ESTIMATE THE CROP YIELD USING DATA ANALYTICS

A PROJECT REPORT

Submitted by

TEAM ID: PNT2022TMID31255

HEERA .Y (621519205013)

ISHWARYA . S (621519205014)

JANANI . P (621519205015)

JANSIRANI . R (621519205016)

MAHESHWARI . R (621519205022)

in the partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

INFORMATION TECHNOLOGY

In

MAHENDRA COLLEGE OF ENGINEERING

ANNA UNIVERSITY: : CHENNAI 600 025

TABLE OF CONTENTS

1.Introduction

1.1 Project overview:

Agriculture is the backbone of Indian Economy. In India, majority of the farmers are not getting the expected crop yield due to several reasons. The agricultural yield is primarily depends on weather conditions. Rainfall conditions also influences the rice cultivation. In this context, the farmers necessarily requires a 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 in their crops. Yield prediction is an important agricultural problem. Every farmer is interested in knowing, how much yield he is about expect. In the past, yield prediction was performed by considering farmer's previous experience on a particular crop. The volume of data is enormous in Indian agriculture. The data when become information is highly useful for manypurposes.

IBM Cognos Business Intelligence is a web-based integrated business intelligence suite by IBM. It provides a toolset for reporting, analytics, score carding, and monitoring of events and metrics. The software consists of several components designed to meet the different information requirements in a company. IBM Cognos has components such as IBM Cognos Framework Manager, IBM Cognos Cube Designer, IBM Cognos Transformer. Cognos Analysis Studio helps business users get fast answers to business-related queries. Reporting studio allows you to create pixel-perfect reports for your organization. Cognos event studio allows you to assign a specific event that sends a notification to the stakeholder in your organization. Cognos Metric Studio allows you to monitor and analyze business metrics of your organization by building a scorecard environment.

1.2 Purpose

Agriculture is the most important sector that influences the economy of India. It contributes to 18% of India’s Gross Domestic Product (GDP) and gives employment to 50% of the population of India. People of India are practicing Agriculture for years but the results are never satisfying due to various factors that affect the crop yield. To fulfill the needs of

Around 1.2 billion people, it is very important to have a good yield of crops. Due to factors like soil type, precipitation, seed quality, lack of technical facilities etc. the crop yield is directly influenced. To focuses on implementing crop yield prediction system by using Machine learning techniques by doing analysis on agriculture dataset. For evaluating performance Accuracy is used as one of the factors. The classifiers are further compared with the values of Precision, Recall and F1score. Lesser the value of error, more accurate the algorithm will work. The result is based on comparison among the classifiers.

Preparing the Dataset:

The demo dataset is now supplied to machine learning model on the basis of this data set the model is trained. Every new detail filled at the time of application form acts as a test data set. After the operation of testing, model prediction based upon the inference it concludes on the basis of the training data sets. Satellite Imagery (Remote Sensing Data), has been widely used for predicting crop yield. This dataset is collected using the sensors mounted on satellites or planes, which detect the energy (electromagnetic waves), reflected or diffracted from surface of the earth. Remote sensing data has a lot of energy bands to offer, but mainly only few of them have been used for crop yield prediction. Yet, there are some people who have tried generating relevant features using the bands which are typically ignored, and they have been successful with improving results with that. In case of this dataset, most people rarely explore the high-order moments of the features. Based on these datasets people have used algorithms like Regression models, Random Forest and Nearest Neighbor etc.

Table shows details of the datasets:

Variable	**Description**
Crop	Crop name
State Name	Indian state name
District Name	District name list of each state
Cost of Cultivation (`/Hectare) C2	Cultivation amount for C2 Scheme
Cost of Production (`/Quintal) C2	Production amount for A2+FL Scheme
Yield (Quintal/ Hectare)	Yield of crop
Crop year	Crop year list
District Name	District name for each state
Area	Total area of each place
Rainfall	Water availability of each crop

Average humidity	directly influences the water relations of plant and indirectly affects leaf growth
Mean Temperature	Climate of each crop
Cost Production of per yield crop	Cost of crop yield

Scope:

The scope of this project is to investigate a dataset of crop records for agricultural sector using machine learning technique. To identifying crop predicting by farmer is more difficult. We try to reduce this risk factor behind selection of the crop

Objectives:

- Data validation
- Data Cleaning/ Preparing
- Data Visualization
- Using more algorithm with comparing to predict more accuracy (Like random forest, Decision tree Logistic classification algorithm

2.LITERATURE SURVEY

General:

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

Title: Estimation of Organic Matter Content in Coastal Soil Using Reflectance Spectroscopy Research

Author: ZHENG Guanghui1, Dongryeol RYU2,∗, JIAO Caixia1 and HONG Changqiao1

Year: 2015

Description:

Rapid determination of soil organic matter (SOM) using regression models based on soil reflectance spectral data serves an important function in precision agriculture. “Deviation of arch” (DOA)-based regression and partial least squares regression (PLSR) are two modeling approaches to predict SOM. However, few studies have explored the accuracy of the DOA-based regression and PLSR models. Therefore, the DOA-based regression and PLSR were applied to the visible near-infrared (VNIR) spectra to estimate SOM content in the case of various dataset divisions. A two-fold cross-validation scheme was adopted and repeated 10000 times for rigorous evaluation of the DOA-based models in comparison with the widely used PLSR model. Soil samples were collected for SOM analysis in the coastal area of northern Jiangsu Province, China. The results indicated that both modelling methods provided reasonable estimation of SOM, with PLSR outperforming DOA-based regression in general. However, the performance of PLSR for the validation dataset decreased more noticeably. Among the four DOA-based regression models, a linear model provided the best estimation of SOM and a cut off of SOM content (19.76 g kg−1), and the performance for calibration and validation datasets was consistent. As the SOM content exceeded 19.76 g kg−1, SOM became more effective in masking the spectral features of other soil properties to a certain extent. This work confirmed that reflectance spectroscopy combined with PLSR could serve as a non-destructive and cost-efficient way for rapid determination of SOM when hyper spectral data were available. The DOA-based model, which requires only 3 bands in the visible spectra, also provided SOM estimation with acceptable accuracy.

Title: Preliminary Study of Soil Available Nutrient Simulation Using a Modified WOFOST Model and Time-Series Remote Sensing Observations

Author: Zhiqiang Cheng 1,2 ID ,Jihua Meng 1,*, YanyouQiao 1, Yiming Wang 1,2, Wenquan Dong 1 and Yanxin Han 1,2

Year: 2017

Description:

The approach of using multispectral remote sensing (RS) to estimate soil available nutrients (SANs) has been recently developed and shows promising results. This method overcomes the limitations of commonly used methods by building a statistical model that connects RS-based crop growth and nutrient content. However, the stability and accuracy of this model require improvement. In this article, we replaced the statistical model by integrating the World Food Studies (WOFOST) model and time series of remote sensing (T-RS) observations to ensure stability and accuracy. Time series of HJ-1 A/B data was assimilated into the WOFOST model to extrapolate crop growth simulations from a single point to a large area using a specific assimilation method. Because nutrient-limited growth within the growing season is required and the SAN parameters can only be used at the end of the growing season in the original model, the WOFOST model was modified. Notably, the calculation order was changed, and new soil nutrient uptake algorithms were implemented in the model for nutrient-limited growth estimation. Finally, experiments were conducted in the spring maize plots of Hongxing Farm to analyze the effects of nutrient stress on crop growth and the SAN simulation accuracy. The results confirm the differences in crop growth status caused by a lack of soil nutrients. The new approach can take advantage of these differences to provide better SAN estimates. In general, the new approach can overcome the limitations of existing methods and simulate the SAN status with reliable accuracy.

Title: Distinguishing Heavy-Metal Stress Levels in Rice Using Synthetic Spectral Index Responses to Physiological Function Variations

Author: Ming Jin, Xiangnan Liu, Ling Wu, and Meiling Liu

Year: 2016

Description:

Accurately assessing the heavy-metal contamination in crops is crucial to food security. This study provides a method to distinguish heavy-metal stress levels in rice using the vari- ations of two physiological functions as discrimination indices, which are obtained by assimilation of remotely sensed data with a crop growth model. Two stress indices, which correspond to dailytotalCO2 assimilationanddry-matterconversioncoefficient were incorporated into the World Food Study (WOFOST) crop growth model and calculated by assimilating the model with leaf area index (LAI), which was derived from time-series HJ1- CCD data. The stress levels are not constant with rice growth; thus, to improve the reliability, the two stress indices were obtained at both the first and the latter half periods of rice growth. To com- pare the stress indices of different stress levels, a synthetic stress index was established by combining the two indices; then, three types of stress index discriminant spaces based on the synthetic index of different growth periods were constructed, in which the two-dimensional discriminant space based on two growth periods showed the highest accuracy, with a misjudgment rate of 4.5%. When the discrimination rules were applied at a regional scale, the average correct discrimination rate was 95.0%.

