Crop Data Analysis of Indian Region using Big Data Techinques for aiding Indian Farmers with Emergencies Alert

Submitted by:

Ashutosh Anand
Roll No: 171500064
Vineet Rathore
Roll No: 171500382
Aman Sharma
Roll No: 171500033

Submitted to:

Mr. Ashutosh Shankdhar
Asst. Professor

Department of Computer Engineering & Applications

Institute of Engineering & Technology

GLA University
Mathura- 281406, India
2020

Department of Computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH2, Mathura-Delhi Road, P.O. – Chaumuhan,
Mathura – 281406

Declaration

*We hereby declare that the work which is being presented in the Big Data Project: "****Crop Data Analysis of Indian Region with Emergencies Alert****", in fulfilment of the requirements for Big Project, is an authentic record of our own work carried under the supervision of* ***Mr. Ashutosh Shankdhar, Asst. Professor, GLA University, Mathura.***

Ashutosh Anand

Vineet Rathore

Aman Sharma

Department of Computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH2, Mathura-Delhi Road, P.O. – Chaumuhan,
Mathura – 281406

CERTIFICATE

*This is to certify that the project entitled “***Crop Data Analysis of Indian Region with Emergencies Alert***” carried out in Big Data Project is a bonafide work done by Ashutosh Anand(171500064), Vineet Rathore(171500382) , and Aman Sharma(171500033) is submitted in fulfilment of the requirements for the award of the degree Bachelor of Technology (Computer Science & Engineering).*

Signature of Supervisor:
Name of Supervisor: Mr. Ashutosh Shankdhar
Date: 12-05-2020

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Big Data Project undertaken during B. Tech. Third Year. This project in itself is an acknowledgement to the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received.

Our heartiest thanks to Dr. (Prof). Anand Singh Jalal, Head of Dept., Department of CEA for providing us with an encouraging platform to develop this project, which thus helped us in shaping our abilities towards a constructive goal.

We owe special debt of gratitude to Mr. Ashutosh Shankdhar, Asst. Professor, Department of CEA, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. He has showered us with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Table of Contents:

Declaration..ii

CERTIFICATE..iii

ACKNOWLEDGEMENT..iv

Chapter 1: Introduction..1

1.1 Overview..1

1.2 Motivation and Objectives..2

Chapter 2: Software Requirements Analysis..3

2.1 Software Requirements..3

2.1.1 Node.js..3

2.1.2 MongoDB..4

2.1.3 Web Browser..4

2.1.4 Visual Studio..5

2.2 Introduction to Libraries..5

2.2.1 Pandas..5

2.2.2 Numpy..6

2.2.3 Seaborn..7

Chapter 3: Data Analysis..8

3.1 Dataset..8

3.2 Data Cleaning..8

3.3 Data Visualization..8

Chapter 4: Methodology..10

Chapter 5: User Interface..12

Chapter 6: Conclusion..17

6.1 Future Prospects..17

6.2 Conclusion..17

References..18

Chapter 1: Introduction

1.1 Overview

Big Data Analytics is a process to uncover some hidden patterns and information by applying big data techniques, by using this technique we can analyse all the data and can get the significant value from it. We know that in current scenario data can be in any form like in the form of some written text, in the form of audio files or in the form of images and it could be of any form so by applying modern big data techniques like hadoop, mapReduce and No SQL database we can store and process these data very efficiently. Nowadays big data is very important, it is very helpful for the business organizations because by using this technique they can understand the customer needs better, they can understand the new market strategies, by using this technique they can take help from social media to understand the customer behaviour better. Now coming to the agriculture benefit of big data, we all know that Agriculture or farming is the largest contributor to the GDP of our country. Approximately 70 percent of rural household's primary source of income is agriculture. But still farmers don't get results worth their efforts. As a matter of fact, the condition of farmers in our country is horrible. Even though the agriculture provides whopping 50 percent of the total employment in India, it only constitutes about 15.4 percent of the Gross Domestic Product (GDP). Due to variations in climatic conditions, improper crop selection or irrigation often leads to fewer yields than expected. On the other note, in the field of Information Technology, Big Data and Machine Learning has come forth as a blazing topic. So by using Big Data approach in analysing the crop data over various factors such as soil type, temperature, water level, humidity, soil pH, fertilizers, we will devise an algorithm that will predict which crop should be sown in which season that both the yield and the profit can be maximized. For fulfilling the above task, we have used Hadoop framework, Machine Learning and NoSQL Database. Furthermore, even if the above factors are kept in mind, we can not deny or predict the forces of nature. One example of such are natural disasters. It is a known fact that most regions of India are prone to natural calamities. Even though farmers do everything right but due to these calamities not only they lose their properties and yield but also their lives.

So as an attempt to minimize these circumstances, this paper proposes a very effective and economical way of alerting to all kinds of security emergencies and the good aspect of this is that it does not require any additional efforts. Including this feature with the help of APIs from Indian Meteorological Department has devised another use case, which is, that, it can also provide weather data. Hence, the farmer can also choose an optimal time for sowing a particular crop based upon its water requirements and rain. This security feature is incorporated within the crop analysis and selection project which makes this a complete package for helping our farmers. Thus, it will maximize the yield and profit on one hand and try reducing the loss of lives and properties.

1.2 Motivation and Objectives

We know that many farmers face various problems to select crop for specific season, what cost it will take in whole season and how much will they earn from it. And many a times they also have to face natural disasters which makes them even more helpless. Some brokers in the middle take profit from it and make fool of poor farmers. Often farmers commit suicides at the failure of their crops. So, with the help of our project farmers will be able to analyse best season for cultivating specific types of crops according to the soil type, water content of the soil, pH value of the soil, temperature, humidity, type of fertilizers used and also how to tackle with the natural calamities.

Chapter 2: Software Requirements Analysis

2.1 Software Requirements

2.1.1 Node.js

Node.js is an open-source, cross-platform, JavaScript runtime environment that executes JavaScript code outside of a web browser. Node.js lets developers use JavaScript to write command line tools and for server-side scripting—running scripts server-side to produce dynamic web page content before the page is sent to the user's web browser. Consequently, Node.js represents a "JavaScript everywhere" paradigm,[6] unifying web-application development around a single programming language, rather than different languages for server- and client-side scripts.

