop, N.Mueller and team.Random forests for global and regional crop yield prediction.PLoS ONE Journal.

[3].Narayanan Balkrishnan and Dr. Govindarajan Muthukumarasamy.Crop production Ensemble Machine Learning model for prediction. International Journal of Computer Science and Software Engineering (IJCSSE).

[4]. S.Veenadhari, Dr. Bharat Misra, Dr. CDSingh.Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics (ICCCI).

[5]. Shweta K Shahane , Prajakta V Tawale.Prediction On Crop Cultivation. IInternational Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 10, October 2016.

[6]D Ramesh ,B Vishnu Vardhan. Analysis Of Crop Yield Prediction Using Data Mining Techniques. IJRET: International Journal of Research in Engineering and Technology.

[7]Subhadra Mishra,Debahuti Mishra, Gour Hari Santra. Applications of Machine Learning Techniques in Agricultural Crop Production. Indian Journal of Science and Technology, Vol 9(38), DOI:10.17485/ijst/2016/v9i38/95032, October 2016.

[8].Konstantinos G. Liakos,Patrizia Busato,Dimitrios Moshou, Simon Pearson ID,Dionysis Bochtis. Machine Learning in Agriculture. Lincoln Institute for Agri-food Technology (LIAT), University of Lincoln, Brayford Way, Brayford Pool,Lincoln LN6 7TS, UK, spearson@lincoln.ac.uk.

[9]. Baisali Ghosh. A Study to Determine Yield for Crop Insurance using Precision Agriculture on an Aerial Platform. Symbiosis Institute of Geoinformatics Symbiosis International University 5th & 6th Floor, Atur Centre, Gokhale Cross Road, Model Colony, Pune – 411016.

[10]. Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher ,Kyungdahm Yun, Ethan E. Butler,Soo-Hyung Kim. Random Forests for Global and Regional Crop Yield

Predictions. Institute on the Environment, University of Minnesota, St. Paul, MN 55108, United States of America.

[11] Ecochem Online. (2009). Soil Health and Crop yields. Last modified January 28th 2009.Retrieved on March 4th 2009 from http://ecochem.com/healthy_soil.html

[12] Food and Agricultural Organization. (2006). The state of Agricultural Commodity Markets. 37-39.

[13]Aditya Shastry, H.A Sanjay And E.Bhanushree,"Prediction of crop yield using Regression Technique", International Journal of computing r12 (2):96- 102 2017,ISSN:1816-914] E.14]E. Manjula , S. Djodiltachoumy,"A Model for Prediction of Crop Yield", International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4, March 2017

[14]Mrs.K.R.Sri Preethaa, S.Nishanthini, D.SanthiyaK.Vani Shree ,"CropYield Prediction",International Journal On Engineering Technology and Sciences – IJETS™ISSN(P): 2349-3968, ISSN (O):2349-3976 Volume III,Issue III, March- 2016

[15]Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data"Majumdar et al. J Big Data (2017) 4:20 DOI 10.1186/s40537-017-0077-4

[16]D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473, 2015.

[17] Yethiraj N G , " Applying data mining techniques in the field of Agriculture and allied sciences", Vol 01, Issue 02, December 2012.

[18] Zelu Zia (2009). An Expert System Based on Spatial Data Mining used Decision Tree for Agriculture Land Grading. Second International Conference on Intelligent Computation Technology and Automation.Oct10-11,Chin