ESTIMATE THE CROP YEILD USING DATA ANALYTICS

NALAIYATHIRAN PROJECT BASED LEARNING

on

PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY AND ENTERPRENUERSHIP

A PROJECT REPORT

Submitted by

MAHALAKSHMI V (113219031083)
RAMYA V S (113219031121)
KEERTHANA C (113219031071)
MAHALAKSHMI L (113219031082)

in partial fulfillment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

Velammal Engineering College

Chennai - 600066

An Autonomous Institution

Affiliated to Anna University, Chennai – 600 025 NOVEMBER

2022

Velammal Engineering College

Chennai- 600066

An Autonomous Institution

INTERNAL MENTOR

Mrs. S Sridevi

ASSISTANT PROFESSOR - II,

Department of Computer Science and Engineering,

Velammal Engineering College.

INDUSTRY MENTORS

Srikanth

Mahidhar

Mohammed Azhar Uddin

ABSTRACT

In order to address issues with food security and lessen the effects of climate change, it is essential to comprehend global agricultural productivity. One of the top UN Sustainable Development Goals for 2030 is to end hunger, and this goal can assist. In the project, we present a scalable, precise, and low cost approach to agricultural production prediction utilizing openly accessible remote sensing data and machine learning. Several months prior to harvest, our deep learning system can estimate crop yield using just globally accessible covariates with high spatial resolution (county-level). We think our technique may be useful for setting suitable food reserve levels, identifying low-yield locations, and enhancing risk management of derivatives related to crops. In India crop yield is season dependent and majorly influenced by the biological and economic cause of an individual crop. Reporting of progressive agriculture yield in all the seasons is an ample task and an advantageous task for every nation with the respect to assesses the overall crop yield and prediction and estimation at present a common issue worldwide is farmers are stressed in producing higher crop yield due to influence of unpredictable climatic changes and significant reduction of water resources worldwide. A study was carried out to collect the data on world climatic changes and the available water resources which can be used to encourage advanced and novel approaches such as big data analytics to retrieve the information of the previous results to the crop yield prediction and estimation. Study imported that the selection and usage of the most desirable crop according to the existing conditions, supports to achieve the higher and enhanced crop yield. An explicit rationale model which can effectively applied at various levels of the availability of quality information for identifying data sources to analyze crop yield and measuring yield gaps at definite geographical locations and works based on the rise in titer approach. The model is highly helpful in retrieving the useful data from the available, poor quality, less rigorous data sources or if the data is not available. A case study was discussed on the application of selected model design to quantify the yield gaps of maize crop in the state of Nebraska (USA), and also at the different geographical locations representing the nations Argentina and Kenya at national scale level. Different geographical locations such as Nebraska (USA), Argentina and Kenya were identified to symbolize the distinct scenarios of Agri based data availability and the quality for the selected variables assessed to predict and estimate the crop yield gaps. The definitive aspiration of the planned method is to afford transparent, easily accessible, reproducible and technically sound and strong guidelines for predicting the yield gaps. The proposed guidelines were also relevant for understanding and to simulate the influence of change in climate conditions and usage of cultivable land changes from national to global scales. As indicated, the better understanding of data.

TABLE OF CONTENTS

INTRODUCTION

1.1 Project Overview

India is a predominantly agricultural nation. Agriculture is currently the most significant emerging sector in the actual world and the key industry and economic pillar of our nation. The discipline of agricultural information technology has recently undergone significant changes that have made crop yield prediction an interesting research topic. Crop yield prediction is a technique for estimating crop yield using many characteristics, including temperature, rainfall, fertilizers, insecticides, and other climatic variables and parameters. Using data analytics to analyze those parameters and provide the patterns or trends that has been followed over the past years in estimating the yield can help farmers to make right choice in the selection of crop varieties,etc.,To make people gasp and use the knowledge represented we finally put the visualizations made in a dashboard and represent it with the most suitable and appropriate charts or graphs or maps.

1.2 Purpose

Analytics is the interpretation of data pattern that assist decision-making and performance improvement. Agriculture Data analytics in crop yield helps in analyzing some important visualization, creating a dashboard and by going through these we will get most of the insights of Crop production in India.

LITERATURE SURVEY

2.1 Existing problem

- **Crop Yield Prediction Using Machine Learning**: A research group investigated the utilization of various information mining methods which will foresee rice crop yield for the data collected from the state of Maharashtra, India. A total of 27 regions of Maharashtra were selected for the assessment and the data was collected related to the principle rice crop yield influencing parameters such as different atmospheric conditions and various harvest parameters i.e Precipitation rate, minimum, average, maximum and most extreme temperature, reference trim cultivable area, evapotranspiration, and yield for the season between June to November referred as Kharif, for the years 1998 to 2002 from the open source, Indian Administration records. WEKA a Java based dialect programming for less challenging assistance with information data sets, assigning design outcomes tool was applied for dataset processing and the overall methodology of the study includes,

    1. pre-processing of dataset
    2. Building the prediction model utilizing WEKA and
    3. Analyzing the outcomes.

Cross validation study is carried out to scrutinize how a predictable information mining method will execute on an ambiguous dataset. Study applied 10-fold higher cross validation study design to assess the data subsets for screening and testing. Identified and collected information was randomly distributed into 10 sections where in one data

section was used for testing while all other data sections were utilized for the preparation information. Study reported that the method applied was supportive in the precise estimation of rice crop yield for the state of P. Priya et al., (2018) has proposed a random Forest Algorithm for predicting the crop yield of particular area considering various parameters such as rainfall, seasonal crop (Rabi and Kharif) district-wise, temperature (max.),crop production in terms of Kgs/tonnes. Area for doing research was Tamil Nadu. Dataset record were collected from Indian Government over 15years for rice production. They proved in experimental results that prediction analysis done using Random Forest Algorithm – a supervised machine learning algorithm will help farmer to predict the yield of the crop before cultivating onto the agricultural field. This algorithm run efficiently on large databases with high classification accuracy.

- **Crop Yield Prediction Using Data Mining Techniques** : Raor ane A.A. and Kul karn i R.V., discussed few data mining techniques in their paper. They concluded that efficient technique can be developed and analyzed using the appropriate data, to solve complex agricultural problems using data mining techniques. Also recommend some of the algorithms and statistical methods that give [8]good results in agriculture growth.

- **Crop yield prediction using Big Data Analytics**: In India crop yield is season dependent and majorly influenced by the biological and economic causes of an individual crop. Reporting of progressive agricultural yield in all the seasons is an ample task and an advantageous task for every nation with respect to assesses the overall crop yield prediction and estimation. At present a common issue worldwide is, farmers are stressed in producing higher crop yield due to the influence of unpredictable climatic changes and significant reduction of water resource worldwide. A study was carried out to collect the data on world climatic changes and the available water resources which can be used to encourage advanced and novel approaches such as big data analytics to retrieve the information of the previous results to the crop yield prediction and estimation. Study imported that the selection and usage of the most desirable crop according to the existing conditions, support to achieve the higher and enhanced crop yield.