Node.js has an event-driven architecture capable of asynchronous I/O. These design choices aim to optimize throughput and scalability in web applications with many input/output operations, as well as for real-time Web applications (e.g., real-time communication programs and browser games).

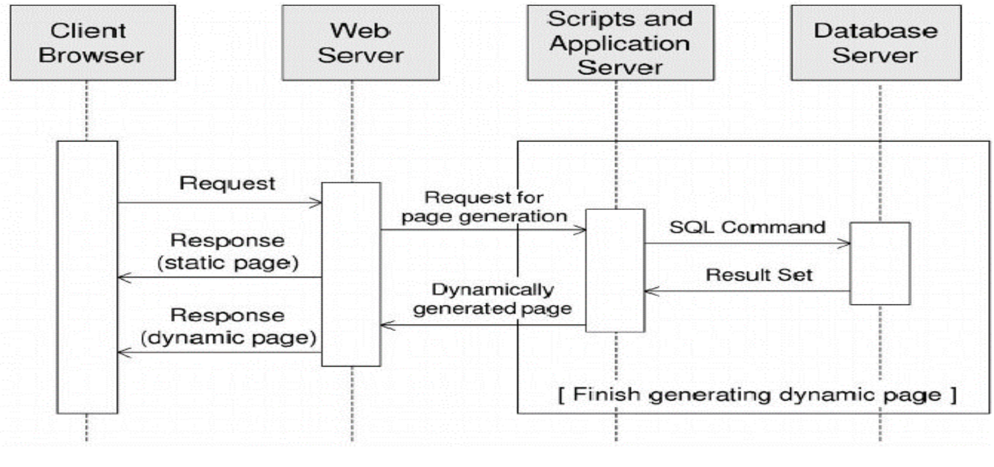

Client Browser
Web Server
Scripts and Application Server
Database Server
Request
Request for page generation
SQL Command
Response (static page)
Result Set
Dynamically generated page
Response (dynamic page)
[Finish generating dynamic page]

Figure 2.1: Traditional Server

2.1.2 MongoDB

MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schema. MongoDB is developed by MongoDB Inc. and licensed under the Server Side Public License (SSPL).

MongoDB supports field, range query, and regular expression searches. Queries can return specific fields of documents and also include user-defined JavaScript functions. Queries can also be configured to return a random sample of result of given size.

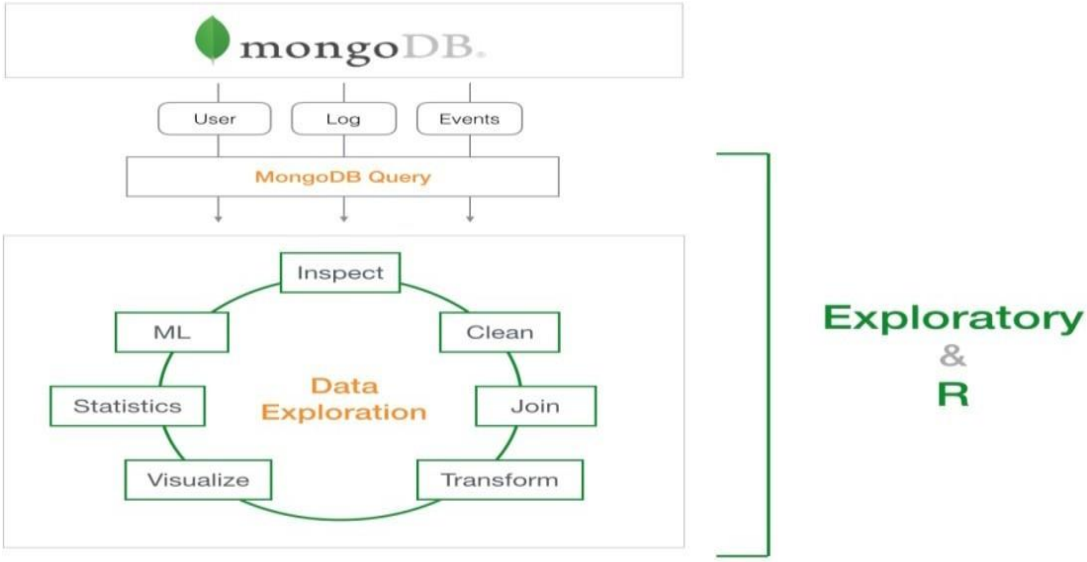


Figure 2.2 MongoDB Block Diagram

2.1.3 Web Browser

A web browser, or simply "browser," is an application used to access and view websites. Common web browsers include Microsoft Internet Explorer, Google Chrome, Mozilla Firefox, and Apple Safari. The primary function of a web browser is to render HTML, the code used to design or "mark up" web pages. Each time a browser loads a web page, it processes the HTML, which may include text, links, and references to images and other items, such as cascading style sheets and JavaScript functions. The browser processes these items, then, renders them in the browser window.

2.1.4 Visual Studio

Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, MacOS and Linux. It comes with built-in support for JavaScript, Typescript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity).Visual Studio Code is an editor first and foremost, and includes the features you need for highly productive source code editing. This topic takes you through the basics of the editor and helps you get moving with your code.

2.2 Introduction to Libraries

2.2.1 Pandas

In computer programming, pandas is a library written in python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three clause BSD license. The name is derived from the term "panel data", and term for data sets that include observations over multiple time periods for the same individuals.

Key Features

1. Intelligent data alignment and integrated handling of missing data: gain automatic label-based alignment in computations and easily manipulate messy data into an orderly form;

2. Flexible reshaping and pivoting of data sets;

3. Intelligent label-based slicing, fancy indexing, and sub-setting of large data sets;

4. Columns can be inserted and deleted from data structures for size mutability;

5. Aggregating or transforming data with a powerful group by engine allowing split-apply combine operations on data sets;

6. High performance merging and joining of data sets;

7. Hierarchical axis indexing provides an intuitive way of working with high-dimensional data in a lower-dimensional data structure;

8. Time series-functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging. Even create domain-specific time offsets and join time series without losing data;

9. Highly optimized for performance, with critical code paths written in C.

10. Python with pandas is in use in a wide variety of academic and commercial domains, including Finance, Neuroscience, Economics, Statistics, Advertising, Web Analytic, and more.

