ESTIMATE THE CROP YIELD USING DATA ANALYTICS

A PROJECT REPORT

Submitted by

TEAM ID: PNT2022TMID42593

PERABATHULA VENKATA SITA MAHA LAKSHMI	**(711119104041)**
PRADEEP C J	**(711119104042)**
PRANESHWAR K	**(711119104043)**
RAGUL J	**(711119104045)**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

JANSONS INSTITUTE OF TECHNOLOGY, COIMBATORE

ANNA UNIVERSITY: CHENNAI 600 025

ABSTRACT

Agriculture is important for human survival because it serves the basic need. A well-known fact that the majority of population (≥55%) in India is into agriculture. Due to variations in climatic conditions, there exist bottlenecks for increasing the crop production in India. It has become challenging task to achieve desired targets in Agri based crop yield. Various factors are to be considered which have direct impact on the production, productivity of the crops. Crop yield prediction is one of the important factors in agriculture practices. Farmers need information regarding crop yield before sowing seeds in their fields to achieve enhanced crop yield. The use of technology in agriculture has increased in recent year and data analytics is one such trend that has penetrated into the agriculture field. The main challengein using big data in agriculture is identification of effectiveness of big data analytics. Efforts are going on to understand how big data analytics can agriculture productivity. The present study gives insights on various data analytics methods applied to crop yield prediction and also signifies the important lacunae points in the proposed area of research.

1. INTRODUCTION

1.1 PROJECT OVERVIEW:

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 desirablecrop according to the existing conditions, support to achieve the higher and enhanced crop yield [11]. The accurate prediction of crop yield certainly benefits the farmers in choosing the right method to reduce the crop damage and gets best prices for their crops.

1.2 PURPOSE:

Agriculture is the widest economic sector and has an important role regarding the frameworkof socio-economic fabric of India. Farming depends on various factors like climate and economic factors like temperature, irrigation, cultivation, soil, rain fall, pesticide and fertilizers. Historical information regarding crop yield provides major input for companies engaged in this domain. The estimation of production of crop helps these companies in planning supply chain decision like production scheduling. The industries such as fertilizers, seed, agrochemicals and agricultural machinery plan production and activities like marketing based on the estimates of crop yield. Farmers experience was the only way for prediction of crop yield in the past days. Technology penetration into agriculture field has led to automation of the activities like yield estimation, crop health monitoring etc.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM:

A) P. VINDHYA "CROP YIELD PREDICTION USING BIG DATA ANALYTICS" ANNA UNIVERSITY, TRICHY, TAMIL NADU, INDIA, 5 MAY 2015.

The proposed system suggests the accurate prediction of crop yield certainly benefits the farmers in choosing the right method to reduce the crop damage and get best prices for their crops. The factors involved in this method are Area under Cultivation (AUC) interims of hectors, Annual Rainfall (AR) rates and Food Price Index (FPI) and to develop relationships among these parameters. Regression Analysis (RA) methodology was applied to examine the selected factors and their impact on crop prediction and final yield. RA methodology is a multivariable investigation practice which can categorize the factors into groups such as explanatory and response variables and helps to assess their interaction to obtain a resolution. Crop yield gaps, measured as difference between expected yields based on the potency and actual farm yield received. In order to achieve the higher crop yield, farmers must tackle the influencing factors such as influence of change in climate conditions on the prospects of crop yields, and change in the usage of agricultural land to assess and ultimately reduce the crop yield gaps. Several researchers reported the applications of bio simulation models to estimate the crop yield gaps in the last decade. The critical challenge remaining with these methods is scaling up of these approaches to assess the data collated between different time intervals from the broader geographical regions.

B) M. A. JAYARAM AND NETRA MARAD, "FUZZY INFERENCE SYSTEM FOR CROP PREDICTION", JOURNAL OF INTELLIGENT SYSTEMS, 2012.

The proposed system suggests an attempt to develop fuzzy inference systems for crop yield prediction. Physio morphological features of Sorghum were considered. A huge database (around 1000 records) of physio morphological features such as days of 50 percent showering, dead heart percentage, plant height, panicle length, panicle weight and number of primaries and the corresponding yield were considered for the development of the model. In order to

and out the sensitivity of parameters, one-to-one, two-to-one and three-to-one combinations of input and output were considered. The results have clearly shown that panicle length contributes forth yield as the lone parameter with almost one-to-one matching between predicted yield and actual value while panicle length and panicle weight in combination seemed to play a decisive role in contributing for the yield with the prediction accuracy rejected by very low RMS value. In hybrid plants, the morphological features such as plant height, panicle length, panicle weight, number of primaries and length of the leaves cannot be determined or predicted accurately. Therefore, sometimes this becomes a failure model.

C) A. D. BOSE, ''BIG DATA ANALYTICS IN AGRICULTURE''.

The proposed system suggests how Big Data Analytics combined with various structured and unstructured data helps in providing insight to farmers to make a decision as to which crops to grow and reduce losses due to unexpected or unpredictable disasters. In Section I the paper states that we can collect the data produced by sensors from the official databases that are usually maintained and governed by institutions. Here the author suggests we can collect and analyze the data in different stages in agriculture and see their influence in the big picture. It is dependent on two major factors, the push and pull factor. Visualization of agricultural data is done to simplify the complex, structured, and unstructured data. Interpretation of data can be done using methods like overviews, verifiable models, or in an Ad-Hoc manner graphs. the implementation of analytic techniques in agriculture had been discussed. The first method is an Intelligent crop recommendation system that considers all the factors such as soil conditions, temperature, rainfall and location. This system is further split into two different systems: the crop predictor, whose main task is to help agriculturists by recommending crops and the rainfall prediction system that predicts the occurrence of rainfall for each month across the year. The next method discussed was Precision Agriculture using Map-Reduce used to allow variable rates and inputs which help in the understanding of time and space variability in criterion. Here the data is obtained and pre-processed. Then map-reduce is performed, and 3D visualization is done to visualize the output. Further crop prediction using various machine learning approaches were discussed. A few of them were 1) Grey wolf optimization (GWO) technique 2)

K-means clustering 3) Apriori algorithm 4) Naive Baye. The author states that obstacles faced for agriculture are usually Technical or Organizational problems. The paper further mentions the problems faced in the big data analysis of agriculture data, majorly, availability, accessibility and scalability of data for analysis.