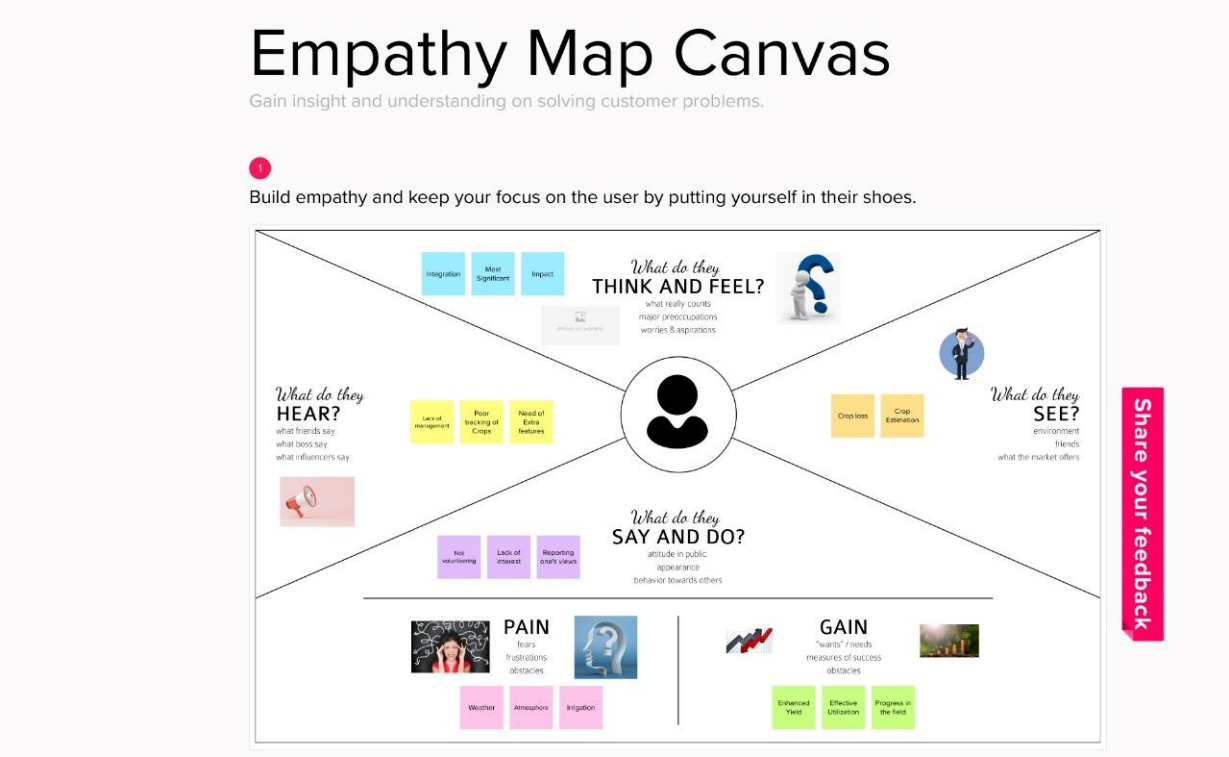
  S. Athmaja, M. Hanumanthappa, and V. Kavitha,a survey of machine learning algorithms has presented effective strategies by for big data analytics. All over the world the agricultural peoples gained some advantages through the comparative knowledge from big data analysis, with machine learning algorithm by using huge data the agricultural peoples get some comparative knowledge and changes in regular agriculture

2.2 Problem Statement Definition

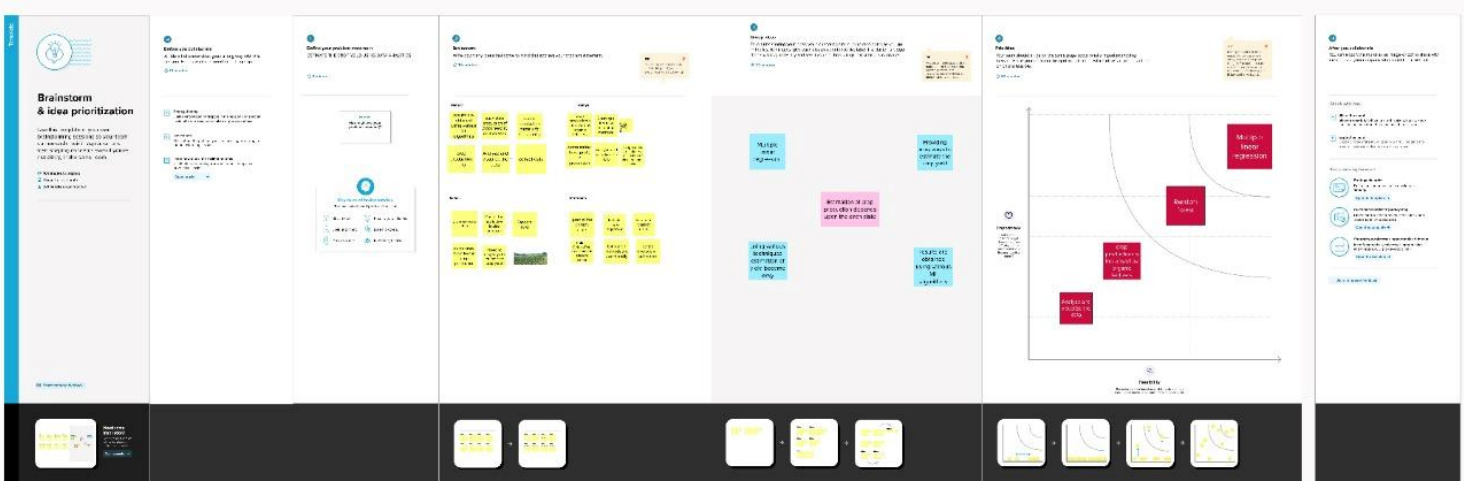
In the agriculture sector the farmers are facing difficulties in analyzing the demand in market and soil quality analysis to achieve high crop yield through technology. The main objective of this project is to predict crop yield that will be extremely useful to farmers to plan for the harvest and sales of harvested grain.

IDEATION & PROPOSEDSOLUTION

3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming



3.3 Proposed Solution

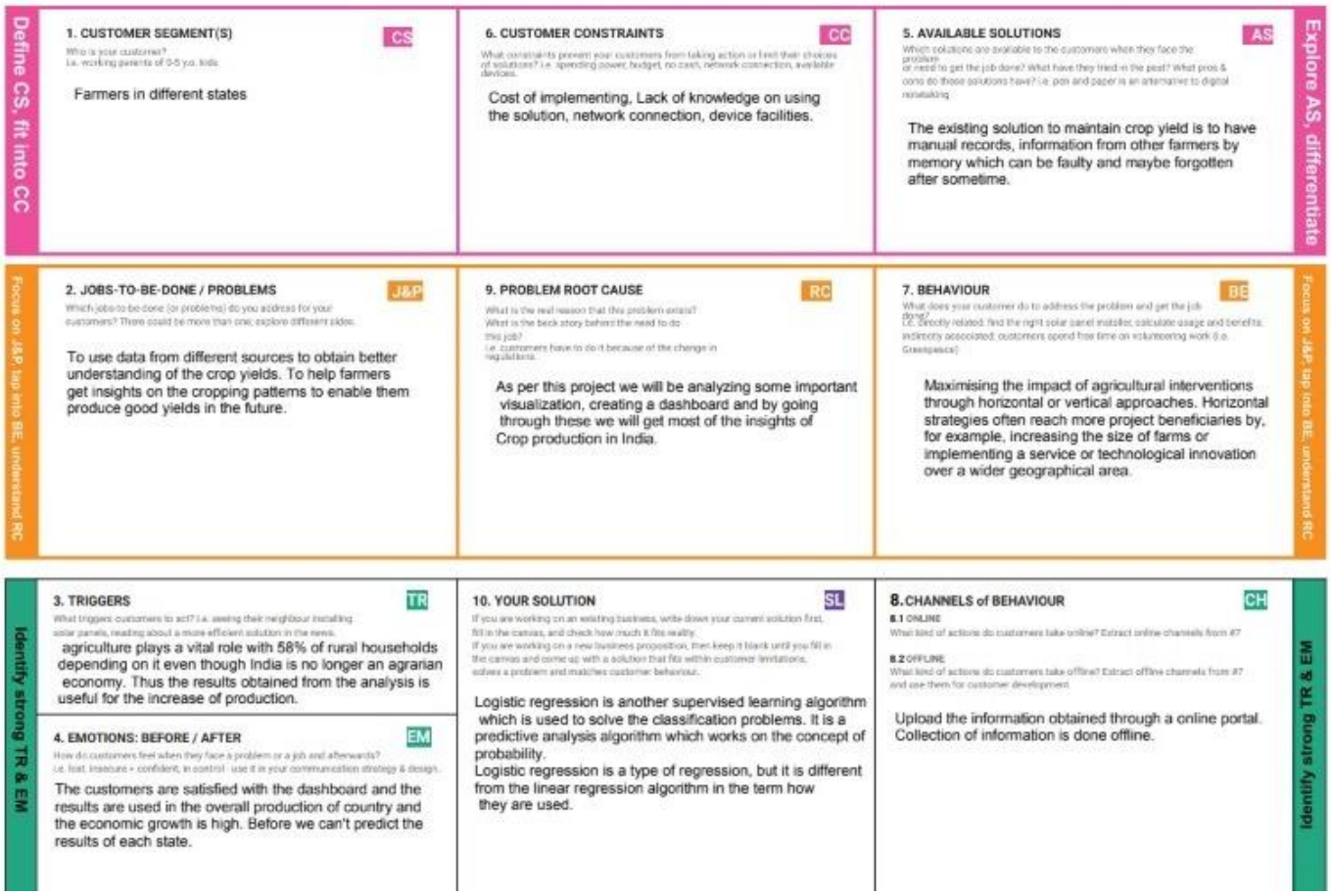
S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	In the agriculture sector the farmers are facing difficulties in analyzing the demand in market and soil quality analysis to achieve high crop yield through technology. The main objective of this project is to predict crop yield that will be extremely useful to farmers to plan for the harvest and sales of harvested grain.
2.	Idea / Solution description	Provide perfect data report after deep analysis of the past data. Helping them out to overcome loss in farming and business.
3.	Novelty / Uniqueness	With this solution we can analyze, visualize data and give the farmers the option to choose which plant/crop to cultivate in which period of time/season to earn more profit from the crop yield.
4.	Social Impact / Customer Satisfaction	Perfect data visuals create a large impact in the crop yield. And hence farmers will be able to gain more profit.
5.	Business Model (Revenue Model)	We can increase/enhance crop production and other raw materials. Also, Increase in productivity will result in increase of Revenue for the farmers.
6.	Scalability of the Solution	With the data visual reports, farmers will be able to cultivate crop according to the area, climate, soil and other features that impact the crop yield and hence enhancing the productivity.