2.1 Existing System:

It present a crop/weeds classification approach based on a three-steps procedure. The first step is a robust pixel-wise segmentation (i.e., soil/plant) and image patches containing plants are extracted in the second step. The third step, a deep CNN for crop/weed classification is used. The extracted blobs in the masked image containing plants information are fed to a CNN classifier based on a fine-tuned model of VGG-16 exploiting the ability of deep CNN in object classification and to reduce the limitations of CNNs in generalizing when a limited amount of data is available. The classification step can then be specialized to the types of plants needed by the application scenario. It evaluated the complete pipeline, including the first background removal phase and the subsequent classification stage. Experimental results demonstrate that can achieve good classification results on challenging data.

Precision agriculture is gaining increasing attention because of the possible reduction of agricultural inputs (e.g., fertilizers and pesticides) that can be obtained by using high-tech equipment, including robots. To focus on an agricultural robotics system that addresses the

We eding problem by means of selective spraying or mechanical removal of the detected weeds. To describe a deep learning based method to allow a robot to perform an accurate weed/crop classification using a sequence of two Convolutional Neural Networks (CNNs) applied to RGB images. The first network, based on encoder-decoder segmentation architecture, performs a pixel wise, plant-type agnostic, segmentation between vegetation and soil that enables to extract a set of connected blobs representing plant instances.

Drawbacks:

- It can't determine to improve the classification accuracy of our pipeline.
- Connecting the bridge manually and some corruption are happened.
- Private sectors domination high, profit low and credits not getting concern farmer.

2.2 REFERENCES

P.Priya, U.MuthaiahM.Balamurugan . Predicting yield of the crop using machinelearning algorithm. International Journal of Engineering Research.

J.Jeong, J.Resop , N.Mueller and team . Random forests for global and regional cropyield prediction.PLoS ONE Journal.

Narayanan Balkrishnan and Dr. Govindarajan Muthukumarasamy . Crop production Ensemble Machine Learning model for prediction. International Journal of ComputerScience and Software Engineering (IJCSSE).

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

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

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

2.3 Problem Statement Definition

Analytics is the interpretation of data pattern that assist decision- making and performance improvement. Agriculture Data analytics in crop yield helps in analysing some important visualization, creating a dashboard and by going through these we will get most of the insights of Crop production in India. IBM Cognos Analytics integrates reporting, modelling, analysis, exploration, dashboards, stories, and event management so we can understand our organization\'s data, and make effective decisions. A dashboard helps us to monitor events or activities at a glance by providing key insights and analysis about our data on one or more pages or screens. In this project, we visualize, analyse and gain most of the insights by creating a dashboard.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas :

An empathy map is a collaborative visualization used to articulate what we know about a particular type of user. It externalizes knowledge about users in order to 1) create a shared understanding of user needs, and 2) aid in decision making

.