2.2REFERENCE:

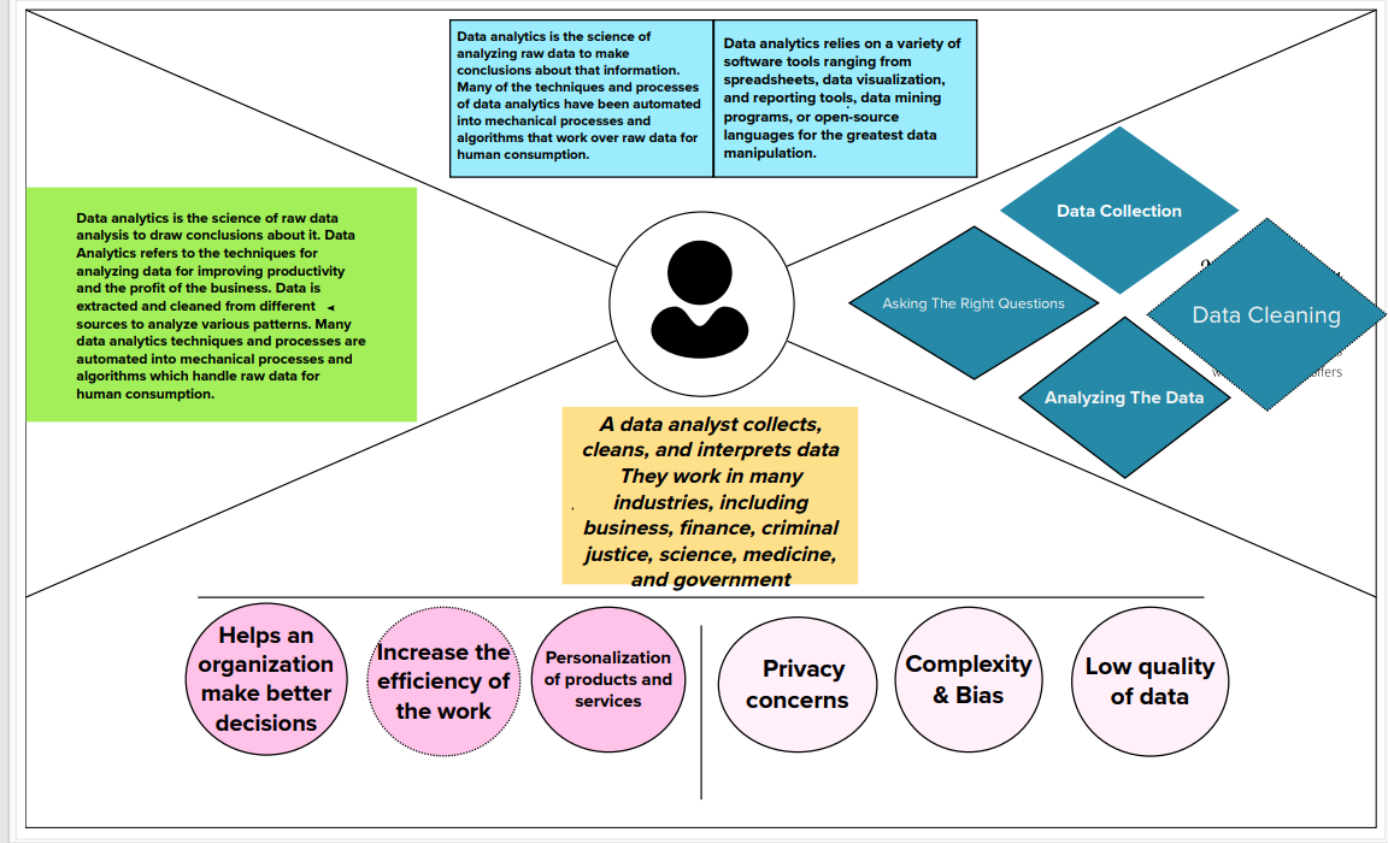
- https://www.degruyter.com/document/doi/10.1515/jisys-2012- 0016/html
- https://ieeexplore.ieee.org/document/8697806
- https://www.researchgate.net/publication/339102917_Big_data_analytics_in_Agriculture

2.3PROBLEM STATEMENT DEFINITION:

To create a dashboard and perform analysis of crop production in India using IBM Cognos analytic platform. Crop production in India is one of the most important sources of income and India is one of the top countries to produce 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.