3.3 Problem Solution fit

Project Title: Estimate the Crop Yield using Data Analytics

Project Design Phase-I - Solution Fit Template

Team ID: PNT2022TMIDxxxxxx



REQUIREMENT ANALYSIS

4.1 Functional requirement

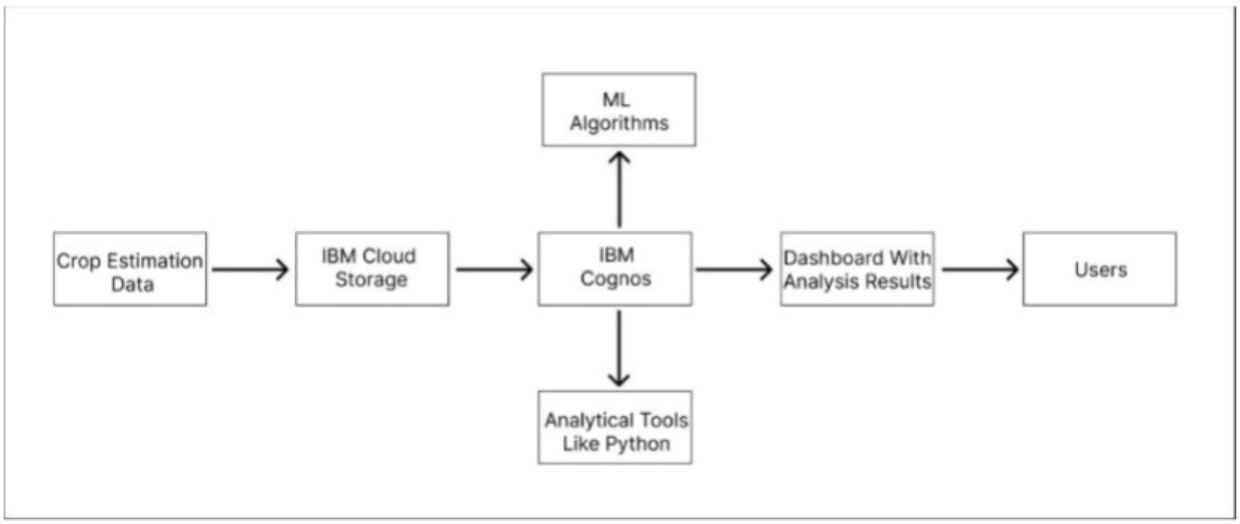
FR No.	Functional Requirement (Epic)	Sub Requirement (Story/ Sub-Task)
FR-1	Knowledge about factors that influence the yield	Behaviour of crops and the yield obtained is highly dependent on factors like rainfall, temperature, soil type, etc., Hence it is significant to know the impact of these factors on the yield with its past history.
FR-2	Estimation module	A prediction of crop yield is to be done based on the user's input data (season ,crop ,production ,area).
FR-3	Analysis	An analysis is done on the given data to gain useful insights on the crop yield.

4.2 Non-Functional requirements

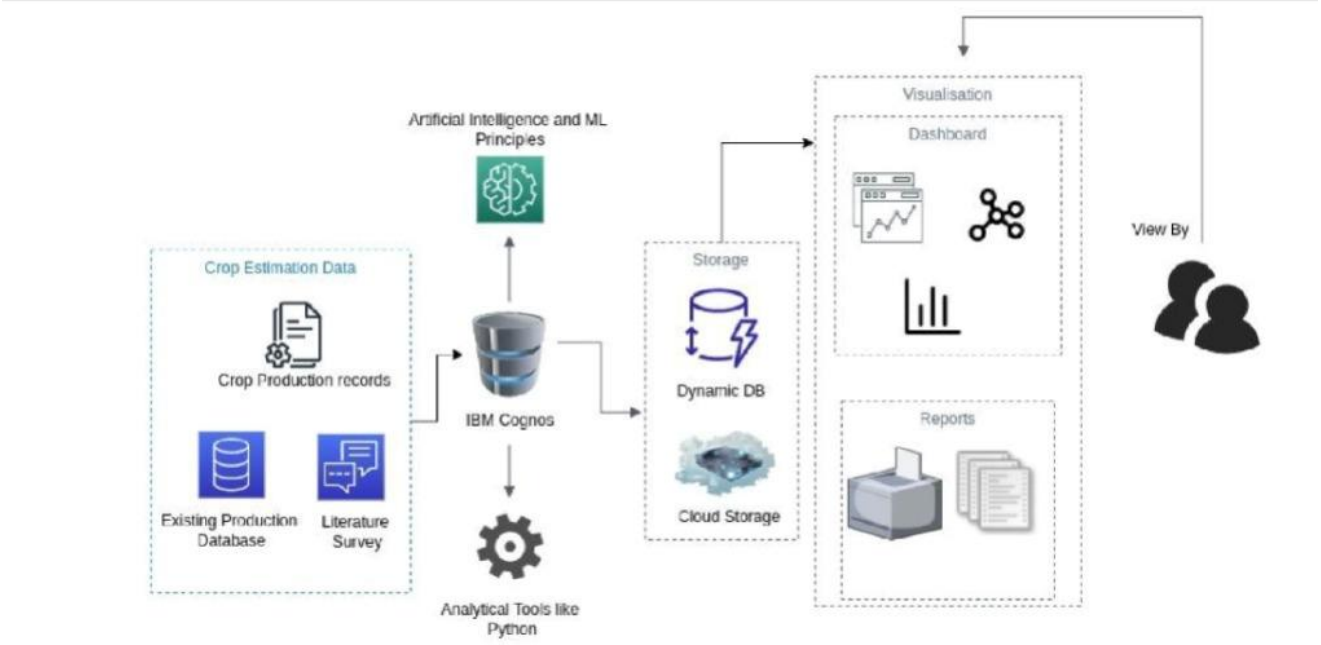
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Provide perfect data report after deep analysis of the past data. Helping farmers to overcome loss in farming and business.
NFR-2	**Security**	The user information is protected by the user login and registration with a secured password.
NFR-3	**Reliability**	Effective tool that all farmers can use, making it reliable by improving the accuracy of the estimation or prediction. This will bridge the gap between farmers and technology.
NFR-4	**Performance**	Multiple technologies and services that will improve the usability in agricultural activities.
NFR-5	**Availability**	Both website and mobile application interface and developed in local language and the content is available in localized language.
NFR-6	**Scalability**	With the data visual reports, farmers will be able to cultivate crop according to the area, climate, soil and other features that impact the crop yield and hence enhancing the productivity.