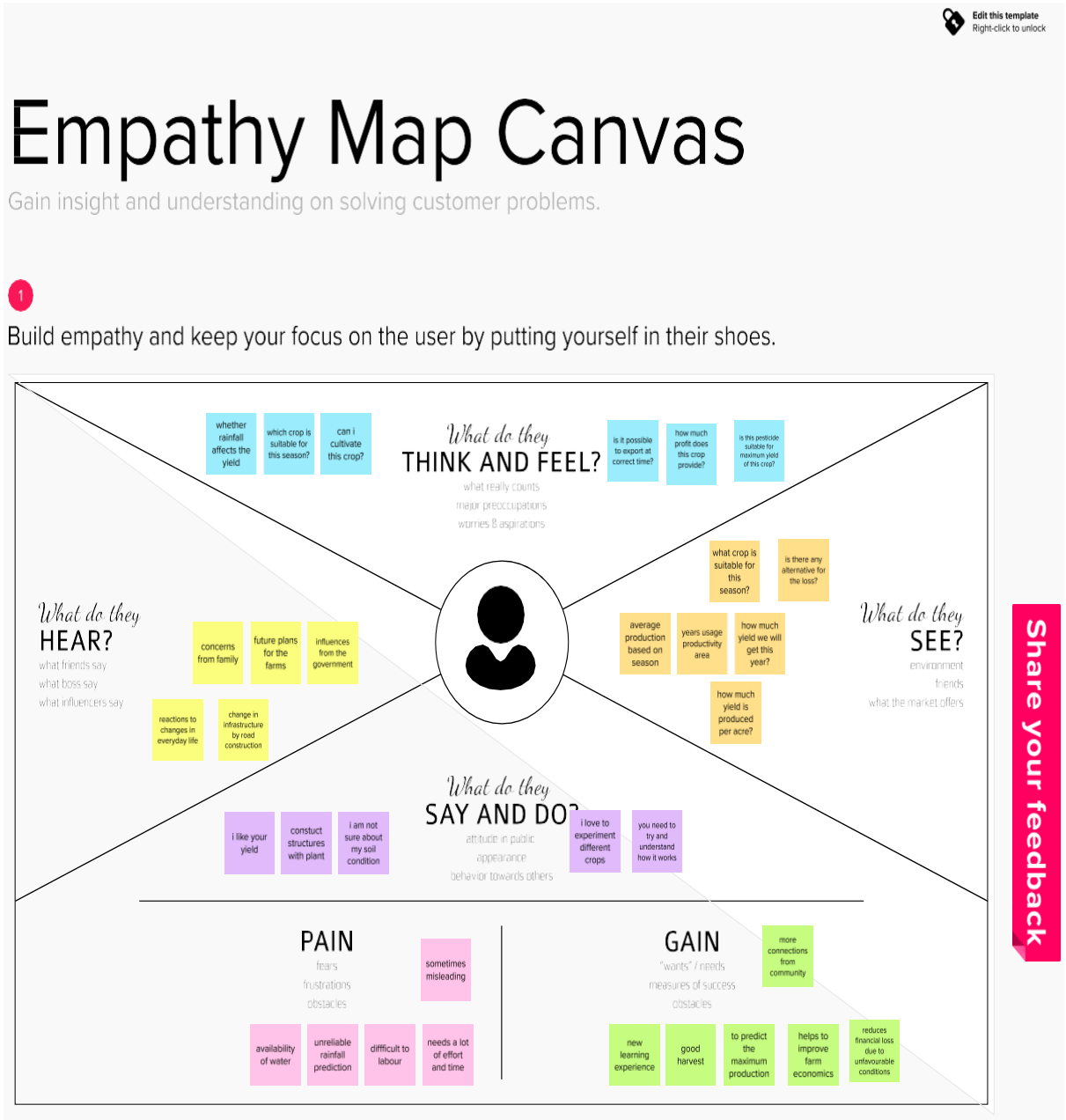


Fig 3.1.1 Empathy Map Canvas

3.2 Ideation and brainstorming

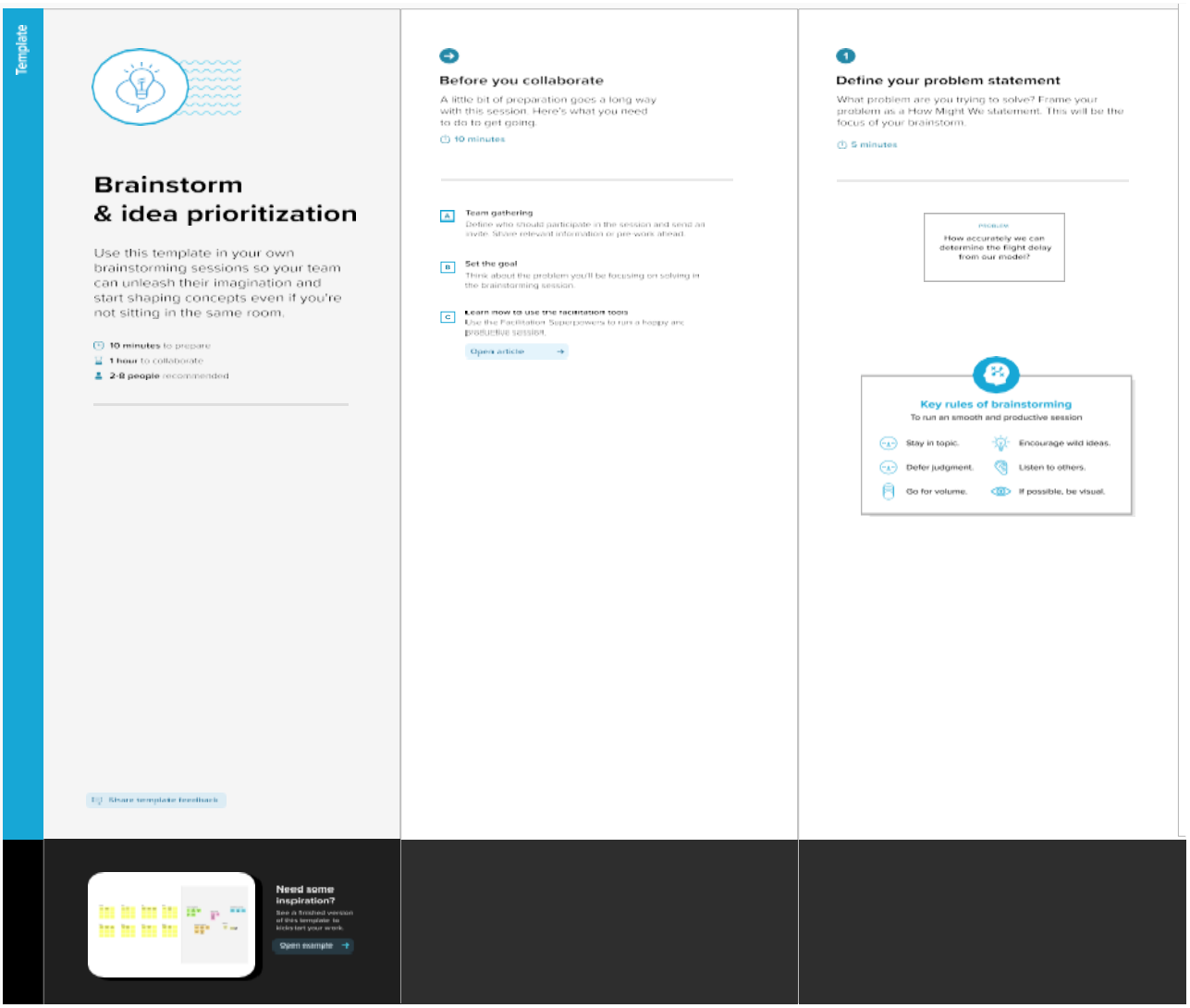


Fig 3.2.1 Ideation & Brainstorming

3.3 Proposed System:

Agriculture is the backbone of Indian Economy. In India, majority of the farmers are not getting the expected crop yield due to several reasons. The agricultural yield is primarily depending on weather conditions. Rainfall conditions also influences the rice cultivation. In this context, the farmers necessarily require a 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 in their crops. As per this project we will be analyzing some important visualization, creating a dashboard and by going through these we will get most of the insights of Crop production in India.

Advantages:

- Our goal is push for assisting farmers, government using our predictions. All these publications state they have done better than their competitors but there is no article or

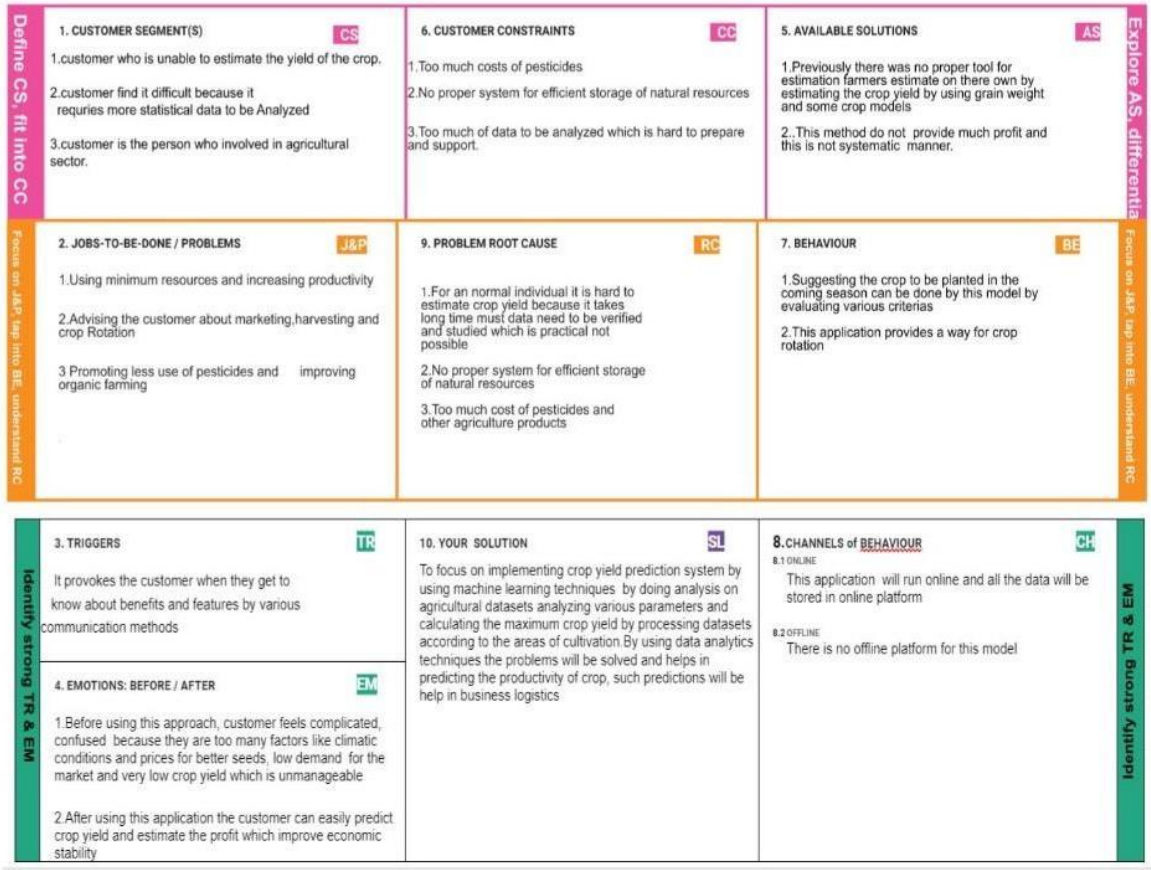
public mention of their work being used practically to assist the farmers. If there are some genuine problems in rolling out that work to next stage, then identify those problems and try solving them.

- It is targeted to those farmers who wish to professionally manage their farm by planning, monitoring and analyzing all farming activities.

Application:

- It is an integrated farm management application using mobile app.
- Agricultural sector to automate to identify the crop prediction process (real time world) and predicting by desktop application / web application.

3.4 Problem Solution Fit



4. REQUIREMENT ANALYSIS

General:

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements
2. Non-Functional requirements
3. Environment requirements
    A. Hardware requirements
    B. software requirements

4.1 Functional requirements:

The functional requirements of this system is the user should login to the IBM cognos analytics platform and then select the crop name to search in the dashboard.

4.2 Non-Functional Requirements:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

Environmental Requirements:

1. Software Requirements:

    Operating System : Windows

    Tool : IBM Cognos Analytics

2. Hardware requirements:

    Processor : Pentium IV/III

    Hard disk : minimum 80 GB

RAM : minimum 2 GB

5.project design

5.1 Data flow Diagram

Overview of the system:

This helps all others department to carried out other formalities. It have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code. The following Involvement steps are,

- Define a problem
- Preparing data
- Evaluating algorithms
- Improving results
- Predicting results

Project Goals:

- Exploration data analysis of variable identification
  - Loading the given dataset
  - Import required libraries packages
  - Analyze the general properties
  - Find duplicate and missing values
  - Checking unique and count values
- Uni-variate data analysis
  - Rename, add data and drop the data
  - To specify data type
- Exploration data analysis of bi-variate and multi-variate
  - Plot diagram of pairplot, heatmap, bar chart and Histogram
- Method of Outlier detection with feature engineering
  - Pre-processing the given dataset
  - Splitting the test and training dataset
  - Predicting on the accuracy

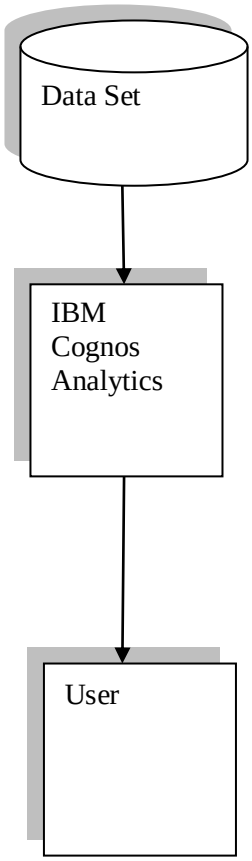
- Analyze the general properties
- Find duplicate and missing values
- Checking unique and count values

➢ Uni-variate data analysis
- Rename, add data and drop the data
- To specify data type

➢ Exploration data analysis of bi-variate and multi-variate
- Plot diagram of pairplot, heatmap, bar chart and Histogram

➢ Method of Outlier detection with feature engineering
- Pre-processing the given dataset
- Splitting the test and training dataset
- Predicting on the accuracy

Work flow diagram



5.2 Solution & Technical Architecture

IBM Cognos Analytics provides dashboards and stories to communicate your insights and analysis. You can assemble a view that contains visualizations such as a graph, chart, plot, table, map, or any other visual representation of data. Explore powerful visualizations of your data in IBM Cognos Analytics and discover patterns and relationships that impact your business. A dashboard helps you to monitor events or activities at a glance by providing key insights and analysis about your data on one or more pages or screens.

