**A**

**PROJECT REPORT**

**ON**

**“CROP YIELD PREDICTION USING CNN”**

**SUBMITTED**

**TO**

**SHIVAJIUNIVERSITY, KOLHAPUR**

**IN THE PARTIAL FULFILLMENT OF REQUIREMENT FOR THE AWARD OF DEGREE BACHELOR OF ENGINEERING IN COMPUTER SCIENCE AND ENGINEERING**

**SUBMITTED BY**

**Miss. Bhakti Ajay Akole 16UCS11003XX**

**Miss. Prajakta Sudhir Bhosale 16UCS11009XX**

**Mr. Sandip Arun Torane 16UCS12059XX**

**Mr. Abhishek Rajendra Urane 16UCS12062XX**

**Mr. Avdhut Balaso Yadav 16UCS12063XX**

**UNDER THE GUIDANCE OF**

**Prof. K. S. KADAM**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**DKTE SOCIETY’S TEXTILE AND ENGINEERING INSTITUTE, ICHALKARANJI**

**(AN AUTONOMOUS INSTITUTE)**

**ACCREDITED WITH ‘A+’ GRADE BY NAAC**

**An ISO 9001: 2015 Certified**

**SHIVAJI UNIVERSITY KOLHAPUR**

**2019-2020**

**D.K.T.E.SOCIETY’S**

**TEXTILE AND ENGINEERING INSTITUTE, ICHALKARANJI**

**(AN AUTONOMOUS INSTITUTE)**

**ACCREDITED WITH ‘A+’ GRADE BY NAAC**

**An ISO 9001: 2015 Certified**

**SHIVAJI UNIVERSITY KOLHAPUR**

**2019-2020**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**



**CERTIFICATE**

**This is to certify that, project work entitled**

**“CROP YIELD PREDICTION USING CNN”**

**Is a bonafied record of project work carried out in this college by**

**Miss. Bhakti Ajay Akole 16UCS11003XX**

**Miss. Prajakta Sudhir Bhosale 16UCS11009XX**

**Mr. Sandip Arun Torane 16UCS12059XX**

**Mr. Abhishek Rajendra Urane 16UCS12062XX**

**Mr. Avdhut Balaso Yadav 16UCS12063XX**

**In the partial fulfillment of award of degree Bachelor in Engineering in Computer Science & Engineering prescribed by Shivaji University, Kolhapur for the academic year 2019-2020.**

**Prof. K. S. KADAM**

**(PROJECT GUIDE)**

**Prof.(Dr.) D.V.KODAVADE Prof.(Dr.) P.V.KADOLE**

**(HOD CSE DEPT.) (PRINCIPAL)**

**EXAMINER: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**DECLARATION**

We hereby declare that, the project work report entitled “**CROP YIELD PREDICTION USING CNN”** which is being submitted to D.K.T.E. Society’s Textile and Engineering Institute Ichalkaranji, affiliated to Shivaji University, Kolhapur is in partial fulfillment of degree B.E.(CSE). It is a bonafide report of the work carried out by us. The material contained in this report has not been submitted to any university or institution for the award of any degree. Further, we declare that we have not violated any of the provisions under Copyright and Piracy / Cyber / IPR Act amended from time to time.

**NAME Roll no. Signature**

1. Miss. Bhakti Ajay Akole 16UCS11003XX
2. Miss. Prajakta Sudhir Bhosale 16UCS11009XX
3. Mr. Sandip Arun Torane 16UCS12059XX
4. Mr. Abhishek Rajendra Urane 16UCS12062XX
5. Mr. Avdhut Balaso Yadav 16UCS12063XX

**ACKNOWLEDGEMENT**

With great pleasure we wish to express our deep sense of gratitude to Mr. K. S. Kadam for his valuable guidance, support and encouragement in completion of this project report. Also, we would like to take opportunity to thank our head of department Dr. D. V. Kodavade for his co-operation in preparing this project report. We feel gratified to record our cordial thanks to other staff members of Computer Science and Engineering Department for their support, help and assistance which they extended as and when required.

Thank you,

Miss. Bhakti Ajay Akole 16UCS11003XX

Miss. Prajakta Sudhir Bhosale 16UCS11009XX

Mr. Sandip Arun Torane 16UCS12059XX

Mr. Abhishek Rajendra Urane 16UCS12062XX

Mr. Avdhut Balaso Yadav 16UCS12063XX

**ABSTRACT**

Deep learning: Deep learning is a branch of Machine Learning which is completely based on artificial neural networks, as neural network is going to mimic the human brain so deep learning is also a kind of mimic of human brain.

Farming is main occupation of India. Crop yield has direct impact on nation and international economies annually and the yield predicted plays significant part in the food management and agriculture sector. The task is to build prediction model for crop production . A prerequisite of intelligent system has brought artificial neural network to become a new technology which provides assorted solution for complex problems in agriculture researches. Performance of agriculture sector mainly hinges on natural forces such as spatio-temporal distribution of rainfall, temperature, climate etc, with the result any deviation of monsoon from the normal pattern brings about in numerous fluctuations in area and production. Crop yield has direct impact on nation and international economies annually and the yield predicted plays significant part in the food management and agriculture sector.

The task is to build prediction model for crop production. The basic principle of ANN architecture, Data Modeling for Prediction involves four stages namely historical data analysis (Descriptive), Data preprocessing, modeling of Data and Performance Estimation. First classify data based different attribute. Regression analysis using CNN, it observes the relation between a independent (predictor) and dependent (target) variables. Based on relation training model will predict crop yield production. application of ANN in predicting crop yield by using various crop performance features as input parameter.