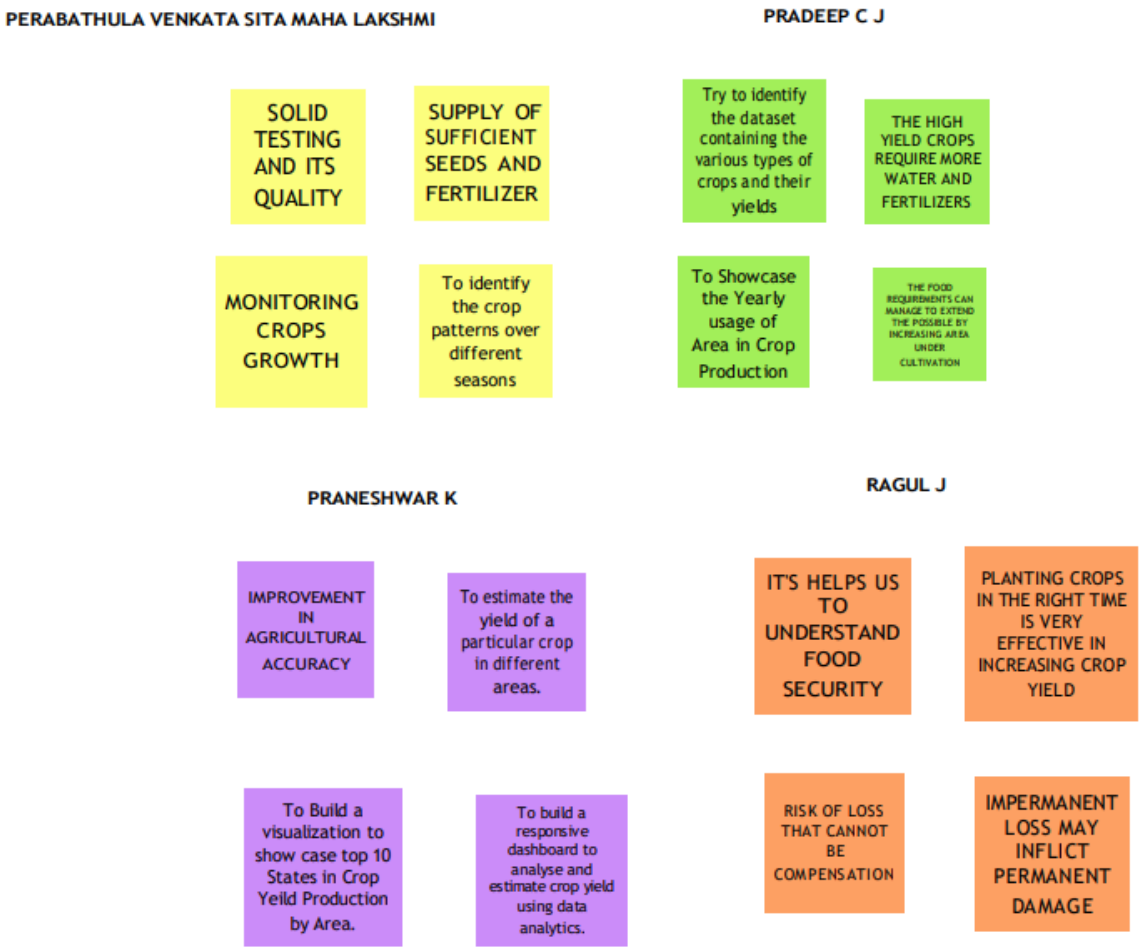
3. IDEATION & PROPOSED SYSTEM

3.1 EMPATHY MAP CANVAS:

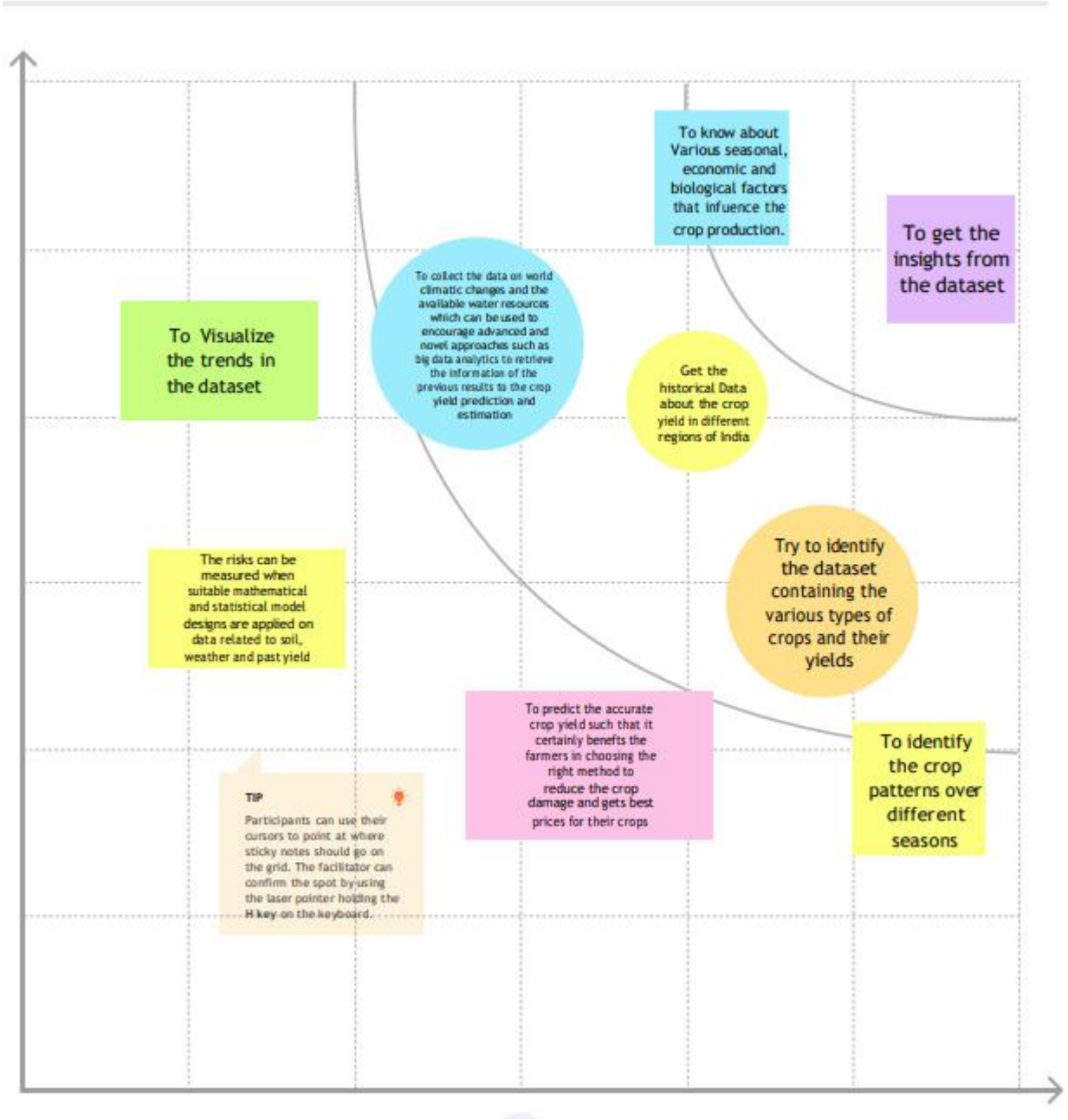


3.2IDEATION & BRAINSTORMING:

A) BRAINSTORMING:



B) IDEA PRIORITIZATION:

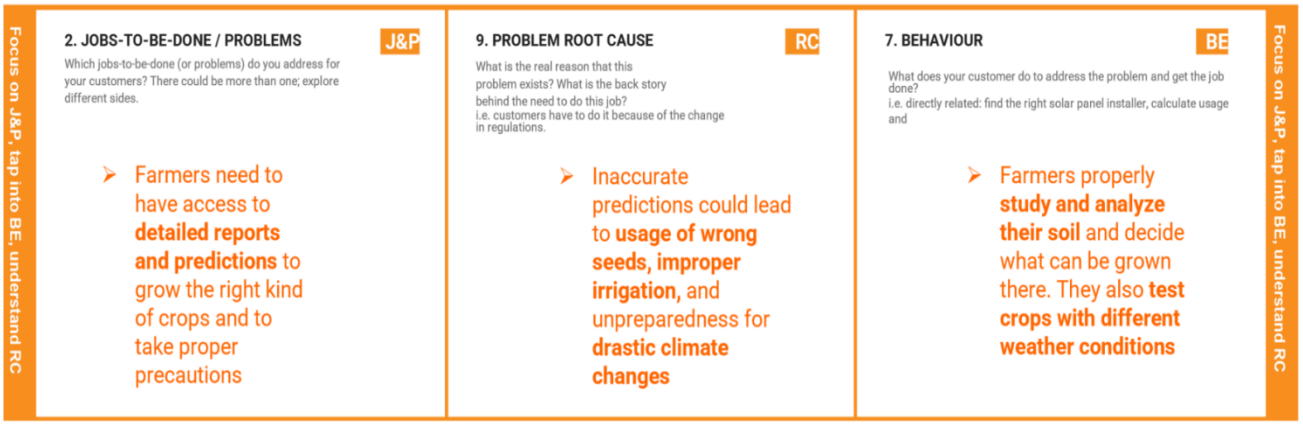


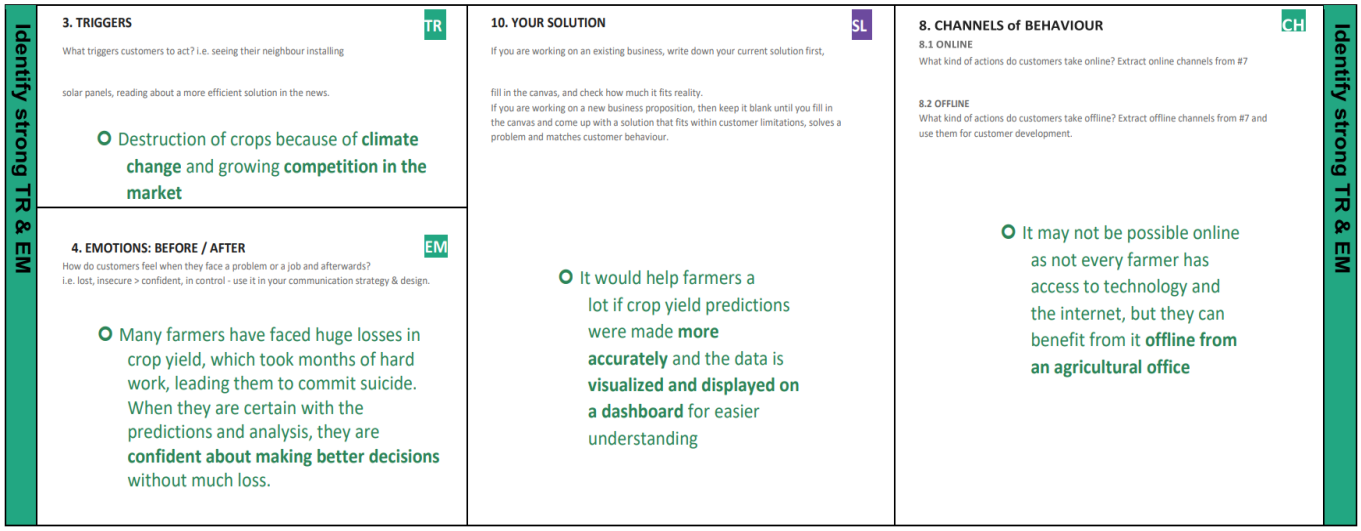
3.3PROPOSED SOLUTION:

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	To Estimate crop yield analysis, using data analysis, to aid farmers in making better decisions in order to have healthy crop production.
2.	Idea / Solution description	To deliver a perfectly analyzed Dashboard of historical agricultural production data from several Indian states so that farmers may forecast their crop yield.
3.	Novelty / Uniqueness	The dataset contains information about the crops in various Districts, States, Seasons, and Areas. Therefore, using all these facts, a thoroughly researched report will assist farmers in making the best crop choice for their region during a specific growing season to increase output.
4.	Social Impact / Customer Satisfaction	The issues that farmers have with yield potential will all be resolved by this report. Therefore, this Dashboard will have a significant impact on farmers, and by adopting the advised crops, they can achieve enormous earnings.
5.	Business Model (Revenue Model)	Profit can be generated by marketing the solution as a freely accessible mobile application that anyone can use. Venture partnerships with the government may yield financial rewards.
6.	Scalability of the Solution	Regarding dataset storage and data gathering, there are no problems. As a result, the system may be readily scaled to manage rising user numbers, traffic, and requirements that must be met.

3.4PROBLEM SOLUTION FIT:

The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you have realized for it actually solves the customer's problem. It helps entrepreneurs, marketers and corporate innovators identify behavioral patterns and recognize what would work and why.