system architecture: -

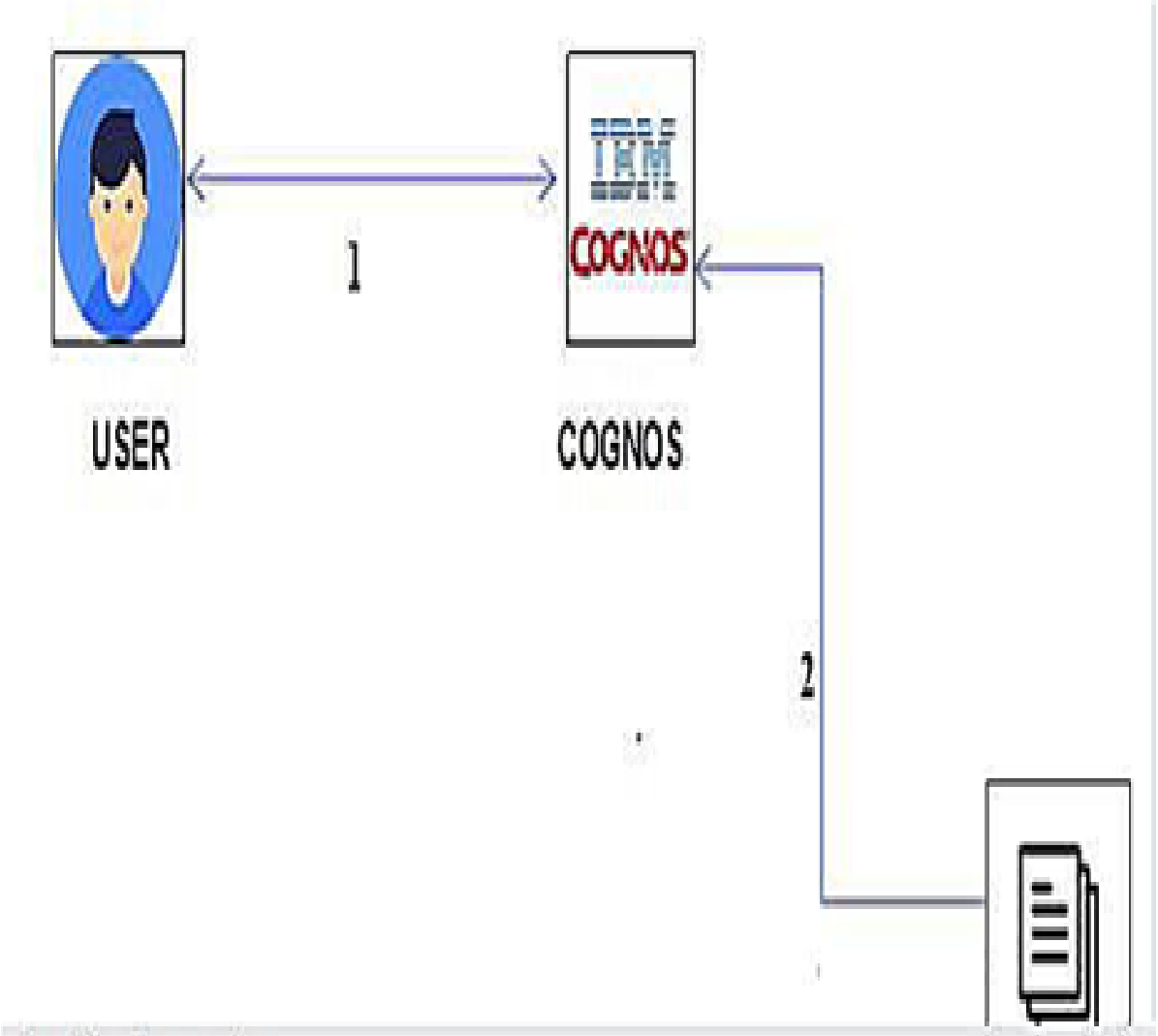
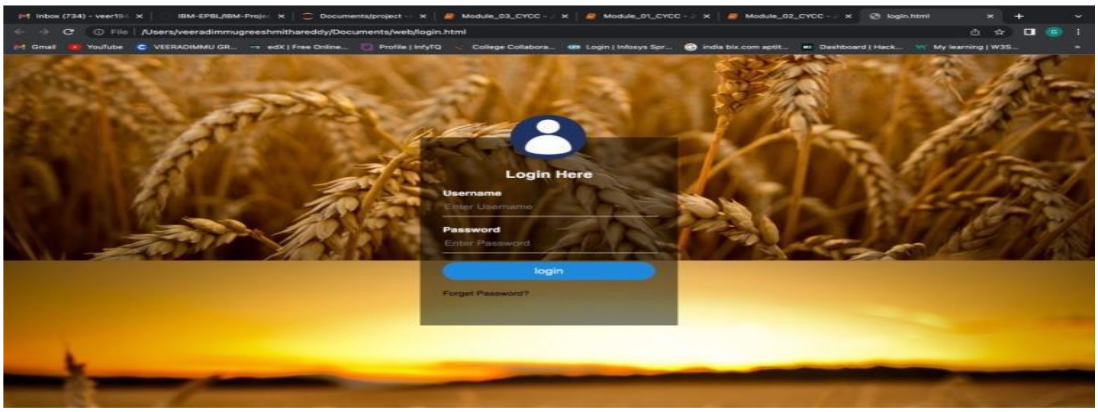
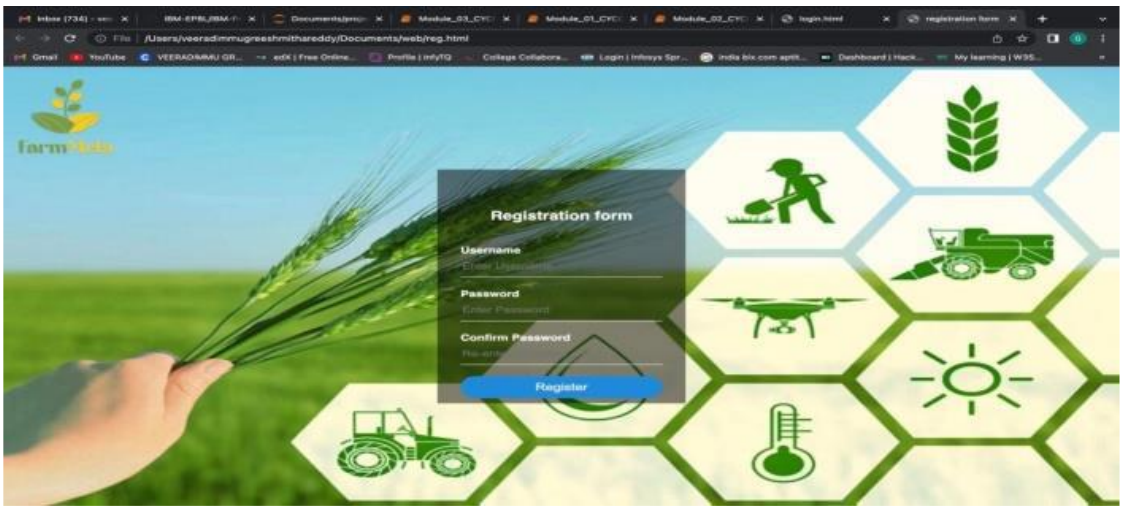