**INDEX**

1. **Introduction 7**
   1. Problem definition 7
   2. Aim and objective of the project 7
   3. Scope and limitation of the project 7
   4. Timeline of the project 8
2. **Background study and literature overview 9**
   1. Literature overview 10
   2. Investigation of current project and related work 11
3. **Requirement analysis 12-14**
4. **System design 15**
   1. Architectural Design 16
   2. Algorithmic description of each modules 17
   3. Structural Design 18
      1. Class diagrams 18
      2. Object diagrams 18
   4. Data Design 19
      1. Data Flow Diagram 19
   5. Behavioral Designs 19
      1. Sequence Diagram 19
      2. Collaboration Diagram 20
      3. State-chart Diagram 20
      4. Activity Diagram 21
   6. Environmental Diagram 22
      1. Deployment Diagram 22
5. **Implementation 23**
   1. Detailed Description of Methods 24
   2. Implementation of Preprocessing 24
   3. Implementation of classification model 24
   4. Implementation of Training Module of yield production model 25
   5. Implementation of Front End Design 25
6. **Integration and Testing 26-27**
7. **Performance Analysis 28-29**
8. **Applications 30-31**
9. **Installation Guide and User Manual 32-35**
10. **Cost Estimation 36-38**
11. **Ethics 39-40**
12. **References 41-42**

**INTRODUCTION**

Farming is main occupation of India. Crop yield has direct impact on national and international economies annually and the yield predicted plays significant part in the food management and agriculture sector. India’s economy mainly depends on agriculture yield growth and their related agro industry. 61% percent of the total geographical area of India comes under agriculture. This model is taking a input as a dataset of the India. This particular dataset includes the data from all the states, districts of India and the crops in the respective regions. So, we are going to study of different states and districts of the India.

A prerequisite of intelligent system has brought artificial neural network to become a new technology which provides assorted solution for complex problems in agriculture researches. Performance of agriculture sector mainly hinges on natural forces such as spatio-temporal distribution of rainfall, temperature, climate etc, with the result any deviation of monsoon from the normal pattern brings about in numerous fluctuations in area and production. Crop yield has direct impact on nation and international economies annually and the yield predicted plays significant part in the food management and agriculture sector. The task is to build prediction model for crop production using data predictive.

1. **Problem definition:**

The problem statement of our project is to Design and develop yield prediction model using CNN.

1. **Aim and objectives of the Project**:

* **Aim**

Design and develop yield prediction model using CNN**.**

* **Objectives**

1. To build the classification model.
2. To build and develop yield prediction model.
3. To improve the accuracy based on MAPE(Mean absolute percentage error) metrics.
4. **Scope and limitation of the Project:**

* **Scope**

The scope of our project is to predict crop production with accuracy Along with .csv file that available for raining and testing. The intended software that shall be required by system Anaconda Navigator, Jupyter Notebook, Online Google collaborator tool freely provided Google for enhancing performance of the separations on the GPU's. This shall enable the system to Machine Learning models in a considerably less time by training testing of this large data set comprising of approximately 248000 records.

* **Limitations**

Categorical inputs such as state, districts, crop, season, should be realistic or valid and known value by model.

1. **Timeline of the Project:**

We have used classic life cycle paradigm also called “Water Fall Model”. For software engineering which is sequential approach to software development that begins at the system level and progress through analysis, design, coding, testing and maintenance. We had completed software requirement analysis by the mid of September 2019 which encompasses both system and software requirement gathering. By the end of November 2019, we had completed project planning and design. On the basis of design prepared in the previous stage by the end of February 2020 we completed coding stage.

After completion of coding stage the important part in the software development which is testing phase carried out in first week of March 2020.

****

**Literature overview**

Literature Overview

1. **Literature Overview:**

Deep learning:

Deep learning is a branch of Machine Learning which is completely based on artificial neural networks, as neural network is going to mimic the human brain so deep learning is also a kind of mimic of humanbrain.

The task is to build prediction model for crop production. The basic principle of ANN architecture, Data Modeling for Prediction involves four stages namely historical data analysis (Descriptive), Data preprocessing, modeling of Data and Performance Estimation. First classify data based different attribute. Regression analysis using CNN, it observes the relation between independent (predictor) and dependent (target) variables. Based on relation training model will predict crop yield production. application of ANN in predicting crop yield by using various crop performance features as input parameter.

1. **Investigation of current Project and Related work:**

**[1]P.Surya, Dr. I.LaurenceAroquiarajPeriyarUniversity, Salem, Tamilnadu, India[2018]** describes In thierproposed work, collected agriculture dataset will be used to get crop yield prediction model using various regression techniques. Regression analysis was tested for the effective prediction or forecast of the agriculture yield for various crops in Tamilnadu state particularly in North Western zone of Tamilnadu. North western zone of tamilnadu state data consist four districts. By the analysis depends on the results of predictor model, in the north western zone, under the area having more cultivated crops are Tapiaco, Sugar cane, Ragi, Maize, Groundnut. In this paper linear regression gives the better solution to predict crop production.

But above solution particularly focused on North Western zone of Tamilnadu, which has similar contents. It has small data, so may gives less accuracy.

**[2]B. Devika, B. Ananthi** they use classification technique to group similar content area wise and year wise. k-nn and linear regression algorithm are used and linear regression gives better accuracy.

But data may be non linear pattern with district wise so this model can gives different output.

[3]**Shakil Ahamed, Navidv Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, RashedurM Rahman** are estimate crop yield for major cereal crops in major districts of Bangladesh. we have considered the effects of environmental(weather) and area of production as factors towards crop production in Bangladesh. Taking these factors into consideration as datasets for various districts, they applied classification techniques to divide regions; and then apply suitable classification techniques to obtain crop yield predictions.

Here proposed methods was linear regression, k-nn, and neural network.

Neural network gives better prediction result.

**Requirement**

**Analysis**

1. System Requirement
2. **Hardware requirement:**

1. Personal computer with

* Chrome or firefox browser
* Internet connection

1. **Software requirements:**

* Microsoft Windows / Linux

1. **Application Development Tools:**

* Python3 with Keras & Flask Library
* Jupiter Notebook with Google Colab
* Tensor Flow Platform for running model

The Dataset was collected from website of agriculture government portal. It contains state wise, district wise, season wise, year wise & crop covered area and production of crop yield in India. In this dataset the attributes of area is measured in hectors and crop production is measured in tones per hector.

1. Functional Requirement
   1. The system shall eliminate redundant feature under the preprocessing phase.
   2. The system shall classify data as district wise, year wise, Area wise, season wise to analyze the data.
   3. The system shall train the neural network model using pre-processed dataset.
   4. The system shall generate prediction module.
2. Analysis
   1. **Preprocessing**

The system shall eliminate redundant data features. Removes null value. Removing of outlier will be done in preprocessing.

* 1. **Classification**

Preprocessed data passed to classification model. Here data will classified district wise, year wise, season wise, crop wise and stored into separate .csv files. In files categorical data is stored with unique values.

* 1. **Training of neural network**

The training dataset first convert into label encoding using label encoding method. We can't use labeling dataset to train so categorical data is embedded into specific category using word embedding. This data is trained to neural network until get better accuracy. Accuracy is measured based on MAPE and MSE metric.