PROJECT DESIGN

5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture



5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user and Laptop users)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard	USN-6	Can use the methods provided in the dashboard		Medium	Sprint-1
	Invest	USN-7	With help of desired results obtained from application ,making profit or loss	Gain or Loss	High	Sprint-2
Administrator	Updating data		Collecting the data and storing it	Checking and updating dataset	High	Sprint-1

PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Loading the dataset	USN-1	Need to load the dataset in the IBM Cognos for the upcoming process	2	High	Mahalakshmi V Mahalakshmi L Keerthana C Ramya V S
Sprint-2	Working with dataset (Removal of outliers etc)	USN-2	Providing dataset with all the required components.	2	High	Mahalakshmi V Mahalakshmi L Keerthana C Ramya V S
Sprint-3	Visualizing the data (Top 10 states with most area)	USN-3	Ensuring that the required fields for visualizing the given dataset are provided by us.	1	High	Mahalakshmi V Mahalakshmi L Keerthana C Ramya V S
Sprint-4	Dashboard creation	USN-4	Make use of the dashboard to see the results of the crop production in respective areas.	2	Low	Mahalakshmi V Mahalakshmi L Keerthana C Ramya V S
Sprint-4	Export the analytics	USN-5	The dashboard is shared as mail or link or pdf such that crop production results can be displayed to others.	2	Medium	Mahalakshmi V Mahalakshmi L Keerthana C Ramya V S

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	10	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	10	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	10	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	10	19 Nov 2022

Data Exploration

7.1 Detecting the presence of null values

jupyter Crop Production Last Checkpoint: Last Saturday at 21:35 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        df  = pd.read_csv("crop_production.csv")

In [2]: print(str('Any missing data or NaN in the dataset:'),df.isnull().values.any())
        Any missing data or NaN in the dataset: True

In [3]: df.isnull().sum()
Out[3]: State_Name          0
        District_Name       0
        Crop_Year           0
        Season              0
        Crop                0
        Area                0
        Production       3730
        dtype: int64
```

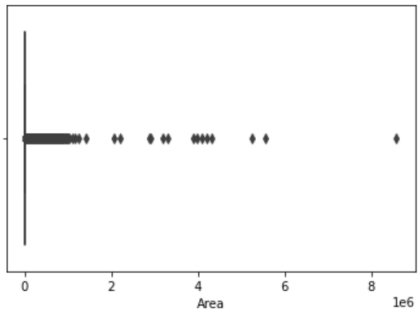
7.2 Dropping the null values

```
In [6]: df1=df.dropna()

In [7]: df1.isnull().sum()
Out[7]: State_Name       0
        District_Name    0
        Crop_Year        0
        Season           0
        Crop             0
        Area             0
        Production       0
        dtype: int64
```

7.3 Detecting and removing outliers

For Area

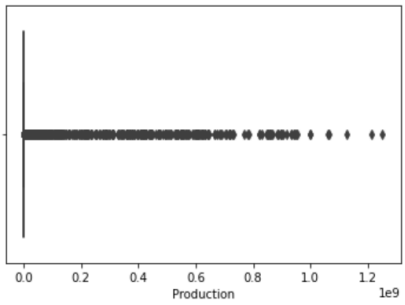
```
In [8]: import seaborn as sns
        sns.boxplot(x=df1["Area"])
Out[8]: <AxesSubplot:xlabel='Area'>
```



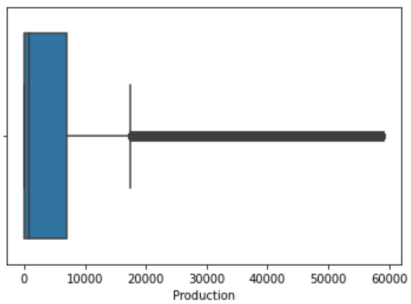
```
In [9]: print(df1['Area'].quantile(0.10))
        print(df1['Area'].quantile(0.90))

        13.0
        27806.0

In [10]: import numpy as np
         df1["Area"] = np.where(df1["Area"] <13.0,13.0 ,df1['Area'])
         df1["Area"] = np.where(df1["Area"] >27806.0,27806.0,df1['Area'])
         print(df1['Area'].skew())

         1.8314510551117793
```

For Production

```
In [13]: sns.boxplot(x=df1["Production"])

Out[13]: <AxesSubplot:xlabel='Production'>
```




```
In [14]: print(df1['Production'].quantile(0.10))
         print(df1['Production'].quantile(0.90))

         10.0
         58941.0

In [15]: df1["Production"] = np.where(df1["Production"] <10.0,10.0 ,df1['Production'])
         df1["Production"] = np.where(df1["Production"] >58941.0, 58941.0,df1['Production'])
         print(df1['Production'].skew())

         1.9483161360423034

In [16]: sns.boxplot(x=df1["Production"])