2.2.2 Numpy

NumPy is a library for the Python programming language, adding support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Num-array into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

Key Features

1. A powerful N-dimensional array object

2. Sophisticated (broadcasting) functions

3. Tools for integrating C/C++ and Fortran code

4. Useful linear algebra, Fourier transform, and random number capabilities

2.2.3 Seaborn

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures. Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Key Features

1. Seaborn is a statistical plotting library
2. It has beautiful default styles
3. It also is designed to work very well with Pandas dataframe objects.
4. High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations
5. Concise control over matplotlib figure styling with several built-in themes
6. Tools for choosing color palettes that faithfully reveal patterns in your data

Chapter 3: Data Analysis

3.1 Dataset

We have taken the dataset from the government website data.gov.in/crop. Our dataset consist of the data of all the states which consist of their districts and the production of crops with their respective seasons. The dataset further consist of area of the districts. The features of dataset are State_Name, District_Name, Season, Crop, Area, Production, P/A(production per area) and Price.

3.2 Data Cleaning

One of the most important aspect of analysing a data requires cleaning of dataset. We removed “crop year” column because of some redundancy. We used “pandas” which is a famous library of python that takes data (like a CSV or TSV file, or a SQL database) and creates a python object with rows and columns called data frame that looks very similar to table in a statistical software. It is mainly used for data manipulation and data analysis. We used group by function of pandas to group our states with their respective districts and seasons. We used aggregate function to retrieve the average of the area and the production of the crop in a particular season of a district. Further we added a new column named as “P/A” which shows about production per area of a particular of a respective season with their respective districts.

3.3 Data Visualization

Data visualization is very important to represent the features of data in graphical form to understand complicated relationship in data. Standardizing the data is essential need before visualization. Data Standardization is a data processing workflow that converts the structure of disparate datasets into a Common Data Format. As part of the Data Preparation field, Data Standardization deals with the transformation of datasets after the data is pulled from source systems and before it's loaded into target systems. Because of the transformation rules engine in Data Exchange operations We have used Min Max Scaler algorithm to standardize our dataset. feature, MinMaxScaler subtracts the minimum value in the feature and then divides b the range is the difference between the original maximum and original minimum. MinMaxScaler preserves the shape of the original distribution. It doesn’t meaningfully change the information embedded in the original data. Note that MinMaxScaler do outliers. The default range for the feature returned by MinMaxScaler is 0 to 1. We have made a bar plot with the help of seaborn library of python of State_Name Vs Production.

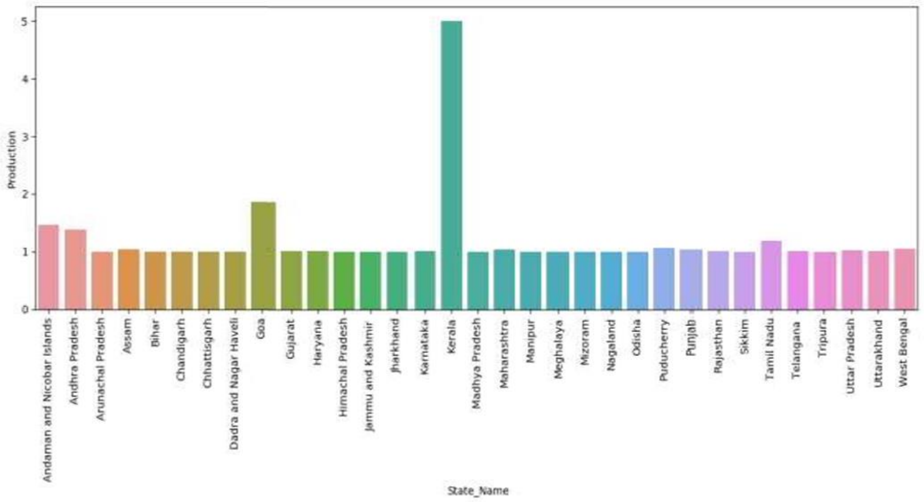


Figure 3.1 State_Name Vs Production

We have made a bar plot with the help of seaborn library of python of State_Name Vs P/A where, P/A represents production of crop per unit area.

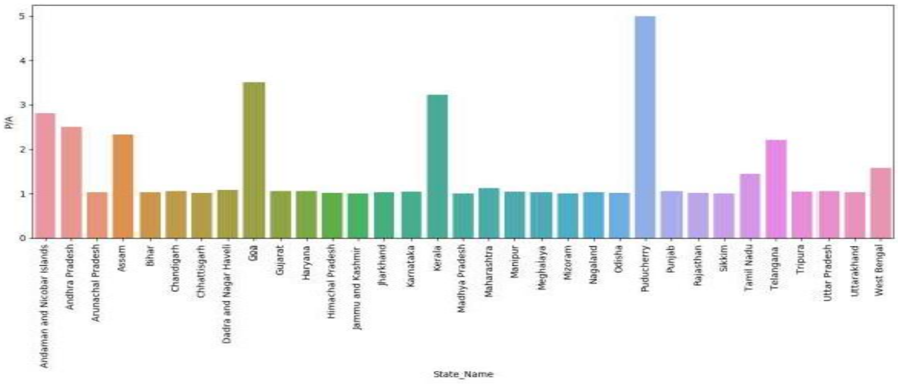


Figure 3.2: Graph of state name v/s “P/A”

Here, P/A represents production of crop per unit area

Chapter 4: Methodology

The trend of social media is increasing rapidly. Almost all the peoples are active on social media platforms like Facebook, Instagram, Twitter etc. On these platforms we post text messages, images, audio files, video files. More than millions of people use these platforms it means more than billon of messages and to store these messages there is a database which is present. These messages can be stored in RDBMS but RDBMS is not efficient to store and process the data which is more than 1 million and this type of data is an unstructured data and RDBMS is not able to store unstructured data so here NOSQL databases comes into picture. NOSQL databases are efficient to store and process the unstructured, semi structured as well as structured data also.

In this project first of all we imported CSV file into mongoDB by using mongo-import command and then it got converted into JSON format (JSON is a format in which all the data in MongoDB is stored).