* 1. **Prediction Model**

After training, design front end for input parameters and display output. This will done using flask and html.

Input parameters convert into label encoding. Give input to the model and predict the output as production in tons per hector.

**System Design**

1. **Architectural Design**

****

1. **Algorithmic description of each modules**
   1. **Preprocessing**
      1. Read the dataset
      2. Remove null values
      3. Remove redundant features
      4. Remove outlier
      5. Filter Dataset with specific range of input
      6. Encode categorical data into numerical using label encoding. Categorical data are district, season, crop.
      7. Split dataset into train and test dataset
      8. Save train.csv and test.csv
   2. **Classification**
      1. Read the preprocessed dataset without encoded
         1. Get unique values from state column
         2. Sort the array
         3. Save as state.csv file
         4. Get unique values from district column
         5. Sort the array
         6. Save as district.csv file
         7. Get unique values from season column
         8. Sort array and save file as season.csv
         9. Get unique values from crop column
         10. Sort array and save file as crop.csv
   3. **Training Module**
      1. Read train and test dataset
      2. Normalize numerical data such as year, area, production
      3. Store each columns data in separate array
      4. Create input layers for each input parameters
      5. Create embedding layer for each categorical parameters- districts, season, crop

Embedding( input\_dim = no\_of\_categories,

Output\_dim = embedded\_size,

Input\_length = 1)

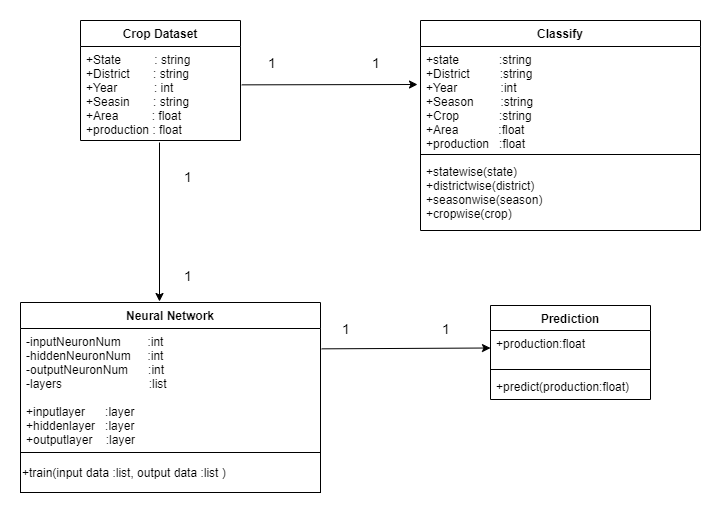
* + 1. Concat layers and attach to CNN
    2. Create neural network model:

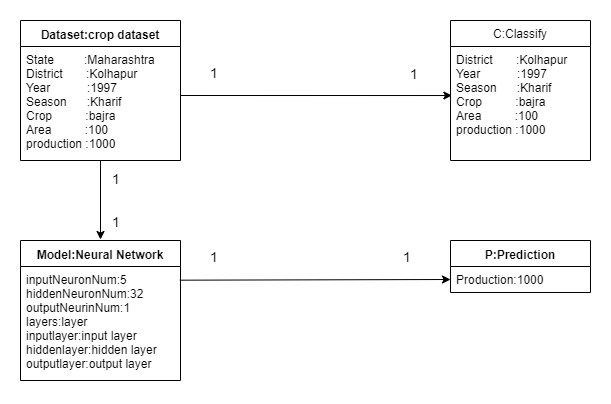
Model( inputs = [districs,year,season,crop,area],

Outputs = [production] )

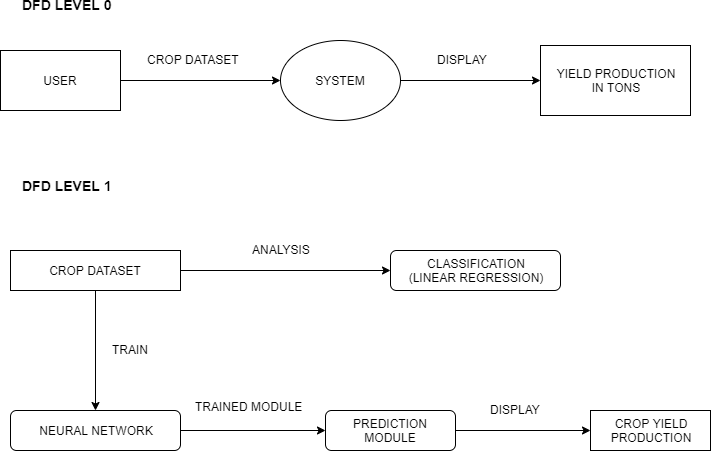
* + 1. Optimize & compile the model
    2. Train the neural network model until get better accuracy based on mape&mse
    3. Evaluate / Test the model
    4. After successful training of model save the trained model

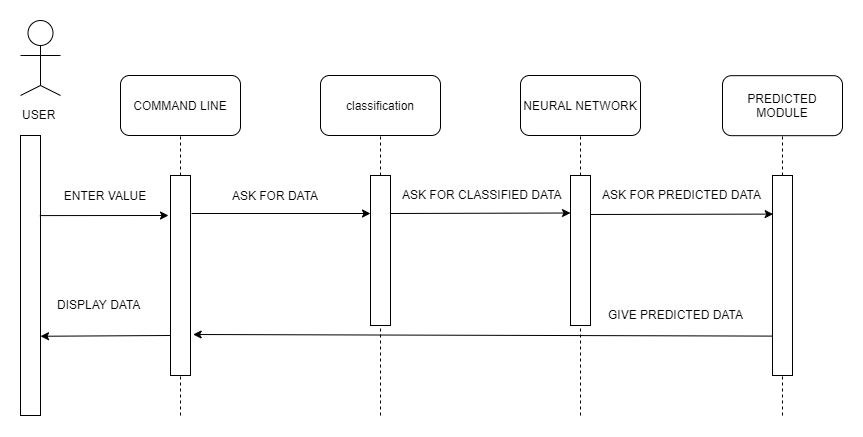
**C. Structural Design**

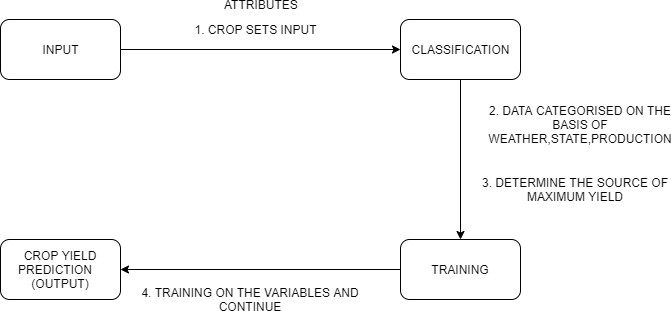
* 1. **Class Diagrams**
  2. **Object Diagrams**