Out[16]: <AxesSubplot:xlabel='Production'>
```



WORKING WITH THE DATASET& DATA VISUALISATION

8.1 Understanding the dataset

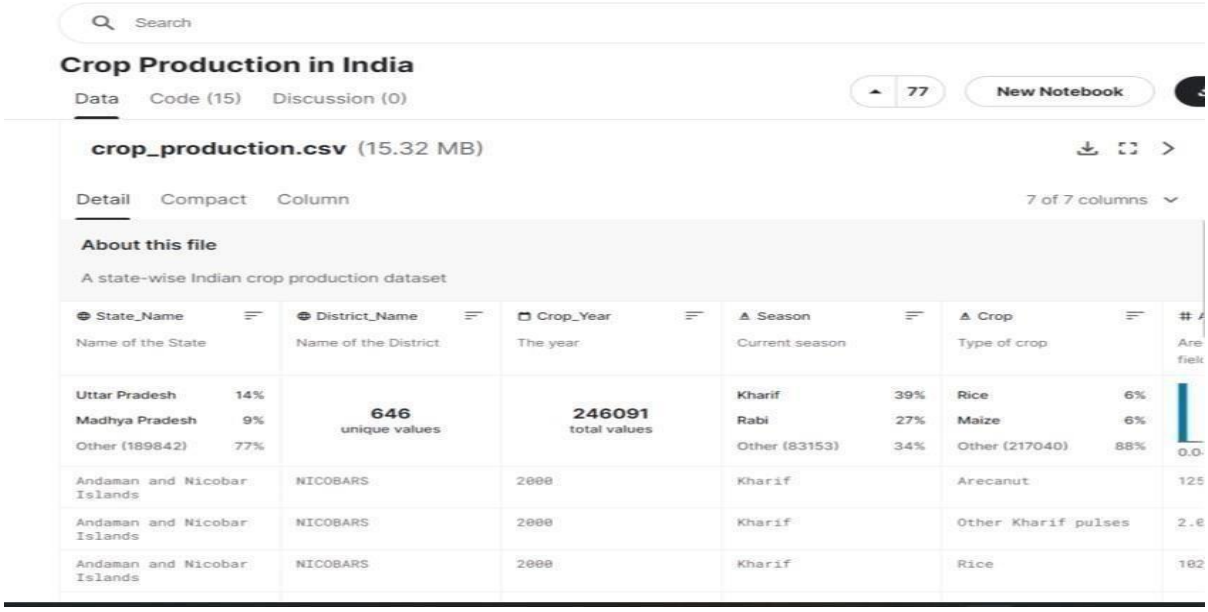
This project is based on a understanding the crop production of India. It has 2,46,092datapoints (rows) and 6 features (columns)describing each crop production related details.

Dataset Link: Dataset

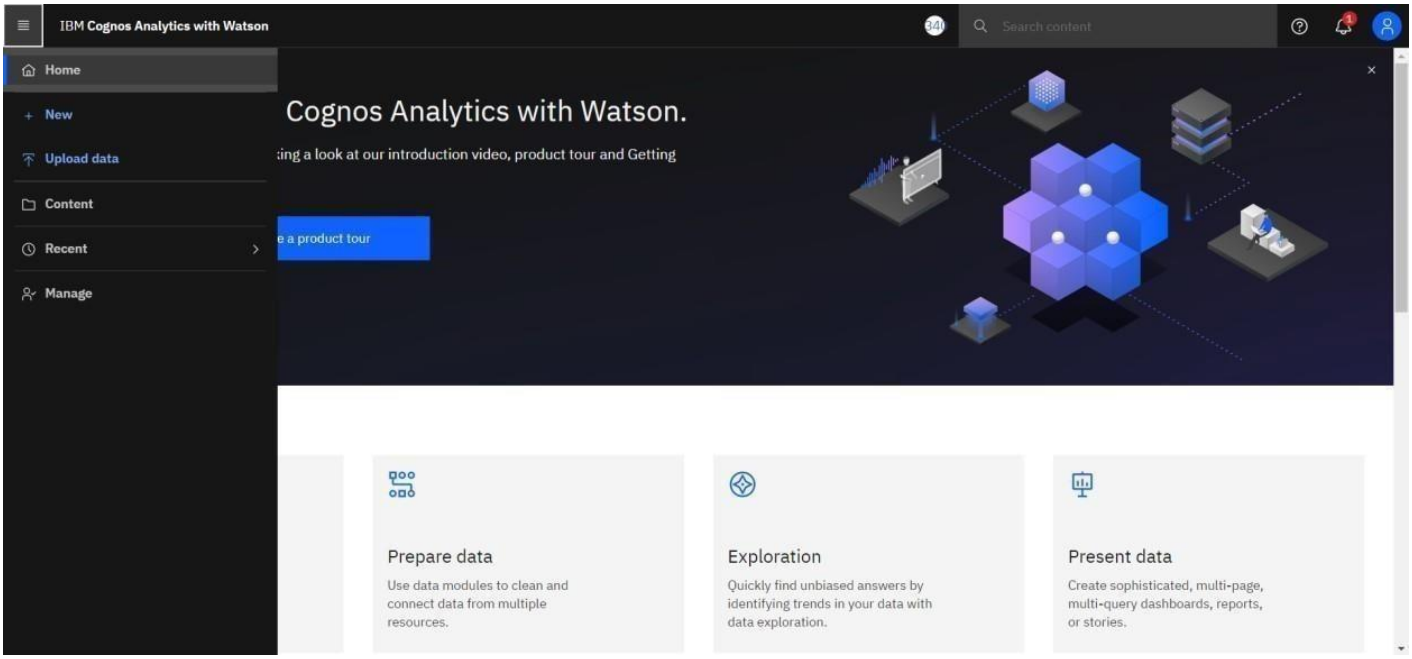
Let's understand the data we're working with and give a brief overview of what each feature represents or should represent

1. StateName - All the IndianState names.
2. District Name -Different District names.
3. Crop Year- containsthe crop years.
4. Season – Different seasons for crop production.
5. Area- Total numberof areas covered.
6. Production- production of crops.

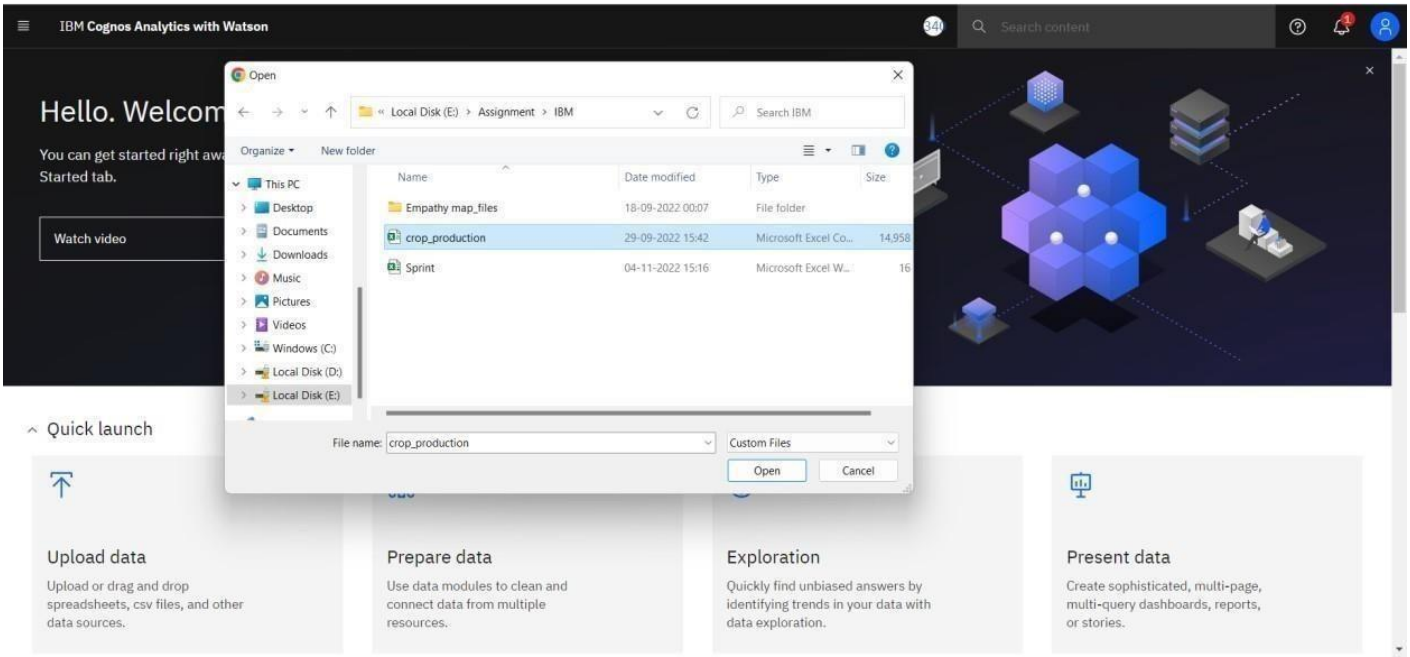
The data format is as shownin the below image:



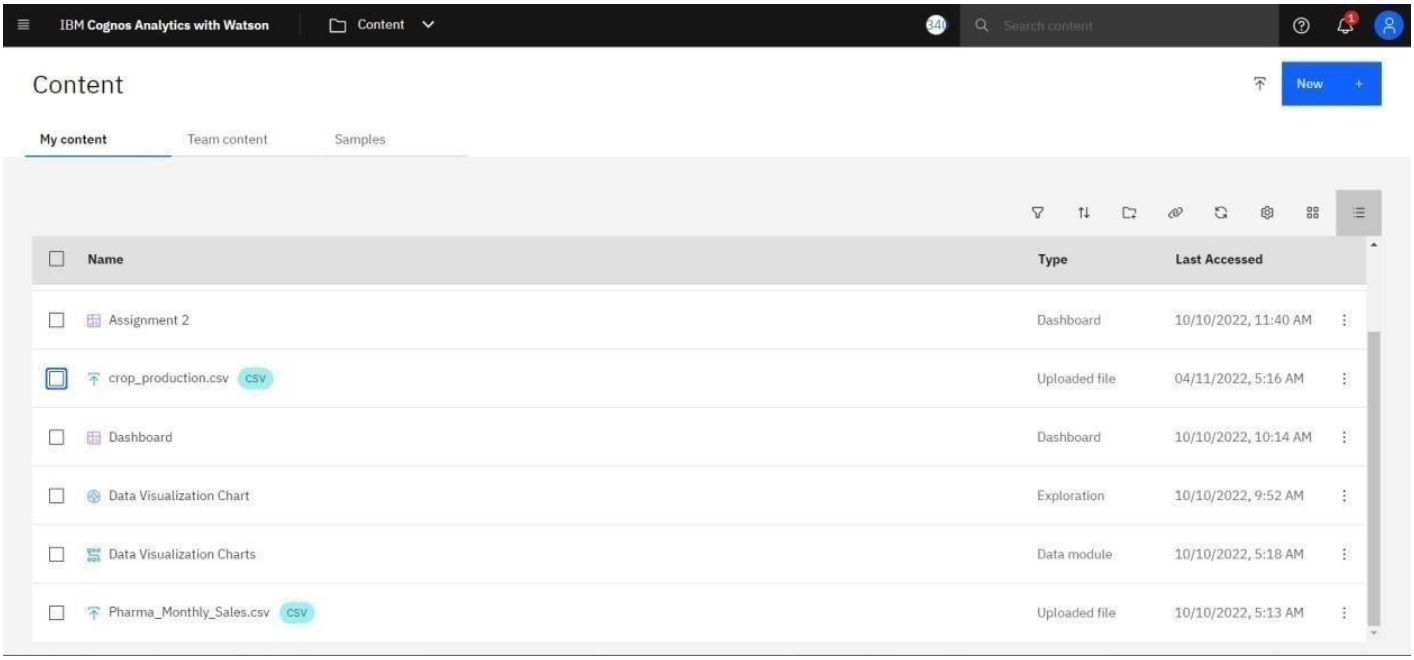
8.2 Loading the dataset



- Click the open menu in the top left corner.
- Select the **Upload Data** in the menu and select the Dataset that you want upload.

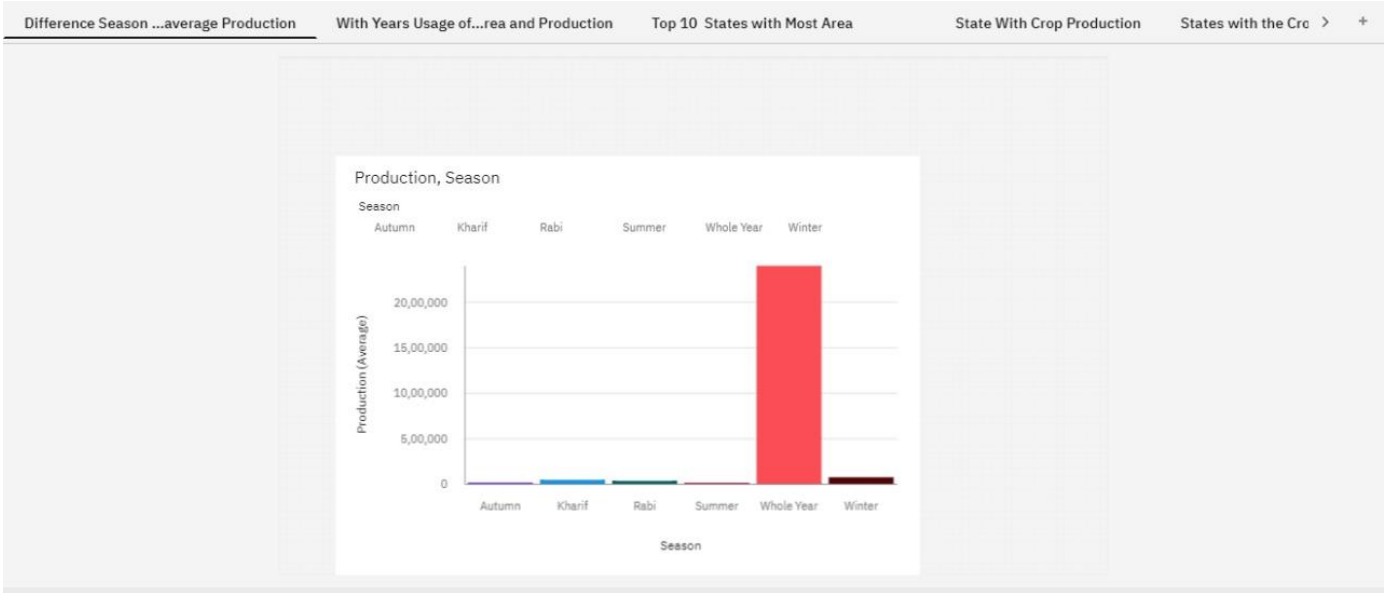


- Once the Dataset is Uploaded it will be displayed in content.

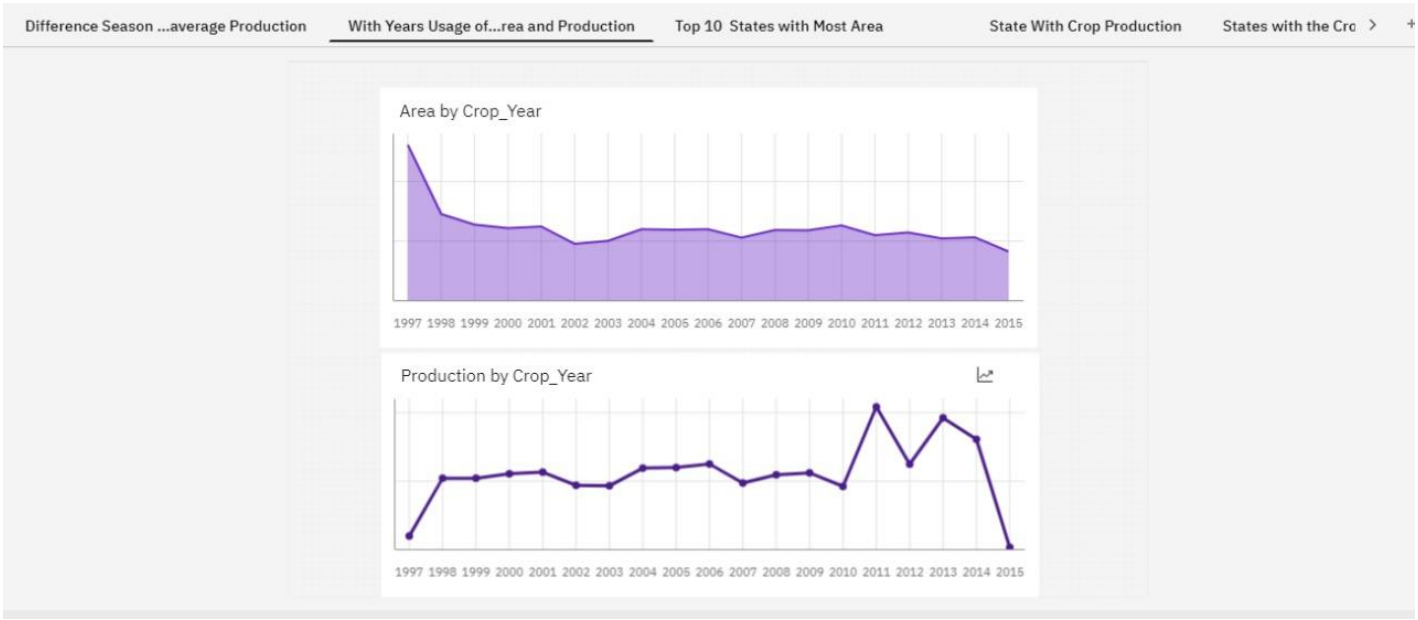