Identify strong TR & EM
3. TRIGGERS
TR
What triggers customers to act? i.e. seeing their neighbour installing
solar panels, reading about a more efficient solution in the news.
Destruction of crops because of climate change and growing competition in the market
4. EMOTIONS: BEFORE / AFTER
EM
How do customers feel when they face a problem or a job and afterwards?
i.e. lost, insecure > confident, in control - use it in your communication strategy & design.
Many farmers have faced huge losses in crop yield, which took months of hard work, leading them to commit suicide. When they are certain with the predictions and analysis, they are confident about making better decisions without much loss.
10. YOUR SOLUTION
SL
If you are working on an existing business, write down your current solution first,
fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.
It would help farmers a lot if crop yield predictions were made more accurately and the data is visualized and displayed on a dashboard for easier understanding
8. CHANNELS of BEHAVIOUR
CH
8.1 ONLINE
What kind of actions do customers take online? Extract online channels from #7
8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.
It may not be possible online as not every farmer has access to technology and the internet, but they can benefit from it offline from an agricultural office
Identify strong TR & EM

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT:

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through Linked IN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	User Profile	User Details Farm Details
FR-4	Required Data	The previous year crop yield data set Farm yield methodology User data of the farmer Details of the Seasons and the Regions
FR-5	Analysis	Cleaning and analysis of the past year crop yields Visualizing the datasets using IBM Cognos
FR-6	Estimation	Creating the perfect data module through attractive stories, dashboard and reports to increase the understandability of data.

4.2NON – FUNCTIONAL REQUIREMENT:

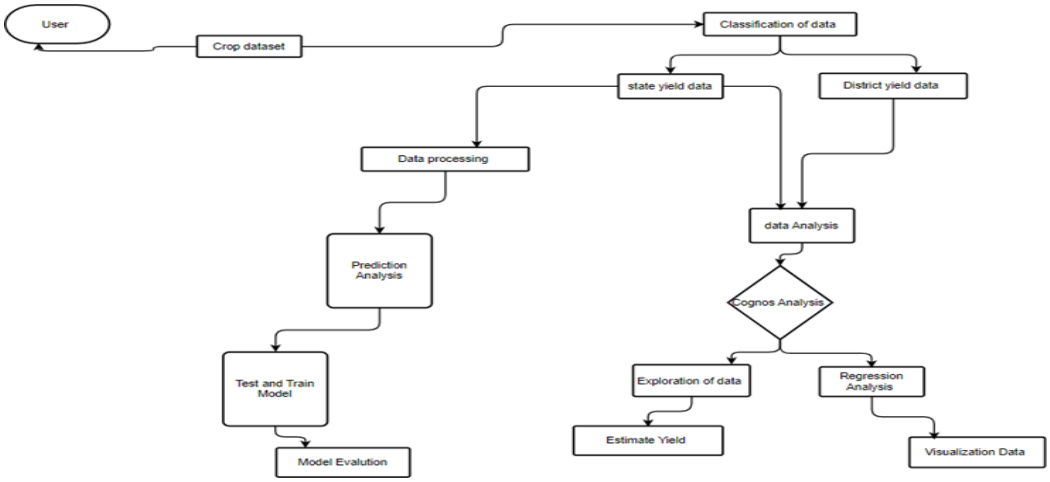
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	From the given datasets, analysis is done and a report iscreated. Accordingly, sowing of crops is recommended.
NFR-2	**Security**	Usage of IBM COGNOS, will provide secure user information (Data Visualization)
NFR-3	**Reliability**	Using the interactive data visual dashboards, we can easily understand the data reports.
NFR-4	**Performance**	Interaction makes better performance between all users and impresses by the data visuals advice.
NFR-5	**Availability**	The dashboard is easily available and accessible in smart phones and PC's.
NFR-6	**Scalability**	Prediction of crops for the forthcoming year can be done. It gives you a variety of crops to choose from our region. Also to know the better profitability of crops.

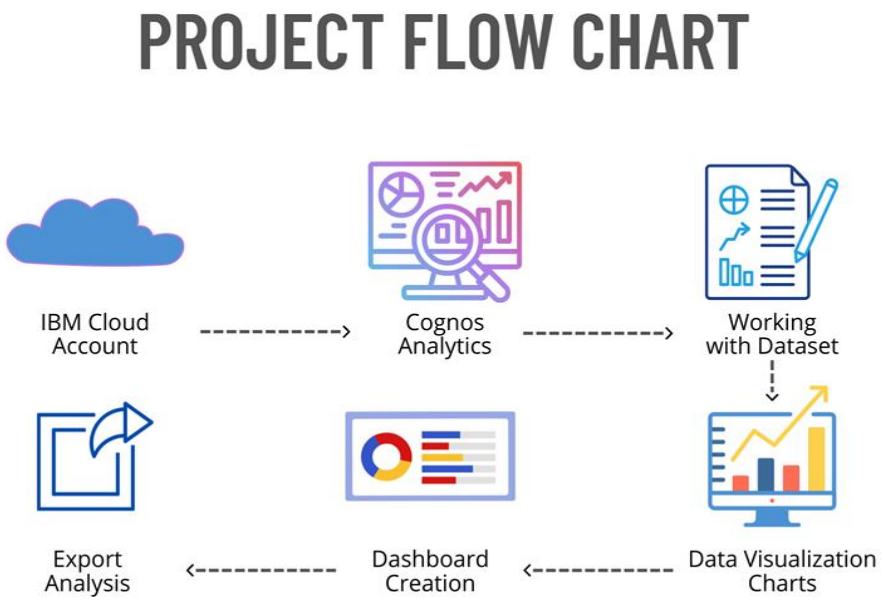
5. PROJECT DESIGN

5.1 DATA FLOW DIAGRAM:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

Project flow for estimating the crop yield using data analytics is shown below.





5.2SOLUTION AND TECHNOLOGY ARCHITECTURE:

The deliverables has include the architectural diagram as below and the information as per the table 1 and table 2