****

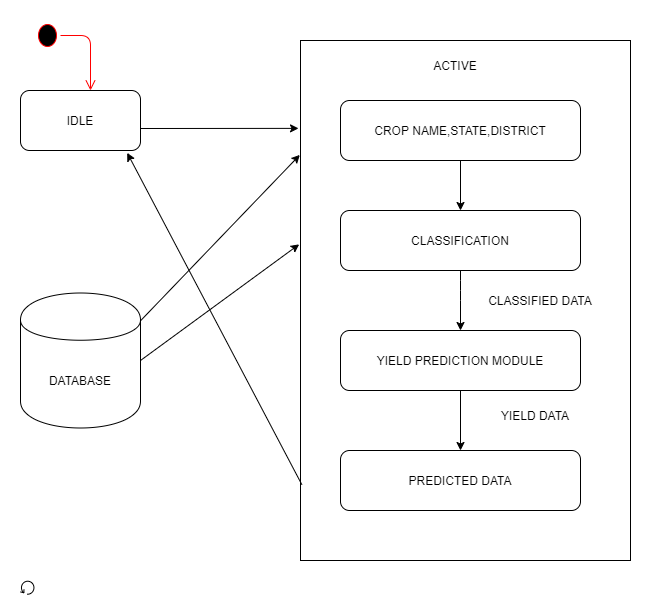
1. **Data Design**
2. **Data-flow Diagram**



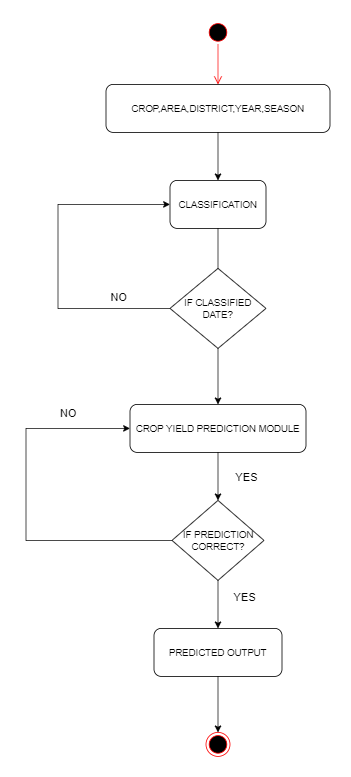
1. **Behavioral Design** 
   1. **Sequence Diagram**
   2. **Collaboration diagram**



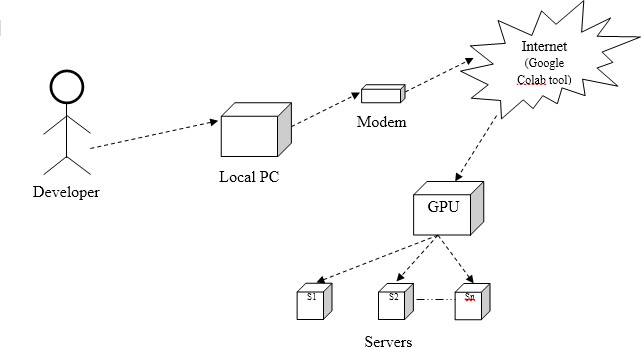
* 1. **State-chart diagram**

****

* 1. **Activity Diagram**

****

1. **Environment Design** 
   1. **Deployment Diagram**



**Implementation**

1. **Detailed Description of Methods**

Google colabinitilization :

Google’s free cloud service for AI Developers with Colab, you can develop deep learning applications on the GPU for free.

The most important feature that distinguish Colab from other free cloud services is; colab provides GPU and is totally free.

* 1. Create new Colab notebook and rename
  2. Setting free GPU

It is so simple to alter default hardware (CPU TO GPU). Just follow edit>Notebook settings>change Runtime type and select GPU as hardware accelator

* 1. Most Colab file with drive so we can access folder and files on Google Drive

1. **Implementation of Preprocessing**
   1. Read the dataset using pandas

Data = pandas.read\_csv(‘dataset.csv’)

* 1. Null values were removed. Eliminate redundant features, outlier were removed. Filters the specific range values
  2. Encode the dataset from categorical values to numeric values using label encoding :

le.fit\_transform(data[‘District\_Name’])

le.fit\_transform(data[‘Season’])

le.fit\_transform(data[‘Crop’])

* 1. Dataset is splited into train and test using skelarn library

Train,test = train\_test\_split(data,test\_size=0,2)

* 1. Save the train and test data frame as train.csv and test.csv file

1. **Implementation of classification model**
   1. Read the dataset using pandas

data = pandas.read\_csv(‘dataset.csv’)

* + 1. Get unique values from state column and store into list

State = data.state\_name.Unique()

* + 1. Sort the array of state
    2. Save file to .csv

df = pd.DataFrame(State)

df.to\_csv(‘/gdrive/state.csv’)

* 1. Repeat the step 2 for district, season, crop.

Save to drive : district.csv, season.csv, crop.csv

1. **Implementation of Training Module of yield production model**
   1. Import tensorflow, keras, pandas import layer
   2. Read the train & test dataset
   3. Normalization :-

Normalize the numerical value of dataset using standard deviation and mean.

crop\_year, Area and Production have the numerical values

cols\_to\_norm = cols\_to\_norm\_mean / standard\_deviation

* 1. Stores each column value in separate array
  2. Build neural networks
     1. Create input layers for each input parameters.

year\_input = Input(shape=(1,), dtype=’float32’)

area\_input = Input(shape=(1,), dtype=’float32’)

district\_input = Input(shape=(1,)dtype=’float32’)

season\_input = Input(shape=(1,)dtype=’float32’)

crop\_input = Input(shape=(1,)dtype=’float32’)

* + 1. First we use one hot encoding for categorical inputs but neural networks becomes bulky for thousands categories. It cant handle. So we use alternative embedding.