Import command in MongoDB

```
 mongoimport –d databasename –c collection_name --type csv --file file_name.csv –
headerline
```

Export Command in MongoDB

```
mongoexport –db database_name –collrction collection_name –type = csv –
fields_id,State_name – out File_Name.csv
```

Here, we are connecting our Mongoose database with node.js with the help of mongoose.js library and modelling (ODM) library that provides a meticulous training environment for the data, imposing schemas as per the requirement while still preserving the flexibility that makes MongoDB powerful and robust. After that we embed mongoDB queries in nodeJS by which it will process the input data given by the user and will print the desired output from database by taking the given input into account.

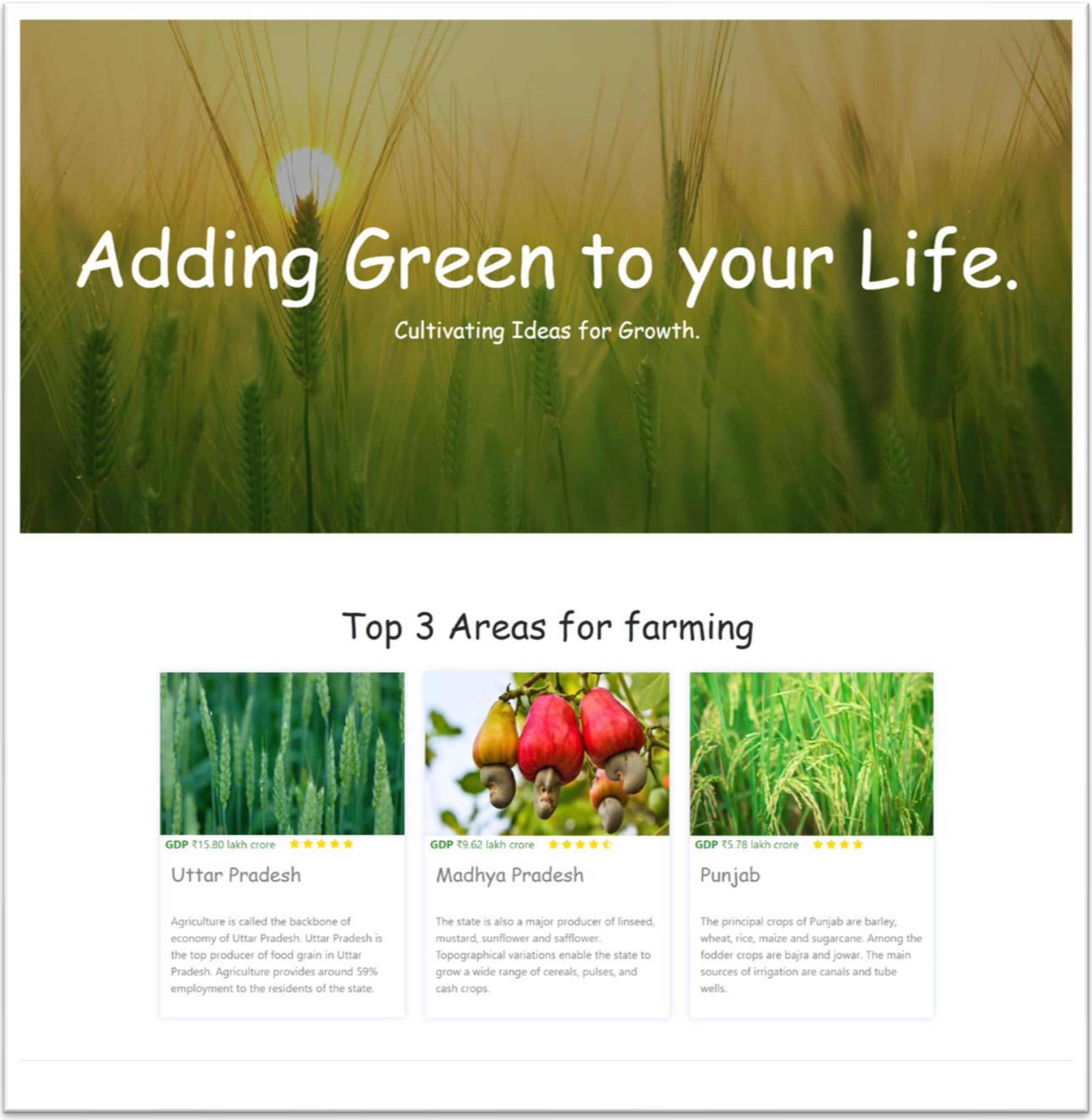
Additionally, to implement the other aspect of the project we took help of the information provided by the Government with respect to the weather, climate, etc. on their website/portals and mainly their APIs. A department of our Government only works on weather forecast, predicting calamities and so on.

The Indian Meteorological Department is an organization of the Ministry of Earth Sciences which comes under the Government of India. It is the chief corporation in charge of the meteorological observations, weather forecasting, seismology and other similar activities. By using the data provided by these agencies, we have incorporated this security alert feature. Whenever a farmer selects his region for crop selection, he will be notified about the conditions of his region within the same webpage. It would have been useful if was present during Karnataka Floods, Kerala Floods, Cyclone Fani, Bihar Floods and many more calamities.

Currently, it the midst of corona pandemic, it alerts the farmers how safe their zone is. Moreover, they are informed about various safety and measure and precautions. Also, whenever such a condition arises, Government announces various programs to help the people and they can be notified about it all in the same place. Similarly, it can work as same for any other calamity. As of now we are able to predict tsunamis, cyclones, etc. ahead of time and if any farmer is situated in its area of impact then they can be notified about the same. Another utilization of this project can be as weather forecast. This can not only predict calamities but also rain and sunshine. Farmers can know when its likely to rain in their region and when it’s going to be sunny and thus can choose the time to sow seeds, plough fields, etc. accordingly.

Chapter 5: User Interface

The Main Page of user interface



Adding Green to your Life.

Cultivating Ideas for Growth.

Top 3 Areas for farming

GDP ₹15.80 lakh crore ★★★★★

Uttar Pradesh

Agriculture is called the backbone of economy of Uttar Pradesh. Uttar Pradesh is the top producer of food grain in Uttar Pradesh. Agriculture provides around 59% employment to the residents of the state.