8.3 Visualization charts

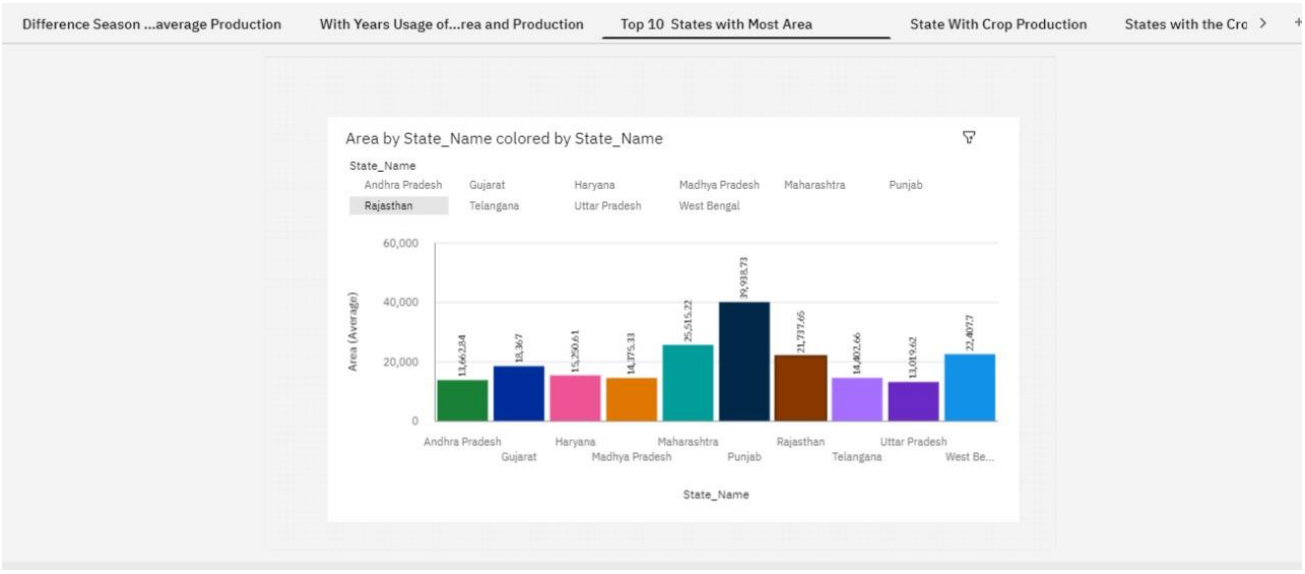
- Seasons with average productions



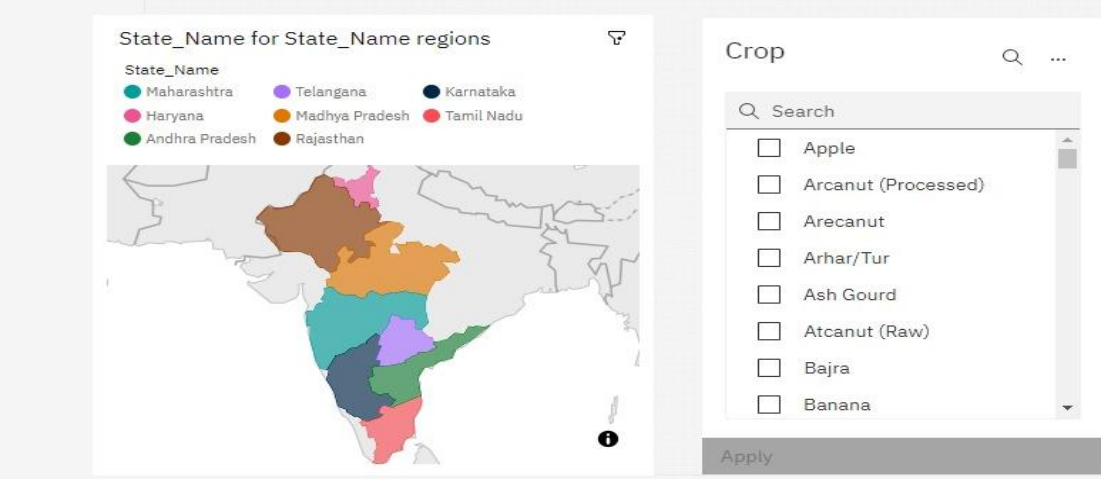
- With years usage of Area and Production



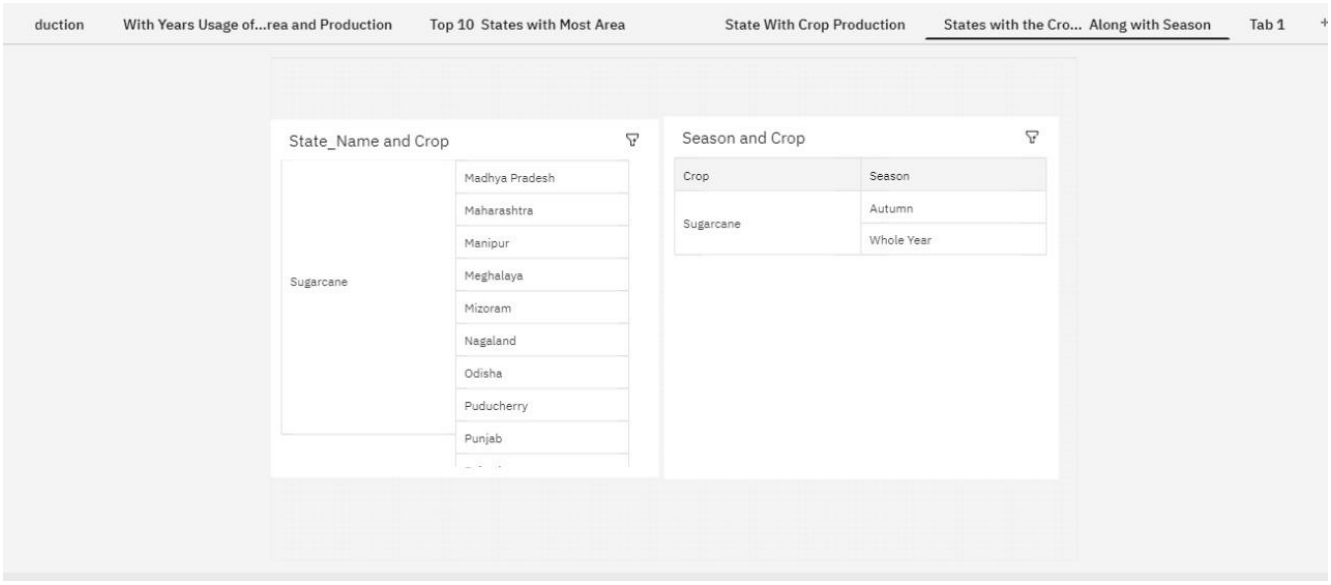
- Top 10 States with most area



- State with crop production

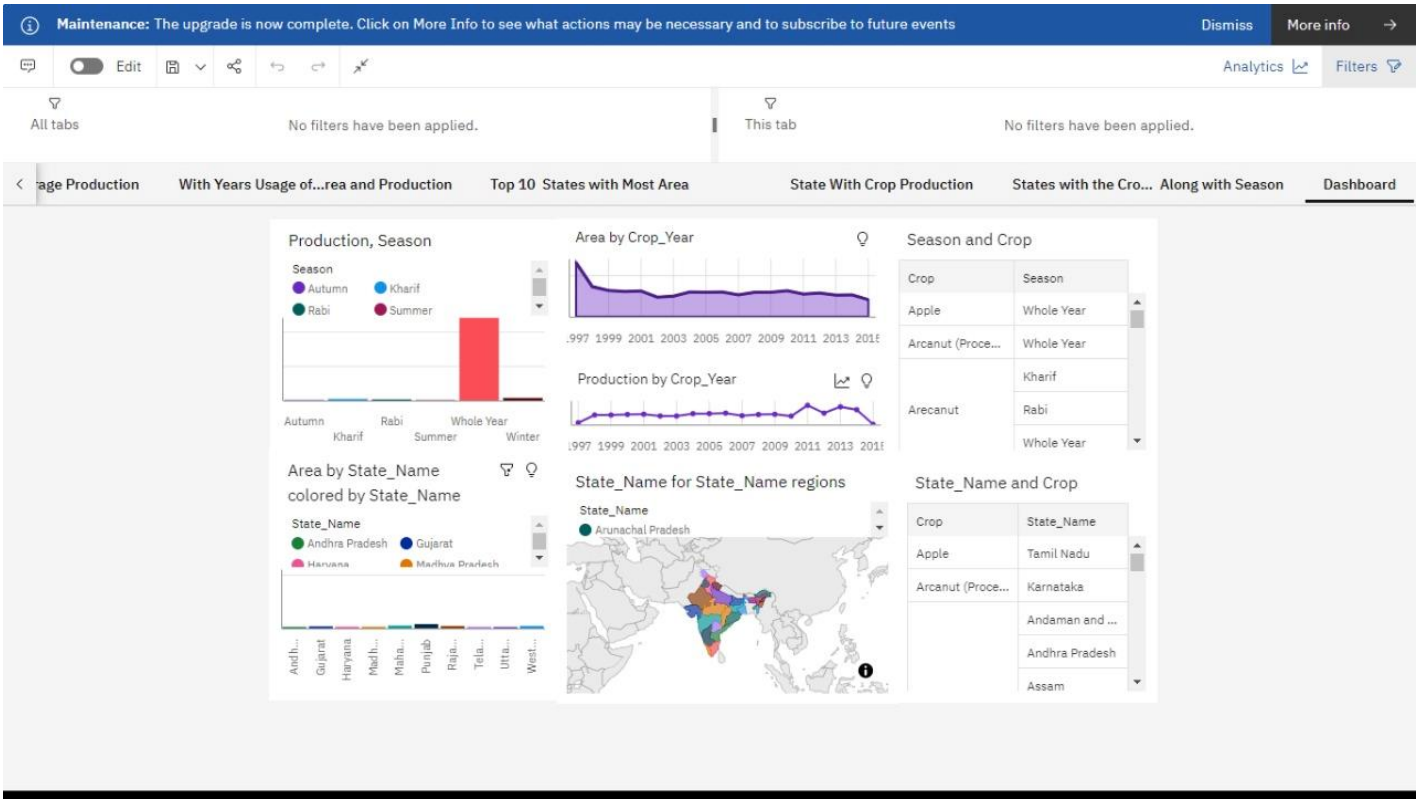


- States with the crop production along with season (Text Table)



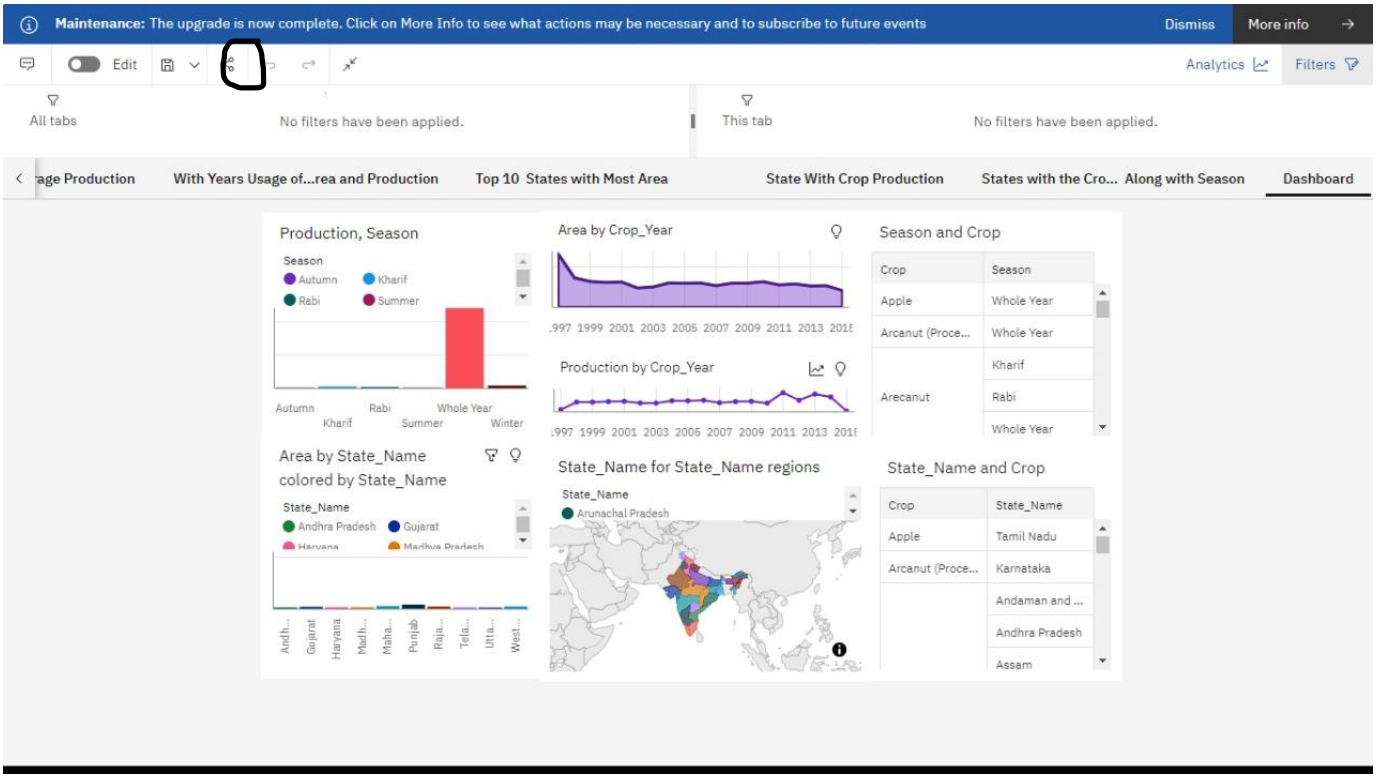
CREATING THE DASHBOARD AND EXPORT THE ANALYTICS

9.1Creating the Dashboard



9.2 Export the Analytics

- Click on the **share icon**
- You can share using Email or Link or Export as pdf.
- Click the Export tab in the Share dialog box.
- You can change the page size and Orientation setting then click **Export**.



ADVANTAGES & DISADVANTAGES

Advantages:

One can easily analyze and understand trends in cropping pattern, seasonal behavior of land in various areas with the created dashboard. With no prior skills and knowledge about the tools that we use for analysis, anyone (literate or illiterate) can easily infer the knowledge that we represent in various charts or graphs or maps. So that it would be helpful to farmers to make appropriate decisions in the future.