Embedding(input\_dim = no\_of\_categories, output\_dim = embedded\_size, input\_length = 1)

Where

Embedded\_size = min(no\_of\_categories/2, 50)

Do this for district\_input, season and crop with their input layers. This is most important step of categorical encoding in the project.

* + 1. Concatenate all layers & CNN and Create model
  1. Compile the model
  2. Train the model using input & output parameters dataset.

Train the model based on mape&mse metrics.

By the hyperparameters tunning improve the accuracy of model. Hyperparameters like layers, neurons, batch\_size, epochs.

* 1. Test the trained model using evaluate
  2. Save the trained model to drive as model.h5 with their weights saved model is useful for future prediction & use

1. **Implementation of Front End Design**
   1. Create input html form to get input
   2. Get the input encode the categorical data to label encoding using classified data.

Do this for district, season, crop using distric.csv, season.csv, crop.csv

* 1. Send the inputs to the neural network model and it returns the predicted output

**Integration**

**And**

**Testing**

Testing Performed :

**Test Plan :**

1. **Time**

We have to improve the efficiency of model. We tried embedding with neural network to train the model. The time taken by Method is :

|  |  |
| --- | --- |
| **Method** | **Time Required (avg)** |
| Neural Network  With Embedding | 0.7 seconds |

1. **Model Accuracy**

We train the neural network by train dataset and test using test dataset. Our main goal of project is to improve the accuracy based on mape (mean absolute percentage error).

We reduce the error 29.43% that mean we achieve accuracy of 71% approx.

1. **Output**
   1. Using one hot encoding before neural network on categorical data :

We test the training module, we get 51% accuracy

* 1. Using embedding in neural network model:

Encode the categorical data by embedding and hyperparameters tunning we get 62% accuracy

After improving batch size, epochs, layers, neurons we achieve accuracy of 71% approx.

* 1. Test input and front-end design:

We get input from html form and give to model for prediction. Here we check validation and selection of options

**Performance Analysis**

1. **Time**

|  |  |
| --- | --- |
| **Method** | **Time Required (avg)** |
| Neural Network with embedding | 0.7 seconds or 700ms |

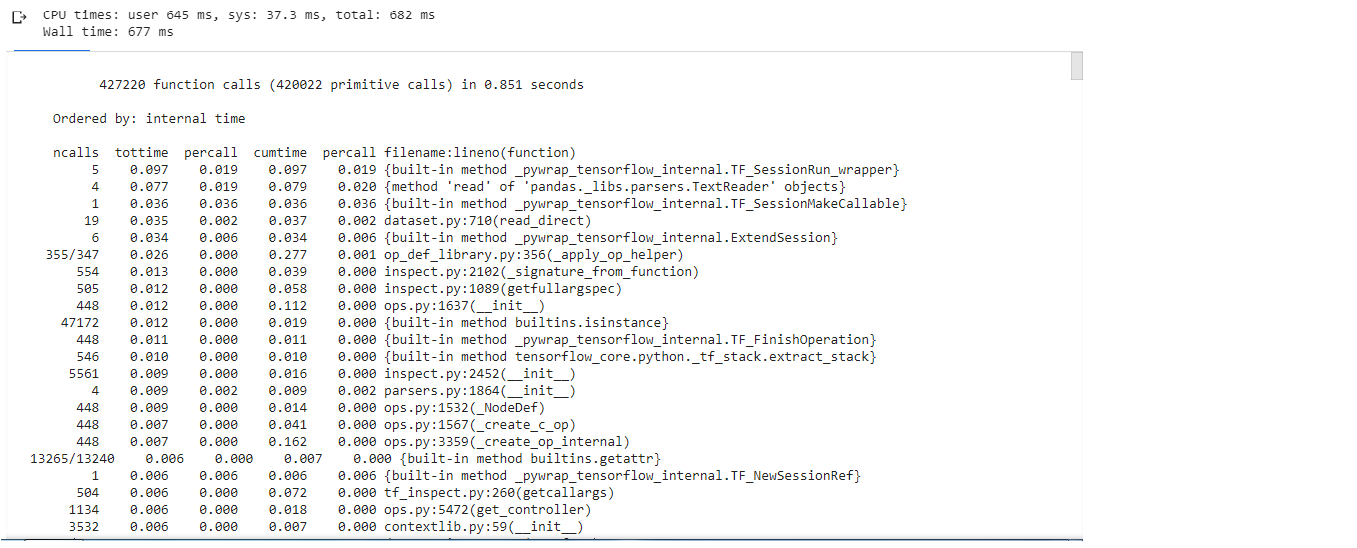
1. **Accuracy**

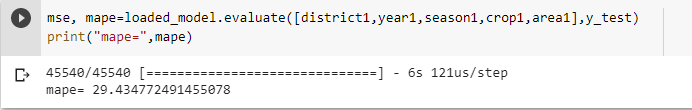
|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| One hot encoding with neural network | 51% |
| Embedding in neural network | 71% |

By using embedding for categorical data having thousands of categories and hyperparameter tunning in neural network, Model gives better result.

Conclusion :

As per above analysis, embedding method & hyperparameter tunning in neural network model gives better result.





**Applications**

**Applications :-**

1. For the farmer to know how much crop production he get before harvesting.
2. Government organization for creation well decision and for make better rules which helps to increased production, economy.