GDP ₹9.62 lakh crore ★★★★☆

Madhya Pradesh

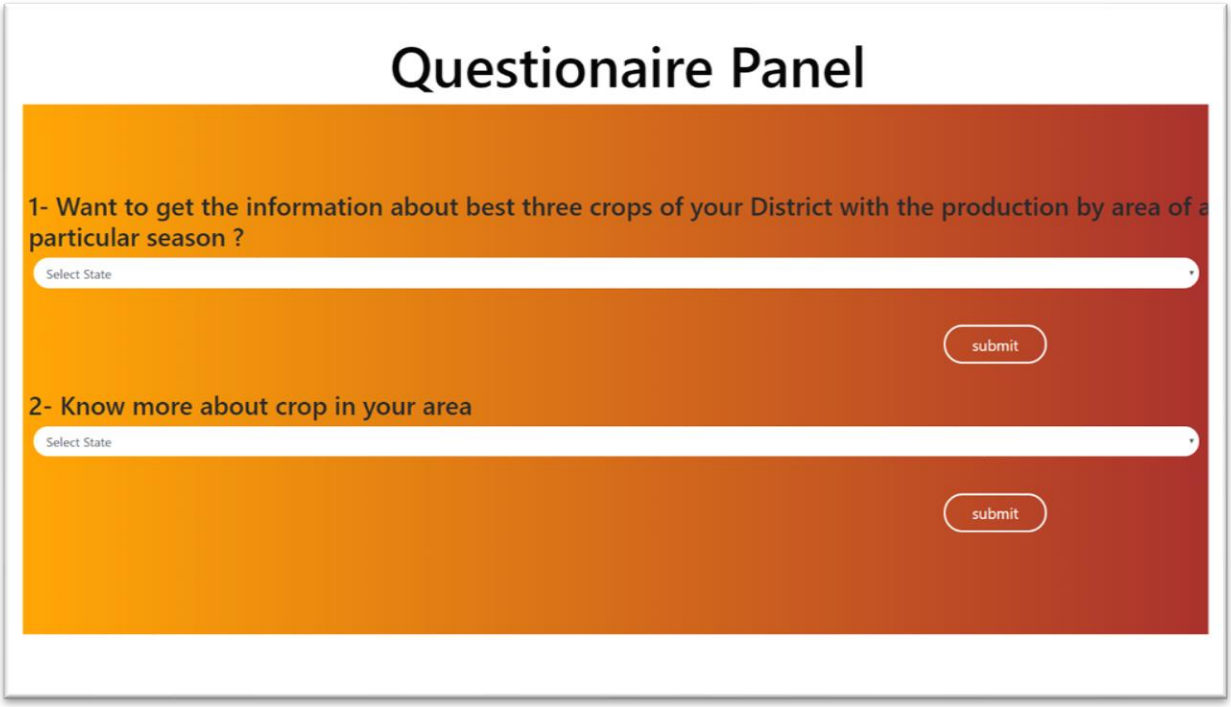
The state is also a major producer of linseed, mustard, sunflower and safflower. Topographical variations enable the state to grow a wide range of cereals, pulses, and cash crops.

GDP ₹5.78 lakh crore ★★★★

Punjab

The principal crops of Punjab are barley, wheat, rice, maize and sugarcane. Among the fodder crops are bajra and jowar. The main sources of irrigation are canals and tube wells.

Questionnaire Panel



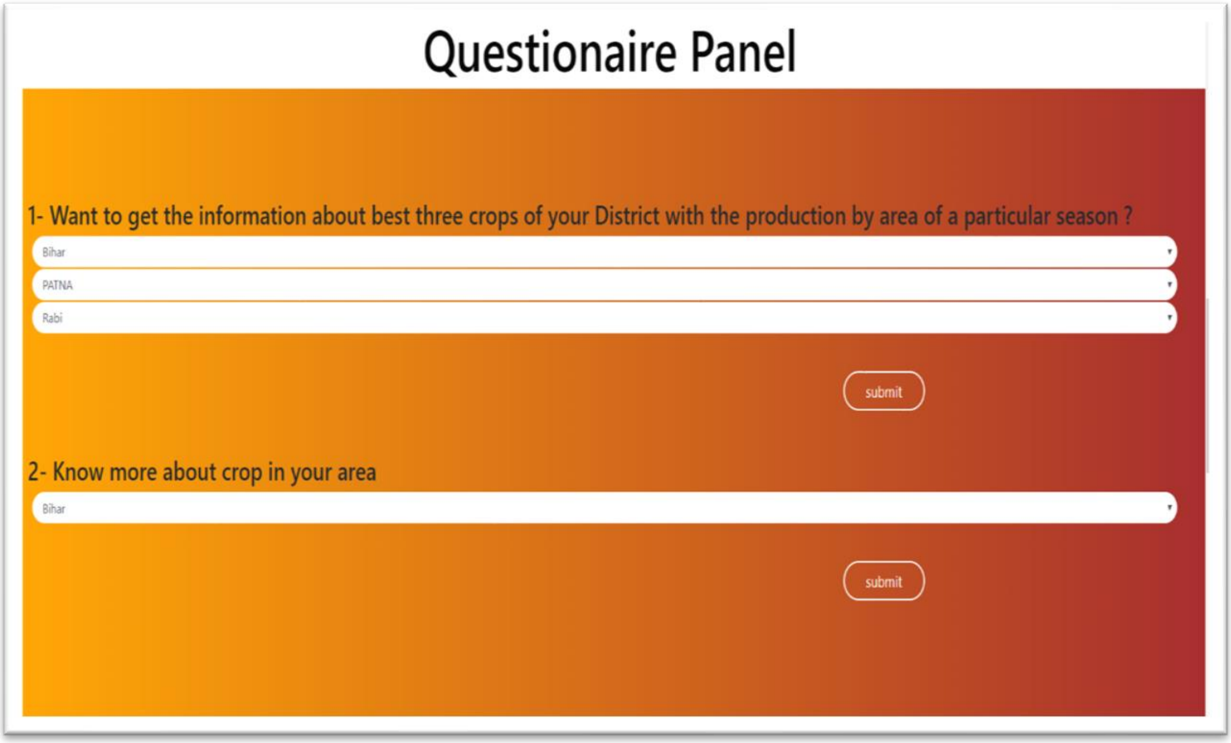
Questionaire Panel

1- Want to get the information about best three crops of your District with the production by area of a particular season ?