Fig. 1 Architecture Diagram

5.3 User stories

USN-1





6.Project Planning and Scheduling

6.1 Sprint Planning and Estimation

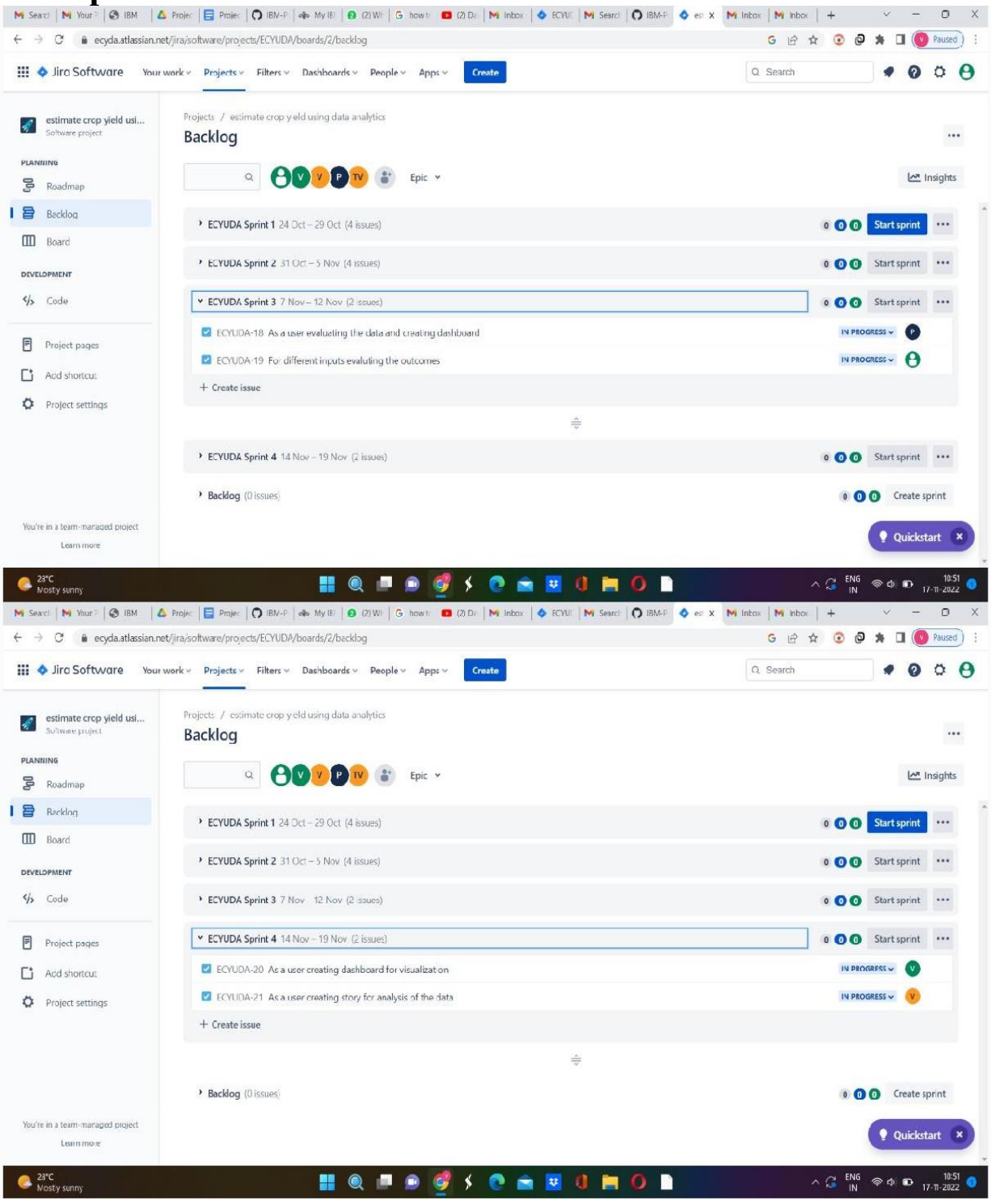
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	**2**	high	JANSIRANI R
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	**1**	high	JANANI P
Sprint-1		USN-3	As a user, I can register for the application through G-Mail or any kind of Mail.	**2**	medium	Maheshwari R
Sprint-1	Login	USN-4	As a user, I can log into the application by entering email & password	**1**	High	Ishwarya S
Sprint-1	Analysis and Estimation(Working and Loading the dataset)	USN-5	As a user, I can view the resource i.e., dataset that is being uploaded or loaded in a platform called IBM Cognos Analytics with Watson Services to view and analyze the data.	**14**	High	Heera Y
Sprint-2	Analysis and Estimation(Data Visualization Charts)	USN-6	As a user, I can visualise the data of crop production to know the insights Where Average Crop Production by Seasons, the Yearly usage of Area in Crop Production, top 10 States in Crop Yield	**20**	High	Janani P

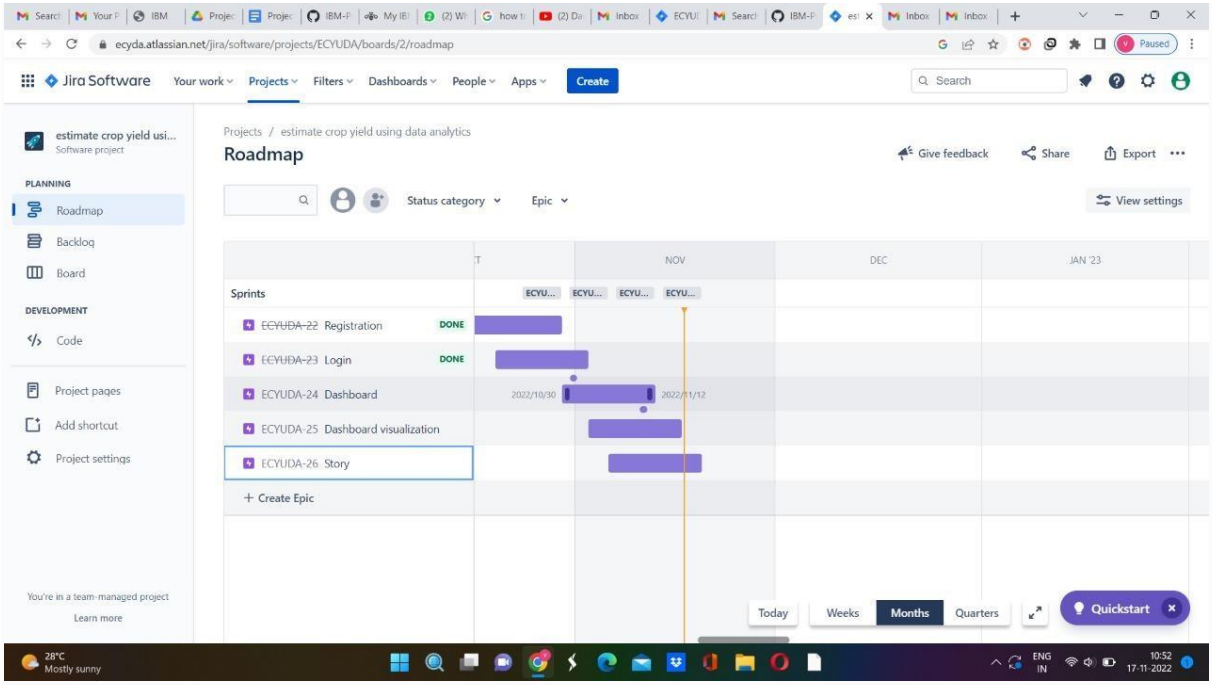
			Production by Area, the Crop Production by State and the Sates with Seasonal Crop Production can be known			
Sprint-3	Dashboard	USN-7	As a User , I can use Cognos Analytics with Watson Services, An interactive dashboard must be created and viewed	**20**	high	Heera Y
Sprint-4	Analysis and Estimation(Exportation /Export The Analytics)	USN-8	As a user, I can view the dashboard and visualization of crop production that is being exported either through email/link/pdf.			JANANI P

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	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

6.3 Reports from JIRA





7. CODING & SOLUTIONING

Modules:

1. Uploading data.(dataset)
2. Cleaning data (prepare data).
3. Analysing and interpreting (exploration).
4. Visualizing data (dashboard creation).

Module-01:

Uploading data(dataset)

In this project we have uploaded crop_production dataset.

Data Pre-processing*:*

The dataset consists of attributes Moisture, rainfall, Average, Humidity, Mean Temp, max Temp, Min temp, alkaline, sandy, chalky, clay, millet, yield, Outcomes. We will be using the .csv to perform the pre-processing.

	State_Name	District_Name	Crop_Year	Season	Crop	Area	rainfall	Average Humidity	Mean Temp	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)	cost of production per yield
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	0.012360	57	62	23076.74	1941.55	9.83	19085.4365
1	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Arecanut	1254.0	0.084119	56	58	12610.85	1691.66	6.83	11554.0378
2	Andaman and Nicobar Islands	NICOBARS	2002	Whole Year	Arecanut	1258.0	0.080064	58	53	32683.46	3207.35	9.33	29924.5755
3	Andaman and Nicobar Islands	NICOBARS	2003	Whole Year	Arecanut	1261.0	0.181051	57	58	13209.32	2228.97	5.90	13150.9230
4	Andaman and Nicobar Islands	NICOBARS	2004	Whole Year	Arecanut	1264.7	0.035446	63	67	22560.30	1595.56	13.57	21651.7492

Fig: Given dataframe

Data Validation/ Cleaning/Preparing Process:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

```
#preprocessing, split test and dataset, split
X = df.drop(labels='CPPY', axis=1)
#Response variable
y = df.loc[:,'CPPY']
```

```
#We'll use a test size of 30%. We also stratif
from sklearn.model_selection import train_test
X_train, X_test, y_train, y_test = train_test_
print("Number of training dataset: ", len(X_tr
print("Number of testing dataset: ", len(X_tes
print("Total number of dataset: ", len(X_train
```

```
Number of training dataset:  163704
Number of testing dataset:  70160
Total number of dataset:  233864
```

Fig: Spliting the given dataset

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set.