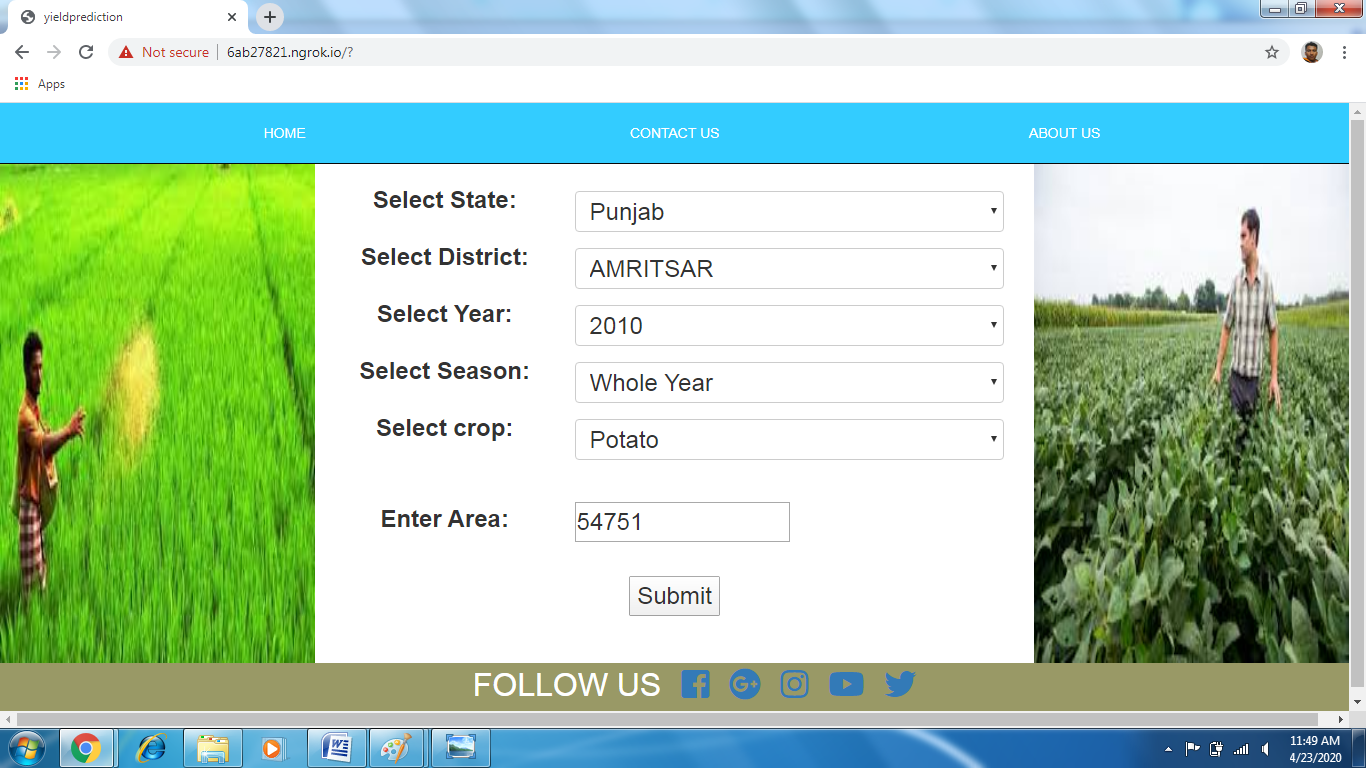
**Installation Guide And**

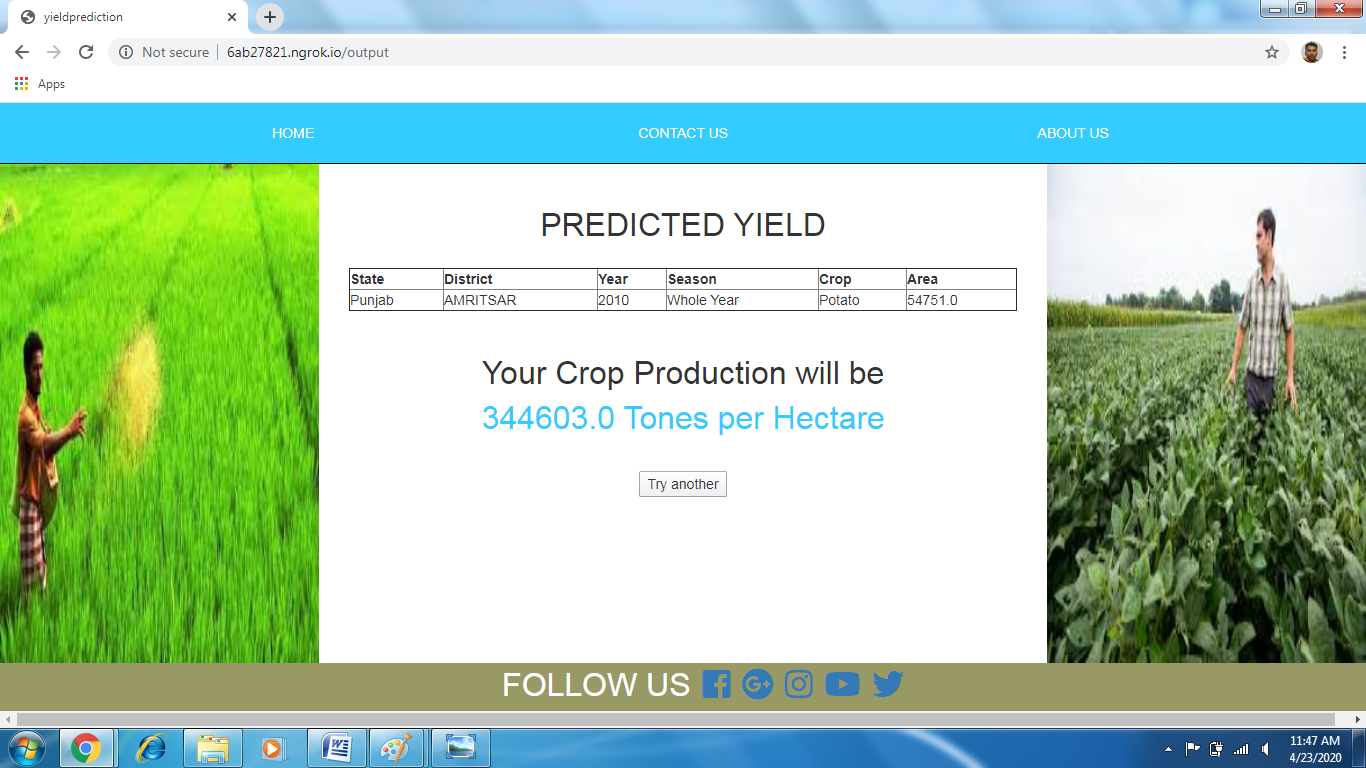
**User Manual**

**Installation guide and user Manual :-**

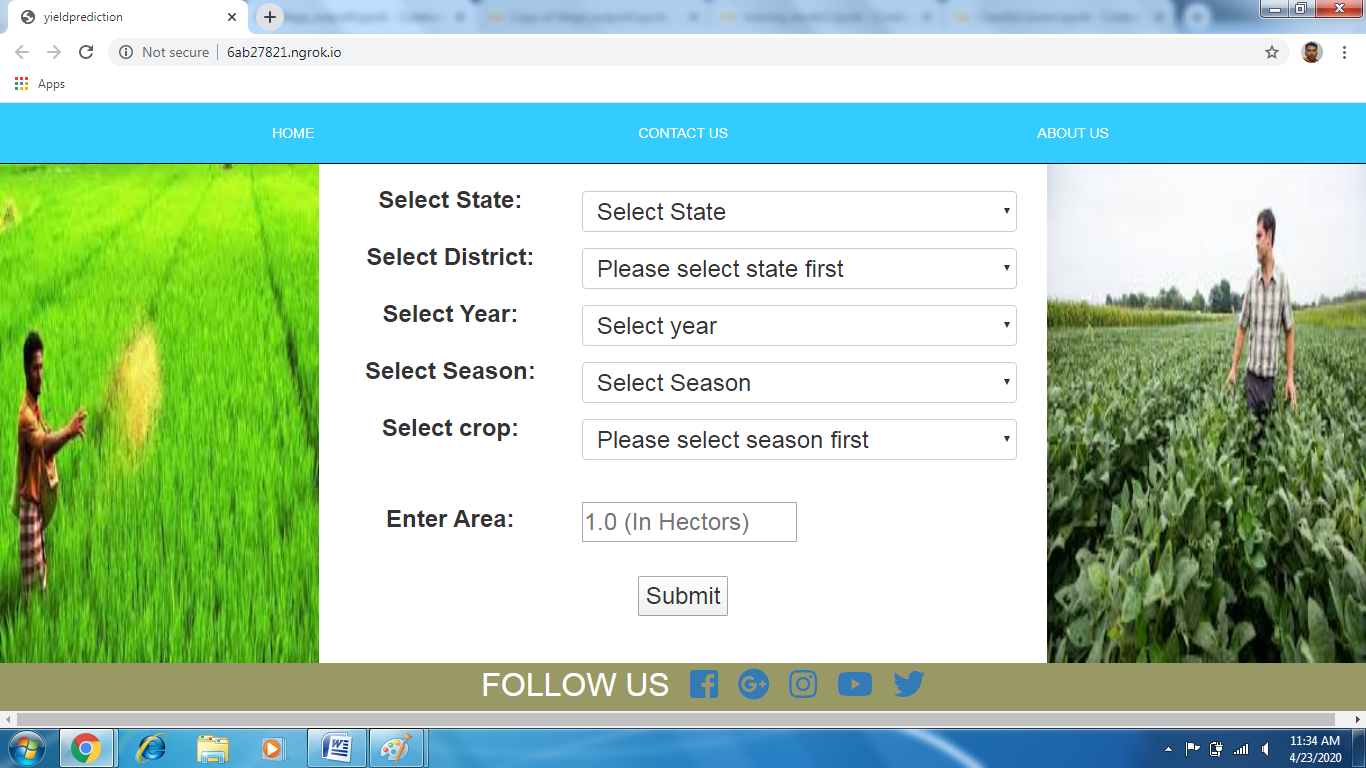
**\*Implementation of front end design:**

1. create input html form to get input.
2. Get the input encode the categorical data to label encoding using classified data. Do this for district, season, crop using district.csv, season.csv, crop.csv.