Select State

2- Know more about crop in your area

Select State



Questionaire Panel

1- Want to get the information about best three crops of your District with the production by area of a particular season ?

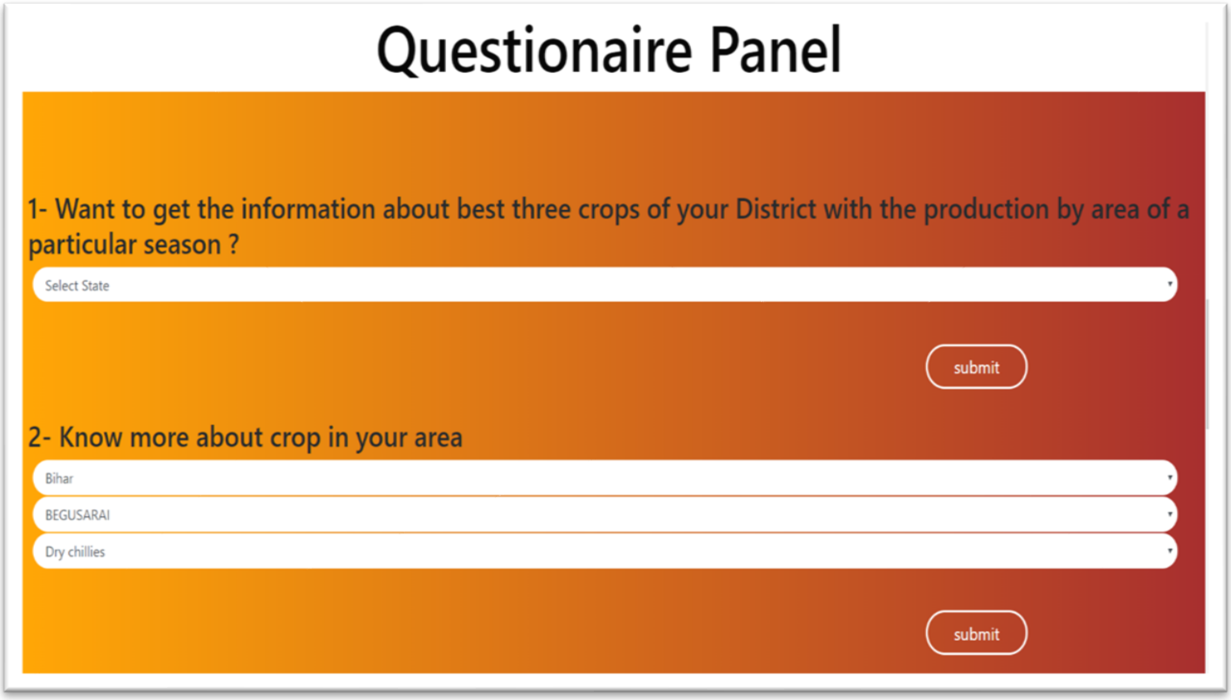
Bihar

PATNA

Rabi

2- Know more about crop in your area

Bihar

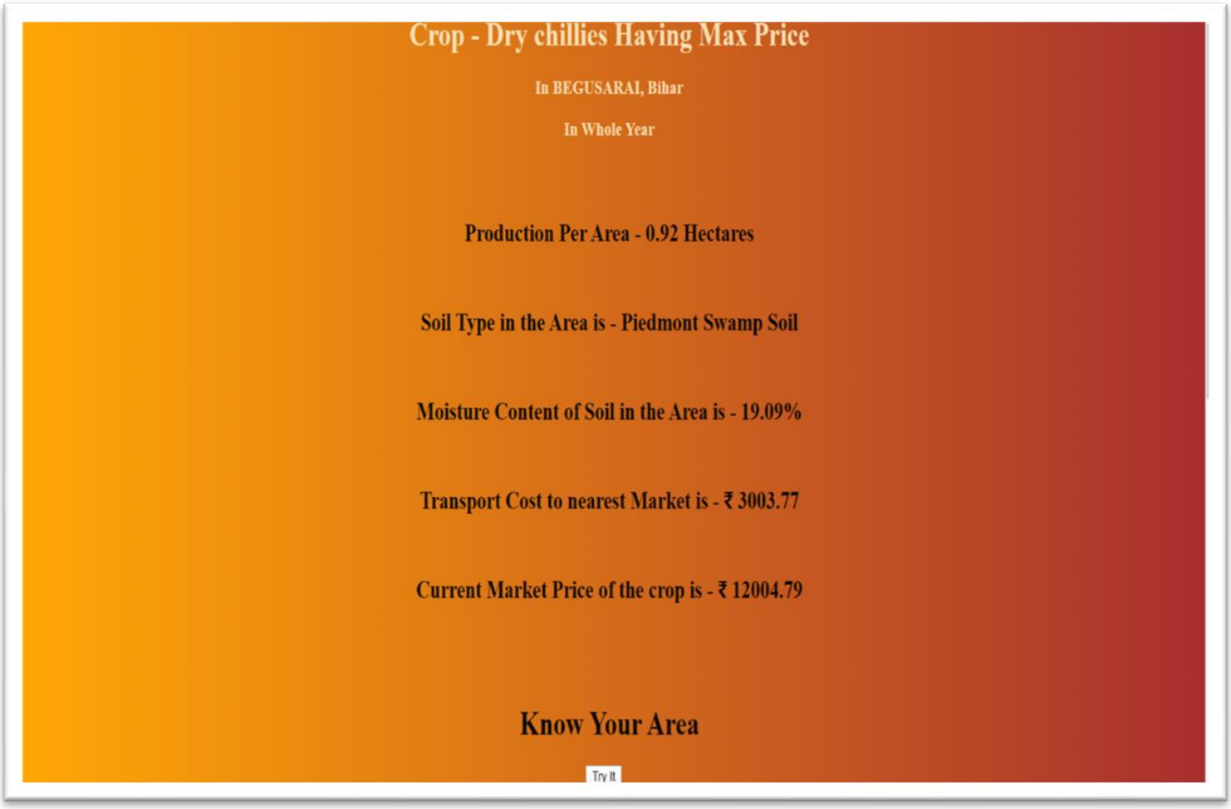


Questionaire Panel

1- Want to get the information about best three crops of your District with the production by area of a particular season ?