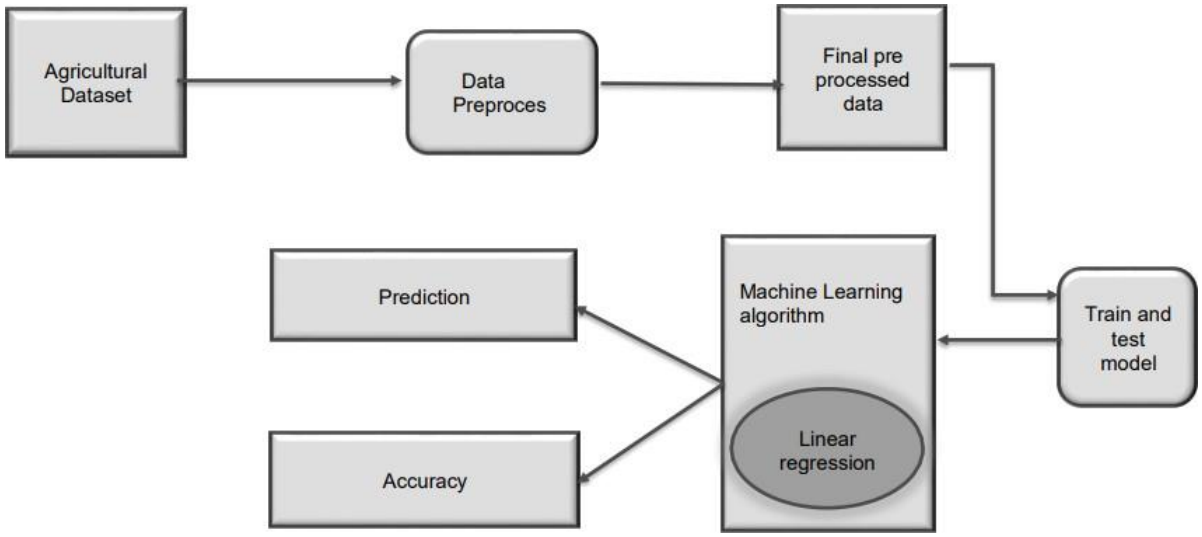


Table-1: Components & Technologies:

S. No	Component	Description	Technology
1.	User Interface	How user interacts with application e.g. Web UI, Mobile App ,Chat bot etc.	HTML, CSS, JavaScript.
2.	Applicationlogic1	Login as a user in the application	Java/Python
3.	Applicationlogic2	Login as admin in the application	IBM Watson STT service
4.	Applicationlogic3	Login as merchants in the application	IBM Watson Assistant
5.	Database	Data related to crop production in previous and also crop data.	MySQL , NoSQL , etc.
6.	Cloud Database	Database Service on Cloud	IBMDB2,IBM Cloudant etc.
7.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local File system
8.	ExternalAPI-1	Weather API are application programming interface that allow you to connect to large databases.	IBM Weather API ,etc.

9.	ExternalAPI-2	Soil testing is a quick and accurate method to determine the relative acidity of the soil and the level ofseveral essential nutrient needed.	Soil API, etc.
10.	Machine Learning Model	It is mostly used for finding out the relationship between variables and forecasting	Linear Regression
11.	Infrastructure(Server/Cloud)	Application Deployment on Local System/Cloud Local Server Configuration CloudServerConfiguration:l1	Local, Cloud Foundry, Kubernete, etc.

Table-2:ApplicationCharacteristics:

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Bootstrap is a free ,open source front-end development frame work	Bootstrap ,React etc.,
2.	Security Implementations	Improves user experience and provides greater security.	Authentication etc.
3.	Scalable Architecture	A3-tier architecture where in application gets data from various sources, manipulates it, stores the min IBM Cloud and Cognos.	IBM Cloud, IBM Cognos.
4.	Availability	The application is being developed is made available to all users	Cognos Analytics

5.	Performance	Multiple technologies and services that will improve the usability in agriculture activities.	Robots, IOT agriculture sensors.

5.3USER STORIES:

User Story Number	User Story / Task
USN-1	Understanding the data set .
USN-2	Loading the data set.
USN-3	Convert the data into required format
USN-4	Explore the data's which is uploaded in the IBM cognos
USN-5	Creating the data visualization chart
USN-6	Creating a dashboard
USN-7	Estimation of accuracy using random forest algorithm
USN-8	Export the analytics

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 by entering my Agri - id card and request.	2	High	Perabathula Venkata sita maha Lakshmi Pradeep CJ Praneshwar K Ragul J
		USN-2	As a user, I can register for the application through Gmail	2	Medium	Perabathula Venkata sita maha Lakshmi Pradeep CJ Praneshwar K Ragul J
	Login	USN-3	As a user, I can Call and request or Approach for dataset	4	High	Praneshwar K Ragul J
	Working with the Dataset	USN-4	To work on the given dataset, Understand the Dataset.	2	High	Perabathula Venkata sita maha Lakshmi Pradeep CJ Praneshwar K Ragul J
		USN-5	Load the dataset to Cloud platform then Build the required Visualizations.	10	High	Perabathula Venkata sita maha Lakshmi Pradeep CJ

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-2	Data Visualization Chart	USN-7	Using the Crop production in Indian dataset, create various graphsand charts to highlight the insights and visualizations. *Build a Visualization to showcase Average Crop Production by Seasons.	4	Medium	Perabathula Venkata sita maha Lakshmi Pradeep CJ Praneshwar K Ragul J
			*Showcase the Yearly usage of Area in Crop Production.	4	Medium	Perabathula Venkata sita maha Lakshmi Pradeep CJ
			Build Visual analytics to represent the States with SeasonalCrop Production using a Text representation.	4	Medium	Perabathula Venkata sita maha Lakshmi Pradeep CJ
Sprint-3	Creating The dashboard	USN-8	Create the Dashboard by using the created visualizations.	20	High	Perabathula Venkata sita maha Lakshmi Pradeep CJ Praneshwar K Ragul J
Sprint-4	Export The Analytics	USN-9	Export the created Dashboard	20	High	Perabathula Venkata sita maha Lakshmi Pradeep CJ Praneshwar K Ragul J

6.2SPRINT 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

6.3REPORT FROM JIRA:

7.CODING AND SOLUTIONING

7.1 FEATURE 1:

Variable Identification Process / data validation process:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data.
(For example to show the data type format of given dataset)

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 decisioning.

```
#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
```

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

	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 pre processing given data frame

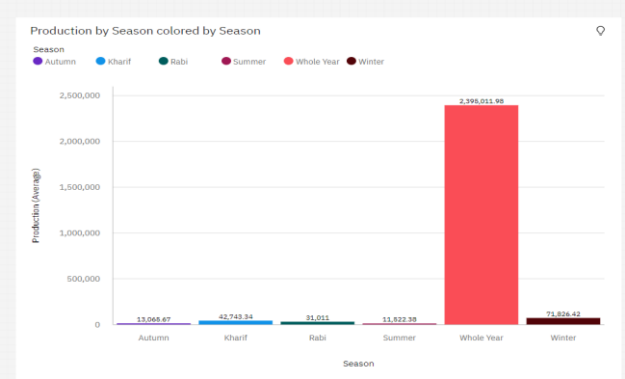
7.2 FEATURE 2:

Exploration data analysis of visualization:

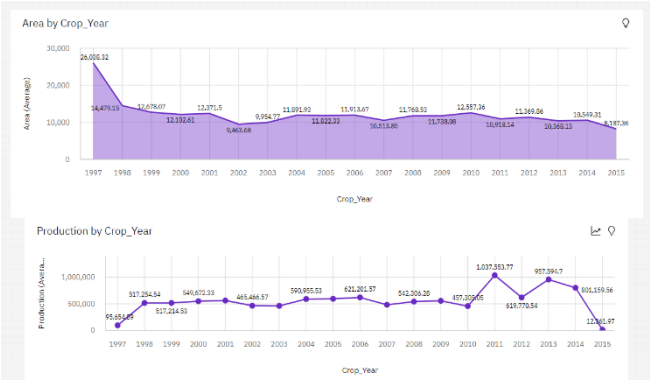
Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and 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.
- How to summarize the relationship between variables with scatter plots.

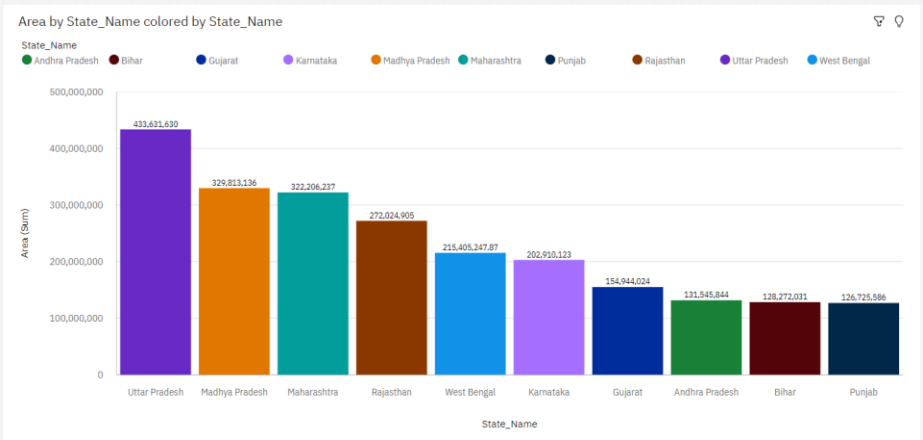
1) Seasons with average production:



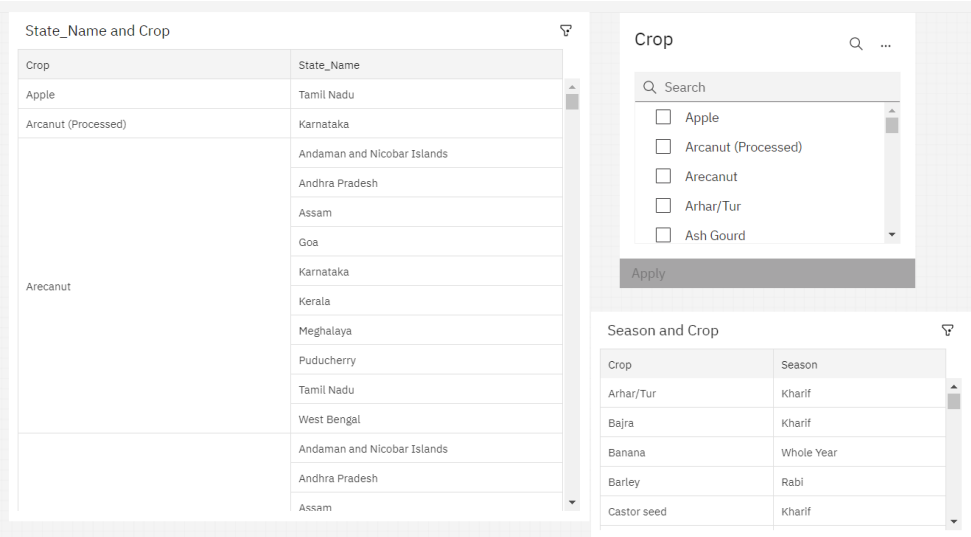
2) With years usage of area and production:

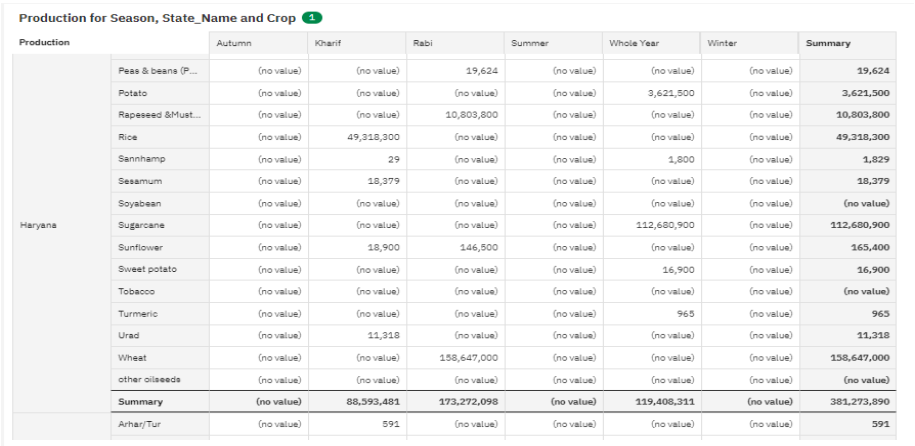


3) Top 10 states with the most area.