3. send the inputs to neural network model and returns the predicted output.

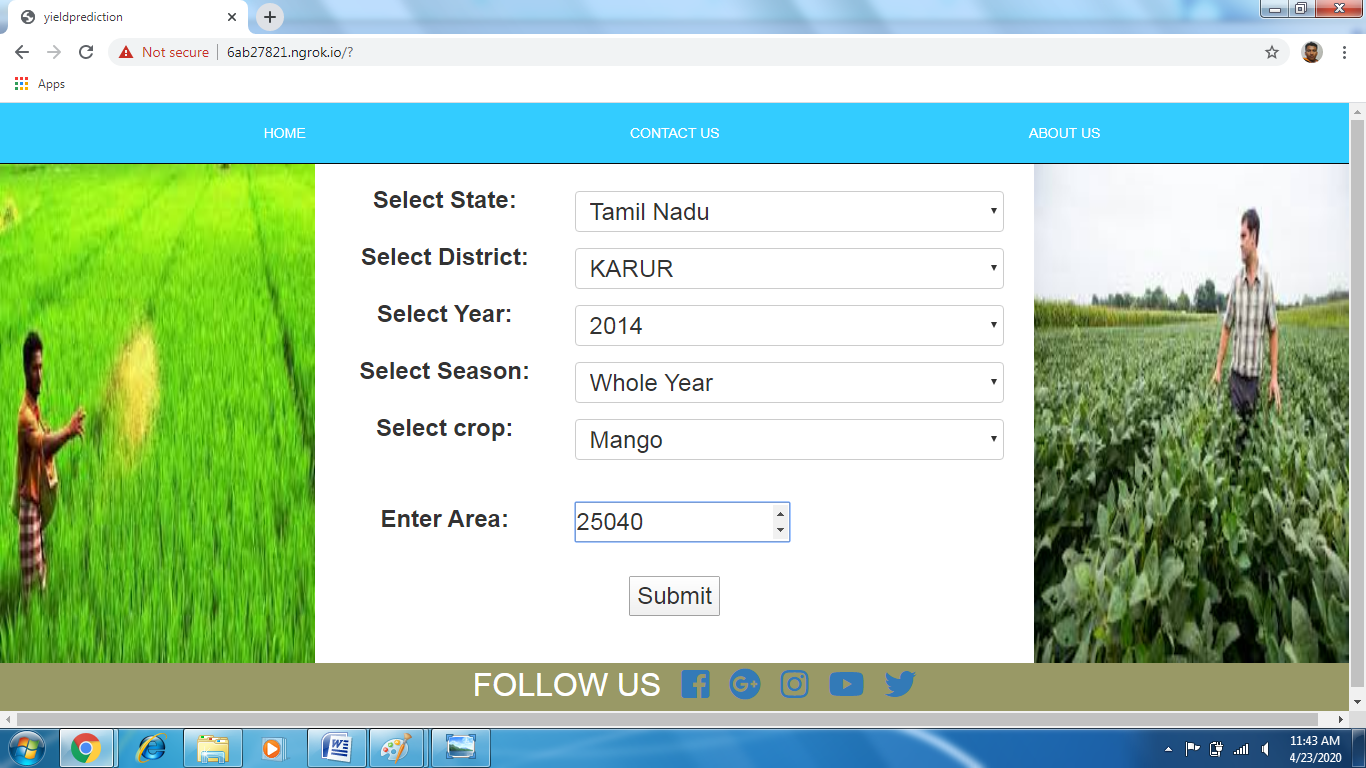




1. No installation required at client side
2. Deployment of system on server
3. Resources needed for a successful installation
   1. High speed connectivity
   2. Web space (future scope)
4. After installing
   1. Open the website