2- Know more about crop in your area

Result Screen



Crop - Dry chillies Having Max Price

In BEGUSARAI, Bihar

In Whole Year

Production Per Area - 0.92 Hectares

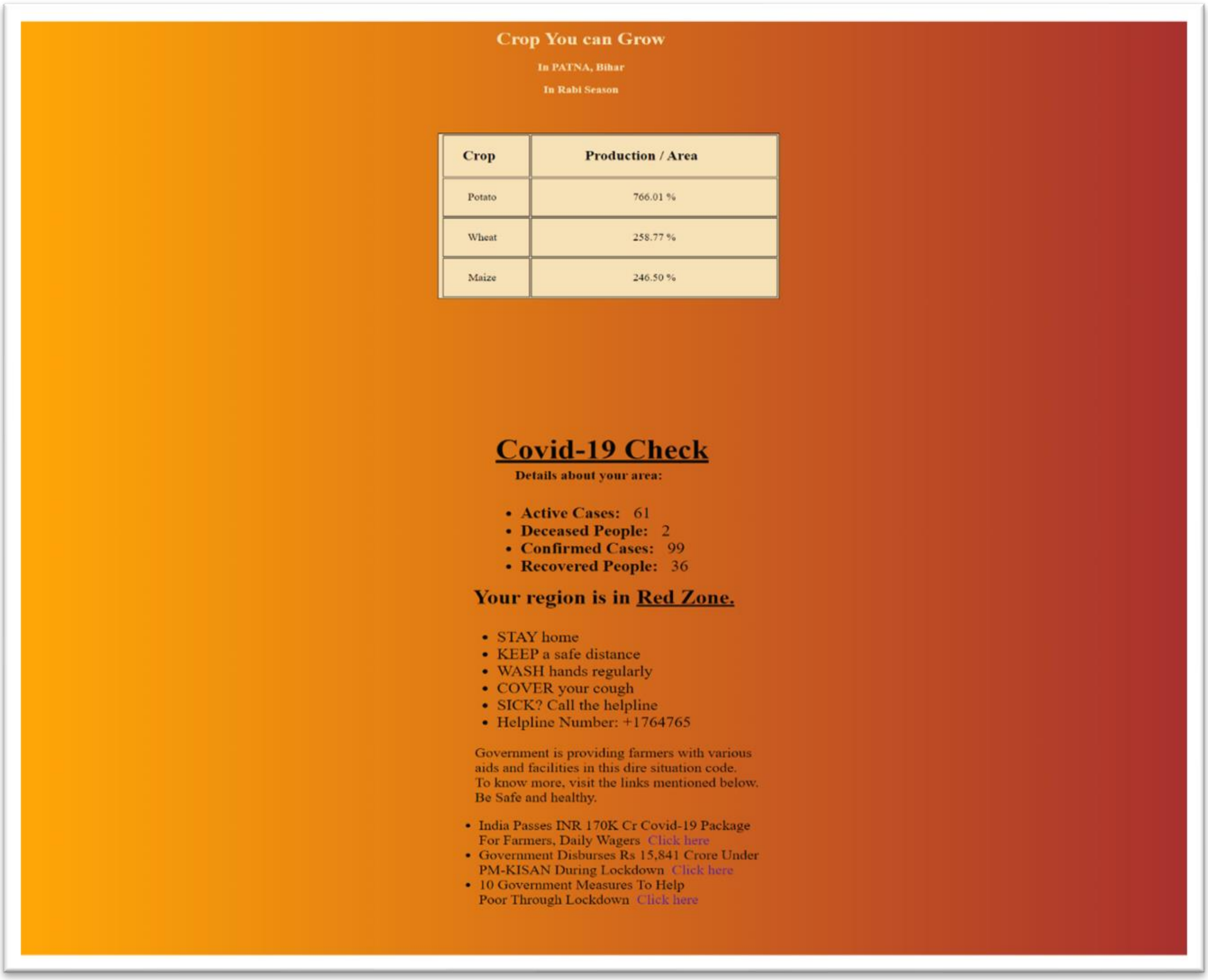
Soil Type in the Area is - Piedmont Swamp Soil

Moisture Content of Soil in the Area is - 19.09%

Transport Cost to nearest Market is - ₹ 3003.77

Current Market Price of the crop is - ₹ 12004.79

Know Your Area



Crop You can Grow

In PATNA, Bihar

In Rabi Season

Crop	Production / Area
Potato	766.01 %
Wheat	258.77 %
Maize	246.50 %

Covid-19 Check

Details about your area:

- **Active Cases:** 61
- **Deceased People:** 2
- **Confirmed Cases:** 99
- **Recovered People:** 36