Disadvantages:

Not all factors influencing the crop yield are being considered for the analysis as we have only taken visible factors into account for the analysis

CONCLUSION

The productivity of agriculture has slightly increased as a result of technology's introduction. New ideas like digital agriculture, smart farming, precision agriculture, etc.
have been made possible by the innovations. From the analysis dashboard, it has been noted that analyses of agricultural productivity and the detection of hidden patterns utilizing data sets related to seasons and crop yields have been conducted. Using IBM Cognos, we have observed and conducted analysis
on various crops grown, area, and productions in various states and districts, including

1. Seasons with average productions. We learn from these analytics which seasons have higher average production and which have lower production.
2. Production split up per crop year. We learn from this study which years have high and low production.
3. District-based production. With the help of these analytics, we may identify the states and districts that farm the chosen crops.
4. Production by area. This will allow us to estimate the yield and determine how much land needs to be planted. After creating the dashboard,

study was done to determine which state, which year, and how much crop area will be produced.

FUTURE SCOPE

Farming is the means of survival as humans require food that is obtained only through farming directly or indirectly. With the growing human population, it is critical to analyze the production in farming every year. So, that we can know the right time, right place and right crop to be cultivated considering all the factors that influence the crop production.

It can be concluded that the research in the field of agriculture with reference to using IT trends like data analytics is in its infancy. As the food is the basic need of humans, the requirement of getting the
maximum yields using optimal resource will become the necessity in near future as a result of growing population. The survey outcomes indicate the need for improved techniques in crop yield analytics. There exists a lot of research scope in this research area.

APPENDIX

Source Code Link:
https://mahalakshmi-v18.github.io/Project/

Demo Link:
https://drive.google.com/file/d/1h4jHBeF2p2sMWMfWNu3fKvybW6TOJOxg/view?usp=sharing

GitHub Link:
https://github.com/IBM-EPBL/IBM-Project-23139-1659869010