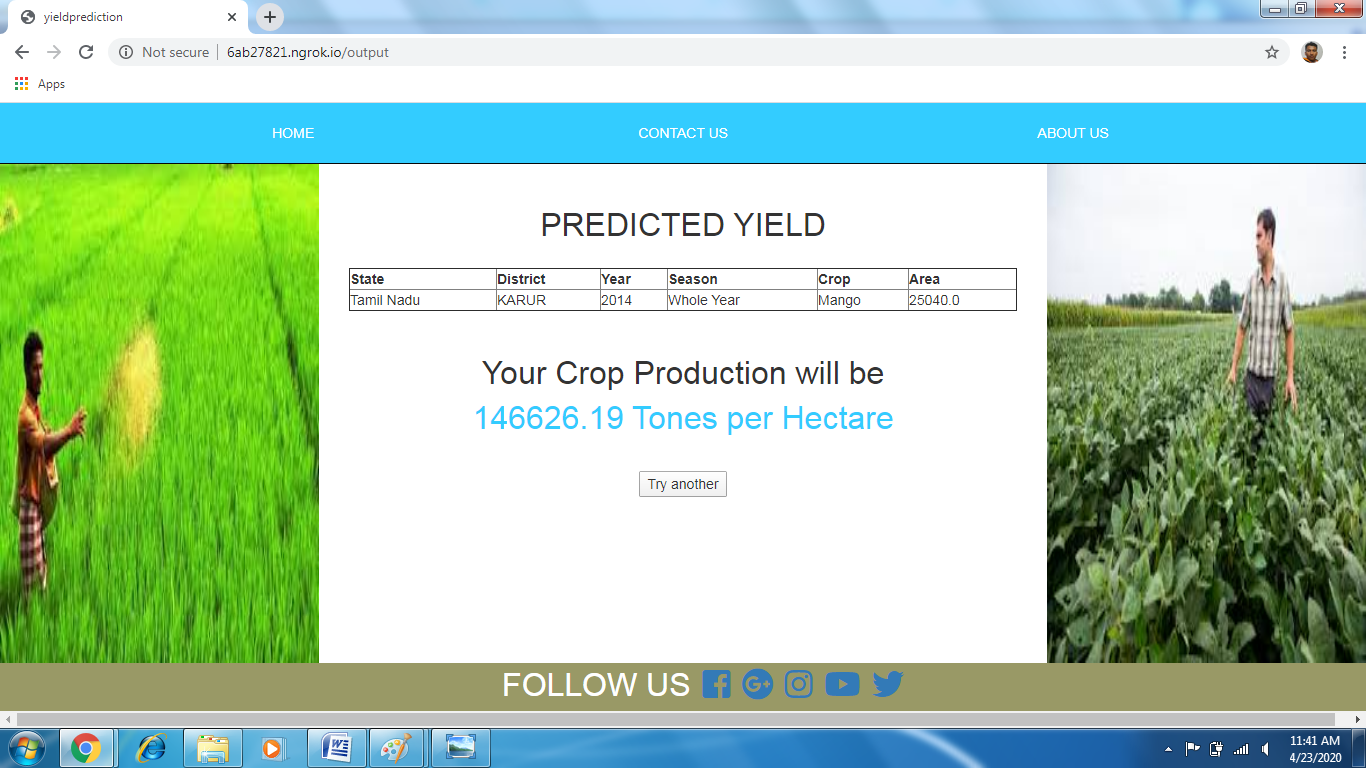


* 1. Provide input to the system such as select state, district, season, crop name and year, area



* 1. Displayed output

c. Displayed output



**Cost Estimation**

**Project Cost :-**

**\*Hardware Cost:**

|  |  |
| --- | --- |
| **Hardware** | **Cost** |
| Computer System | **Rs. 40,000/-** |
| Total | **Rs. 40,000/-** |

In this project the Cost Estimation based on COCOMO (Constructive Cost Model) the formula for the this Model is follows

Effort = Constant × (Size) scale factor× Effort Multiplier

Effort in terms of person-months

Constant: 2.45 in 1998 based on Organic Mode –

Size: Estimated Size in KLOC –

Scale Factor: combined process factors

Effort Multiplier (EM): combined effort factors

The basic COCOMO equations take the form

Effort Applied (E) = ab(KLOC)b b [ man-months ]

Development Time (D) = cb(Effort Applied)d b [months]

People required (P) = Effort Applied / Development Time [count]

Where, KLOC is the estimated number of delivered lines (expressed in thousands) of

code for project. The coefficients ab, bb, cb and db are given in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Software Project** | **ab** | **bb** | **cb** | **db** |
| **Organic** | 2.4 | 1.05 | 2.5 | 0.38 |
| **Semidetached** | 3.0 | 1.12 | 2.5 | 0.35 |
| **Embedded** | 3.6 | 1.20 | 2.5 | 0.32 |

**Semidetached Model** :

EFFORT = 3\*(5)^1.12 = 18.19

Development Time = 2.5\*18.19^0.35 = 6.9

People Required = 18.19/6.9 =2.78 ~ 5 people

**Ethics**

**Declaration of Ethics:**

As A Computer Science & Engineering Student, I believe it is Unethical To,

1. Surf the internet for personal interest and non-class related purposes during classes.

2. Make a copy of software for personal or commercial use

3. Make a copy of software for a friend

4. Loan CDs of software to friends

5. Download pirated software from the internet

6. Share a pirated copy of software

7.Install a pirated copy of software

**REFERENCES**

1. P.Surya, 2 Dr. I.Laurence Aroquiaraj,” Crop yield prediction in agriculture using data mining predictive analytic techniques", IJRAR, Volume 5, Issue 4,page no 783, December 2018.
2. B. Devika, B. Ananthi," analysis of crop yield prediction using data mining technique to predict annual yield of major crops", International Research Journal of Engineering and Technology (IRJET), Volume 5, December 2018.
3. Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman," Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh ", IEEE SNPD,June 2015
4. Dataset-data.gov.in
5. Keras.io