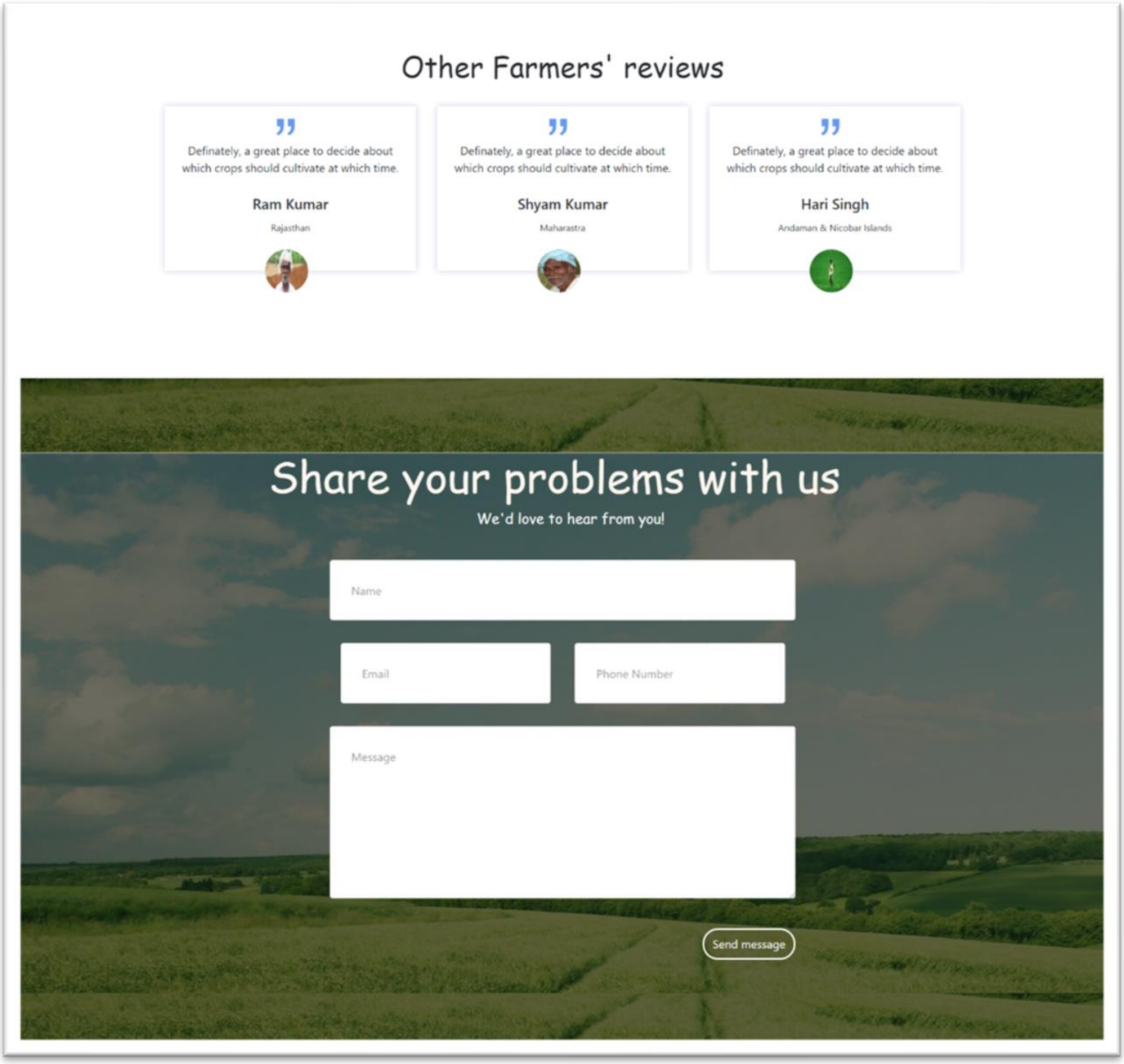
Your region is in Red Zone.

- STAY home
- KEEP a safe distance
- WASH hands regularly
- COVER your cough
- SICK? Call the helpline
- Helpline Number: +1764765

Government is providing farmers with various
aids and facilities in this dire situation code.
To know more, visit the links mentioned below.
Be Safe and healthy.

- India Passes INR 170K Cr Covid-19 Package For Farmers, Daily Wagers Click here
- Government Disburses Rs 15,841 Crore Under PM-KISAN During Lockdown Click here
- 10 Government Measures To Help Poor Through Lockdown Click here

Review Screen



Other Farmers' reviews

Definately, a great place to decide about which crops should cultivate at which time.

Ram Kumar

Rajasthan

Definately, a great place to decide about which crops should cultivate at which time.

Shyam Kumar

Maharastra

Definately, a great place to decide about which crops should cultivate at which time.

Hari Singh

Andaman & Nicobar Islands

Share your problems with us

We'd love to hear from you!

Chapter 6: Conclusion

6.1 Future Prospects

The technology in our time is changing and improving fast. Life has come to smart phones and tablets from desktops and laptops and everyone own smart phones. As of now this application is only available for farmers as a web application but the android application of this project will also be made by which some more problems of farmers will get reduced. Android application will be based on location-based services, by which farmers will have to make less efforts, they won’t need to type or select anything, they will just open up the application and on the basis of their location and choices, results will get displayed. Moreover, during an emergency they can be notified by push notifications and does not need upon the consumer to open the app to get informed and they can also be notified about the weather conditions of their regions periodically in a harvesting Season.

In future, on this project ML lib will be applied which is an apache spark's scalable machine learning library by using this library many algorithms like regression, clustering, collaborative and classification will get implemented on this application by which it will give better results and will predict more results.

6.2 Conclusion

Currently, a large number of people are relying on cultivation rather than office jobs due to their illiteracy. Unluckily their lack of training displays on their methods of cultivation. Here our society needs a higher supervision through technology. Agriculture demonstrator states that who do well do no longer achieve true effects. So, no boom is located in lives of cultivators. Within the destiny, this study might be scaled up in terms of data size and crop variations. We have made a project which helps to identify which crop will be helpful in which season and how to tackle natural calamities, so as to increase the production of the farmer.

References

1. C. Sekhar, C. &. Udaykumar, J. &. Kumar, B. &. Sekhar and Ch, "Effective use of Big Data Analytics in Crop planning to increase Agriculture Production in India.," 2018.

2. S. Sonka, "Big Data and the Ag Sector: More than Lots of Numbers," *IFAMA,* pp. 1-20, 2014.

3. G. N. Fathima, "Agriculture Crop Pattern Using Data Mining Techniques," pp. 781-786, 5 May 2014.

4. M. Muppidathi, "An Analysis for Big Data and its Technologies," *International Journal of Computer Science Engineering and Technology (IJCSET),*, pp. 412-418, 2014.

5. Misra, D. a. Mishra, A. a. Babbar, S. a. Gupta and Varun, "Open Government Data Policy and Indian Ecosystems," in *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance*, New Delhi , Association for Computing Machinery, 2017, pp. 218-227.

6. CSI, "IEEE International Conference by CSI on Big Data on IT in Business, Industry and Government," in *IEEE*, 2014.

7. A. Pal, K. Jain, P. Agrawal and S. Agrawal, A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop, Bhopal: 4th International Conference on Communication Systems and Network Technologies (CSNT), 2014.

8. e. Yang, "Data Clustering on Taiwan Crop Sales under Hadoop Platform," in *Proceedings of the Institute of Industrial Engineers Asian Conference*, 2013.

9. J. O. Chan, "An Architecture for Big Data Analytics," *Communications of the IIMA,* pp. 1-14, 2013.

10. RapidAPI, "Covid-19 India Data by ZT API Documentation," March 2020. [Online]. Available: Covid-19 India Data by ZT API Documentation.