	State_Name	District_Name	Crop_Year	Season	Crop	Area	R	H	T	CC	CP	Y	CPPY
0	0	410	3	1	2	2025	33	45	10	21	30	13	126
1	0	410	4	1	2	2025	121	44	6	3	26	7	72
2	0	410	5	4	2	2030	118	46	1	33	45	11	200
3	0	410	6	4	2	2033	172	45	6	4	36	4	78
4	0	410	7	4	2	2037	75	51	15	20	24	23	144

Fig: After preprocessing given data frame

Module-02:

Exploration data analysis of visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

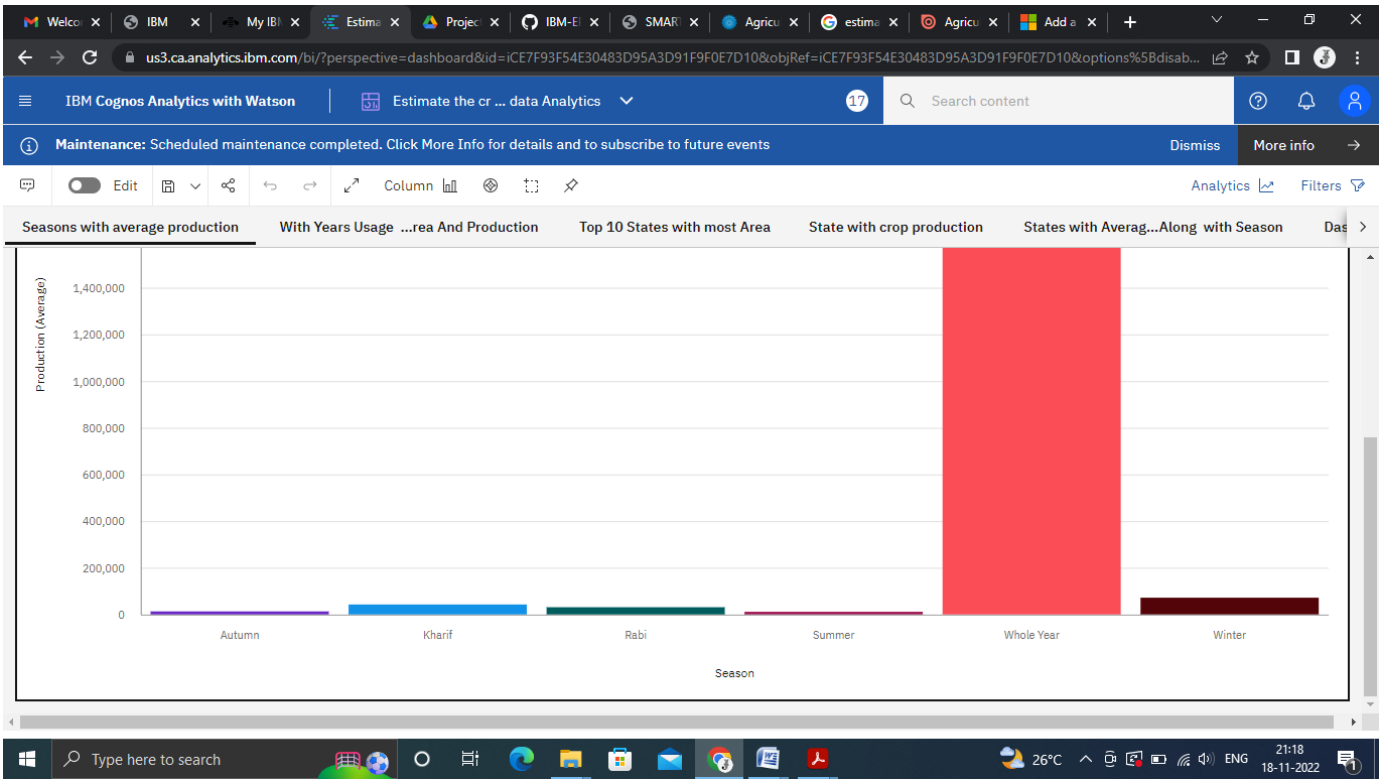
Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

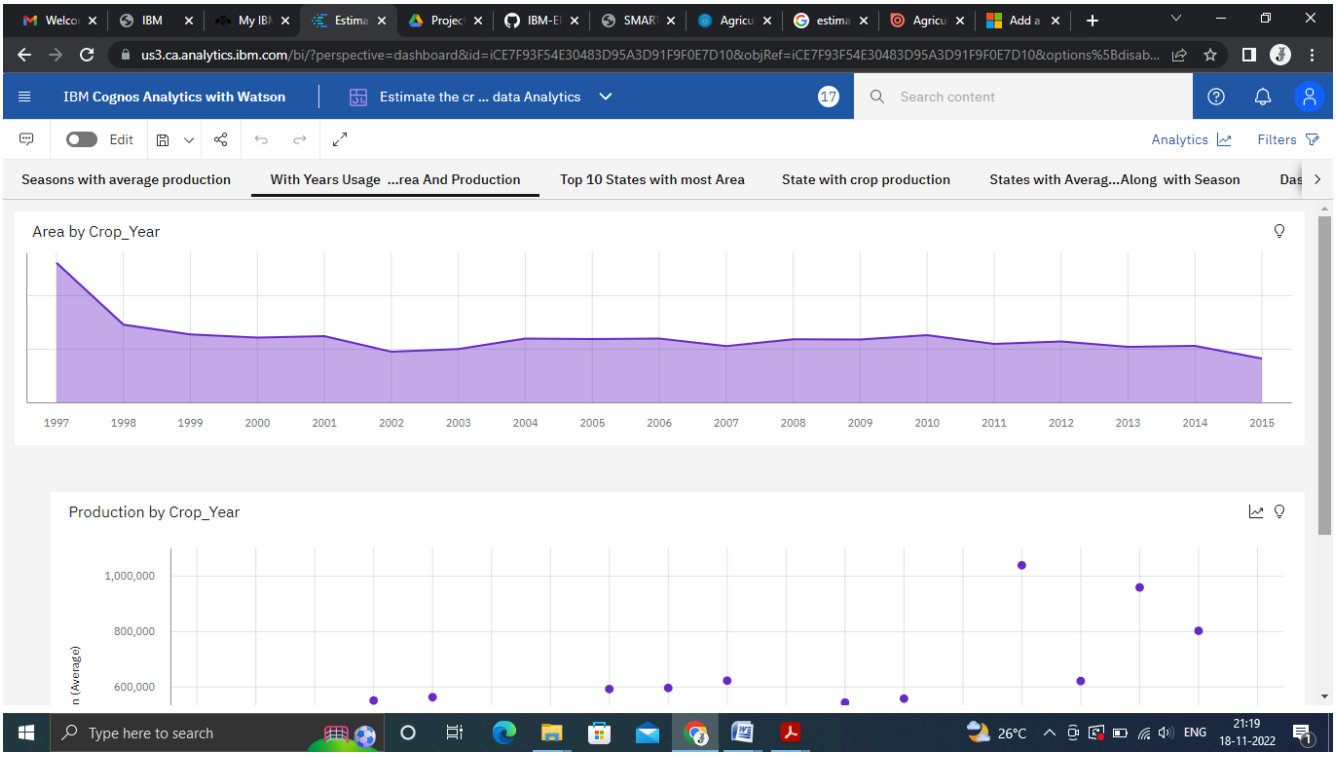
creation of data visualization charts

- At the bottom of the workbook, click the “New Dashboard” icon.
- From the sheets list at left, drag views to the dashboard at right onto the Dashboard Workspace in the indicated location
- Change the target size of the dashboard by making a selection from the size drop- down list in the Dashboard section of the navigation menu on the left and adjust the object sizes accordingly.
- Remove unnecessary filters from the dashboard and make the essential filters as floating type and arrange them accordingly.
- Click on the first object on the Dashboard and click on “Use as Filter”.

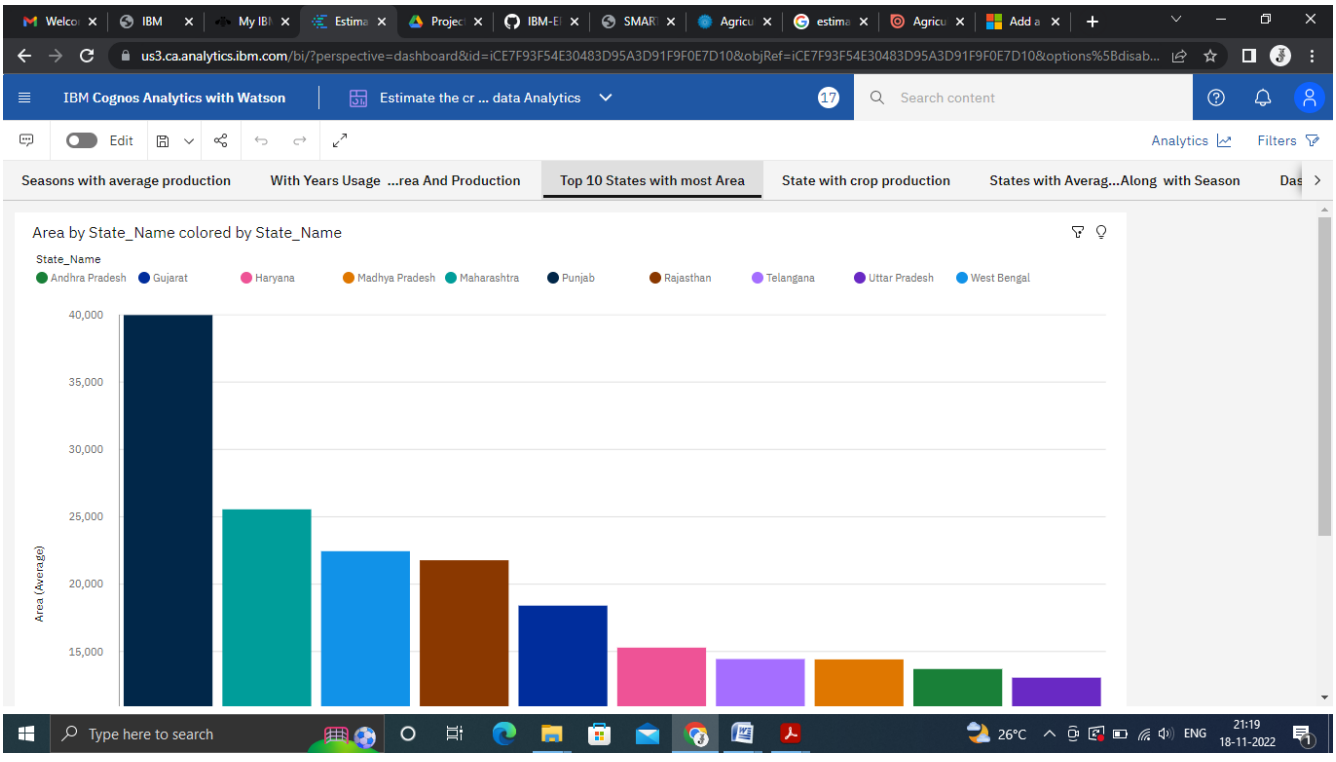
1. Build a Visualization to showcase Average Crop Production by Seasons.



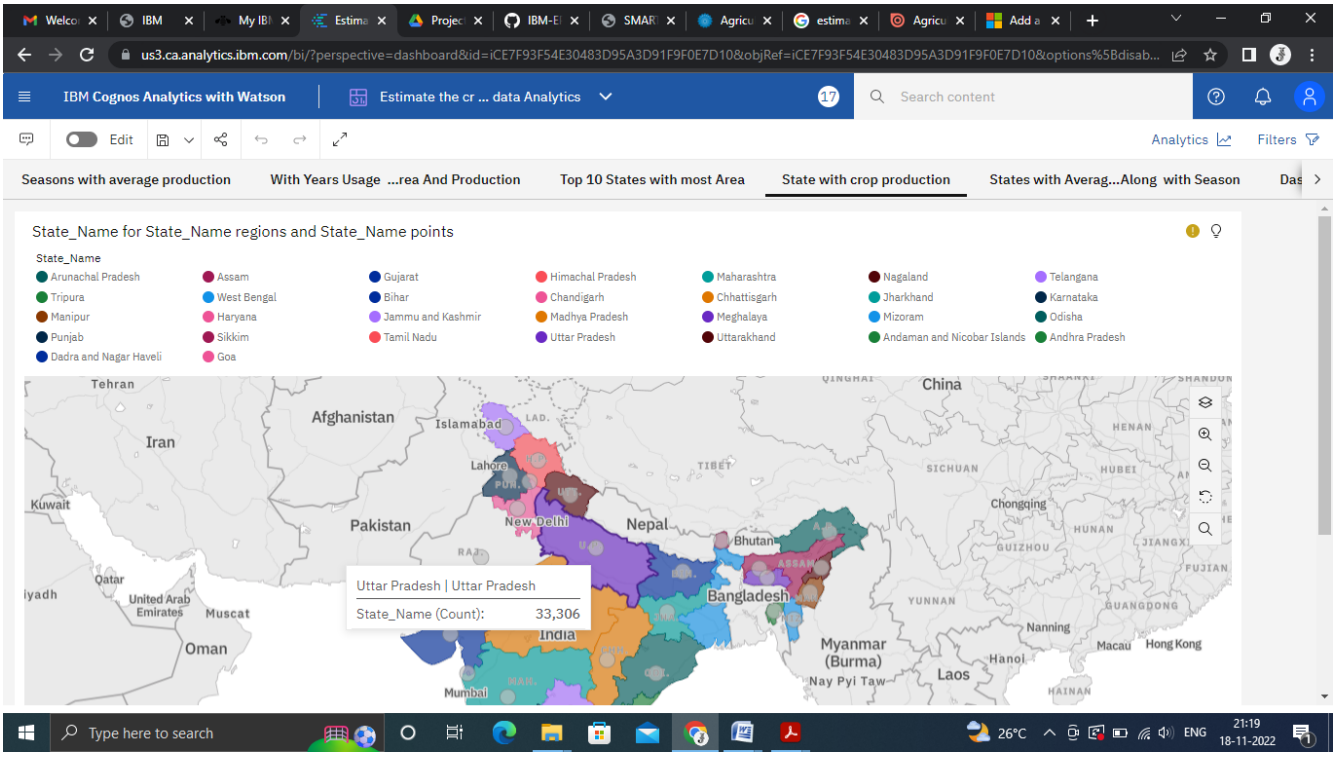
2.Show case the Yearly usage of Area in Crop Production.



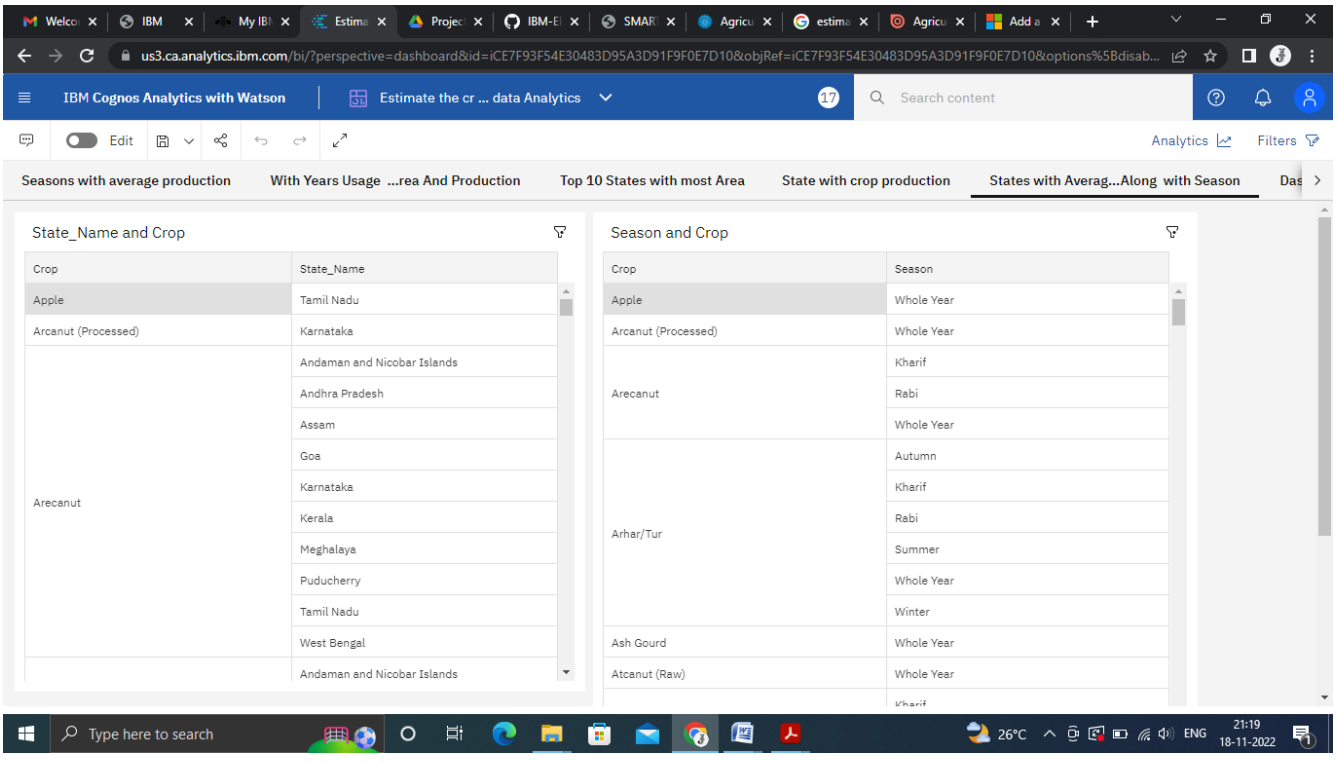
3.Build a visualization to show case top 10 States in Crop Yield Production by Area



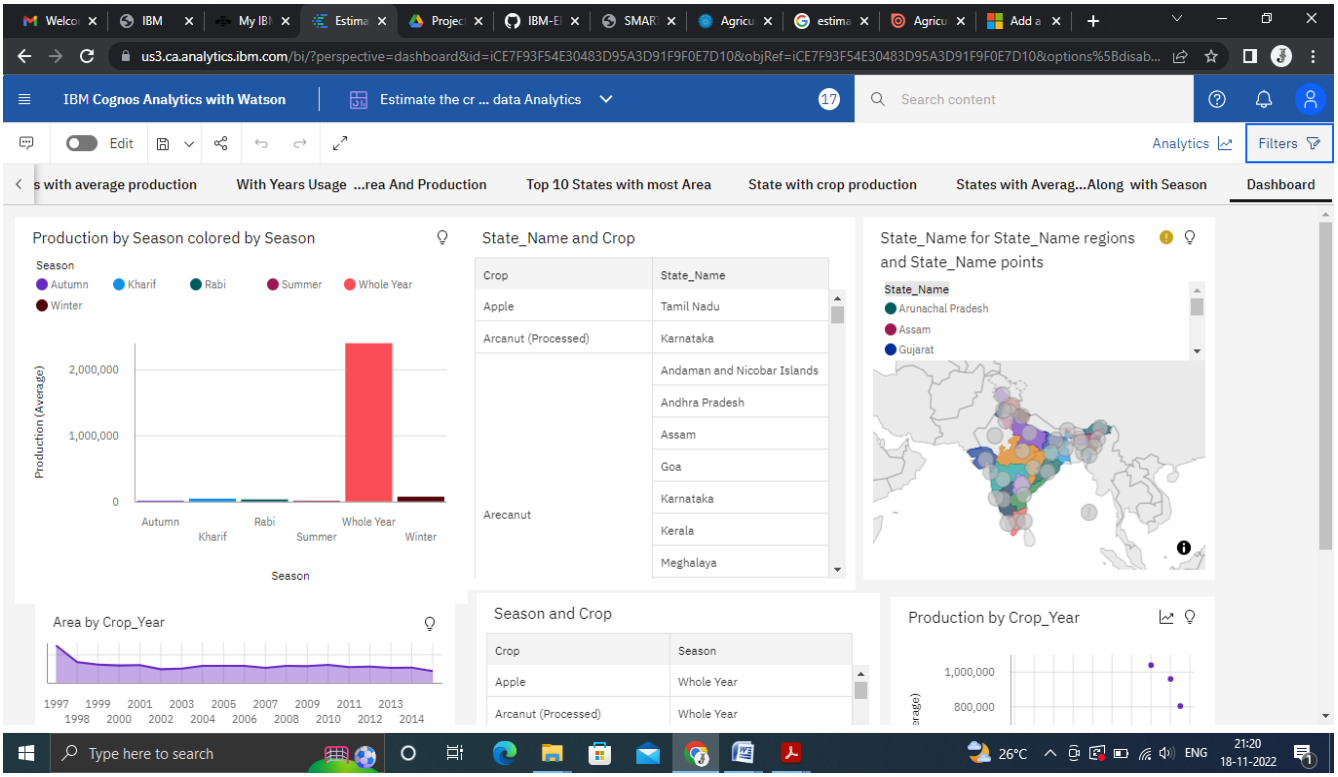
4.Build the required Visualization to showcase the Crop Production by State.



5.Build Visual analytics to represent the Sates with Seasonal Crop Production using a Text representation.



Dashboard creation



8. TESTING

Test Cases

Testing Levels:-

All major activities of various testing level are described below.

1.Unit Testing

2.Integration Testing

3.Functional Testing

4.System Testing

5.White box Testing

6. Black Box Testing

1. Unit Testing:-

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive.

2. Integration Testing:-

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

3. Functional Testing:-

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

4. System Testing:

System testing ensures that the entire integrated software systemmeets requirements. It tests a configuration to ensure known and predictable results.An example of system testing is the configuration oriented system integration test.System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

5.White Box Testing:

White Box Testing is a testing in which in which the softwaretester has knowledge of the inner workings, structure and language of the software, orat least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

6.User Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

9.RESULTS

9.1 Performance Metrics

Sensitivity:

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall. This implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (and, thus, could also be termed as the false negative). This can also be represented in the form of a false negative rate. The sum of sensitivity and false negative rate would be 1. Let's try and understand this with the model used for predicting whether a person is suffering from the disease. Sensitivity is a measure of the proportion of people suffering from the disease who got predicted correctly as the ones suffering from the disease. In other words, the person who is unhealthy actually got predicted as unhealthy.

Mathematically, sensitivity can be calculated as the following:

Sensitivity = (True Positive) / (True Positive + False Negative)

The following is the details in relation to True Positive and False Negative used in the above equation.

- True Positive = Persons predicted as suffering from the disease (or unhealthy) are actually suffering from the disease (unhealthy); In other words, the true positive represents the number of persons who are unhealthy and are predicted as unhealthy.
- False Negative = Persons who are actually suffering from the disease (or unhealthy) are actually predicted to be not suffering from the disease (healthy). In other words, the false negative represents the number of persons who are unhealthy and got predicted as healthy. Ideally, we would seek the model to have low false negatives as it might prove to be life-threatening or business threatening.

The higher value of sensitivity would mean higher value of true positive and lower value of false negative. The lower value of sensitivity would mean lower value of true positive and higher value of false negative. For healthcare and financial domain, models with high sensitivity will be desired.

Specificity:

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negative, which got predicted as positive and could be termed as false positives. This proportion could also be called a false positive rate. The sum of specificity and false positive rate would always be 1. Let's try and understand this with the model used for predicting whether a person is suffering from the disease. Specificity is a measure of the proportion of people not suffering from the disease who got predicted correctly as the ones who are not suffering from the disease. In other words, the person who is healthy actually got predicted as healthy is specificity.

Mathematically, specificity can be calculated as the following:

Specificity = (True Negative) / (True Negative + False Positive)

The following is the details in relation to True Negative and False Positive used in the above equation.

- True Negative = Persons predicted as not suffering from the disease (or healthy) are actually found to be not suffering from the disease (healthy); In other words, the true negative represents the number of persons who are healthy and are predicted as healthy.
- False Positive = Persons predicted as suffering from the disease (or unhealthy) are actually found to be not suffering from the disease (healthy). In other words, the false positive represents the number of persons who are healthy and got predicted as unhealthy.

The higher value of specificity would mean higher value of true negative and lower false positive rate. The lower value of specificity would mean lower value of true negative and higher value of false positive.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Stories the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

10.ADVANTAGES & DISADVANTAGES

Advantages:

- Our goal is push for assisting farmers, government using our predictions. All these publications state they have done better than their competitors but there is no article or public mention of their work being used practically to assist the farmers. If there are some genuine problems in rolling out that work to next stage, then identify those problems and try solving them.

- It is targeted to those farmers who wish to professionally manage their farm by planning, monitoring and analyzing all farming activities.

- Achieving the maximum crop at minimum yield is the ultimate Aim of the project.

- Early detection of problems and management of that problems can help the farmers for better crop yield.

- For the better understanding of the crop yield, we need to study of the huge data with the help of machine learning algorithm so it will give the accurate yield for that crop and suggest the farmer for a better crop.

Disadvantages:

- The obtained result for the crop yield prediction using SMO classifier gives less accuracy when compared to naïve Bayes, multilayer perceptron and Bayesian network.

- Previously yield is predicted on the bases of the farmers prior experience but now weather conditions may change drastically so they cannot guess the yield.

11.CONCLUSION:

As a result of penetration of technology into agriculture field, there is a marginal improvement in the productivity. The innovations have led to new concepts like digital agriculture, smart farming, precision agriculture etc. In the literature, it has been observed that analysis has been done on agriculture productivity, hidden patterns discovery using data set related to seasons and crop yields data. We have noticed and made analysis about different crops cultivated, area and productions in different states and districts using IBM Cognos some of them are 1) Seasons with average productions. In this analytics we come to know in which seasons the average production is more and in which seasons the production is less. 2) Production by crop year. In this analysis we come to know in which years the production is high and low. 3) Production by District. With this analytics we can aware of the districts with the selected crops cultivated and states too.4) Production by Area. From this we can know how much area should be cultivated and the production will be getting will be estimated. Finally created the dashboard and made analysis that in which state and in which year with crop area and to what extent the production will be are analysed.

12.FUTURE SCOPE:

- Remaining SMLT algorithms will be involve to finding the best accuracy with applying to predict the crop yield and cost.
- Agricultural department wants to automate the detecting the yield crops from eligibility process (real time).
- To automate this process by show the prediction result in web application or desktop application.
- To optimize the work to implement in Artificial Intelligence environment.

13.APPENDIX

GitHub Repo Link: https://github.com/IBM-EPBL/IBM-Project-6041-1658822521.git

Demo Link:..\Videos\Captures\MAHENDRA COLLEGE OF ENGINEERING - Google Chrome 2022-11-19 07-17-05.mp4