4) States with crop production with seasons:

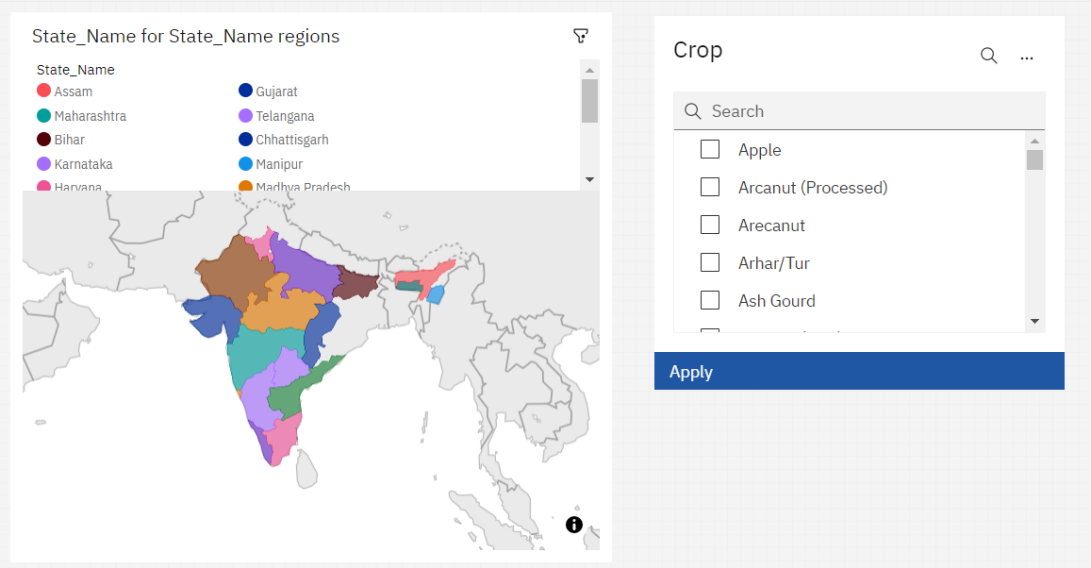




Production for Season, State_Name and Crop 1

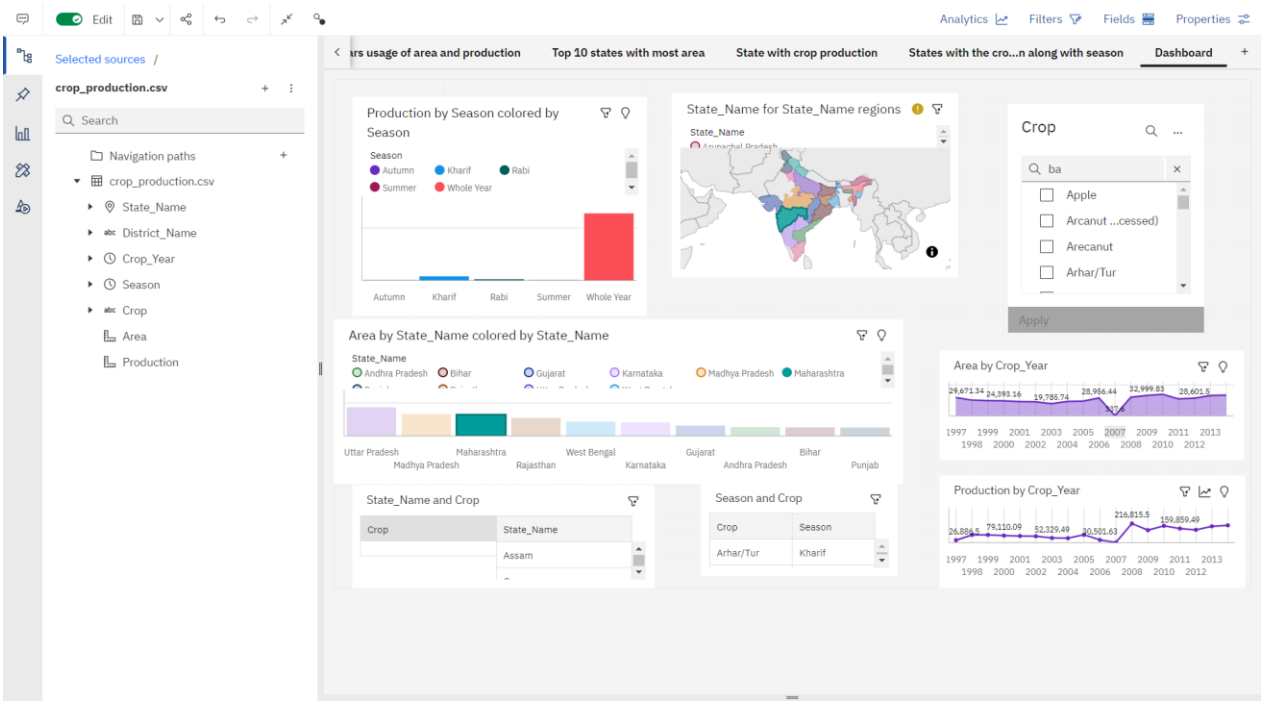
Production		Autumn	Kharif	Rabi	Summer	Whole Year	Winter	Summary
Haryana	Peas & beans (P...	(no value)	(no value)	19,624	(no value)	(no value)	(no value)	19,624
	Potato	(no value)	(no value)	(no value)	(no value)	3,621,500	(no value)	3,621,500
	Rapeseed &Must...	(no value)	(no value)	10,803,800	(no value)	(no value)	(no value)	10,803,800
	Rice	(no value)	49,318,300	(no value)	(no value)	(no value)	(no value)	49,318,300
	Sannhamp	(no value)	29	(no value)	(no value)	1,800	(no value)	1,829
	Sesamum	(no value)	18,379	(no value)	(no value)	(no value)	(no value)	18,379
	Soyabean	(no value)	(no value)	(no value)	(no value)	(no value)	(no value)	(no value)
	Sugarcane	(no value)	(no value)	(no value)	(no value)	112,680,900	(no value)	112,680,900
	Sunflower	(no value)	18,900	146,500	(no value)	(no value)	(no value)	165,400
	Sweet potato	(no value)	(no value)	(no value)	(no value)	16,900	(no value)	16,900
	Tobacco	(no value)	(no value)	(no value)	(no value)	(no value)	(no value)	(no value)
	Turmeric	(no value)	(no value)	(no value)	(no value)	965	(no value)	965
	Urad	(no value)	11,318	(no value)	(no value)	(no value)	(no value)	11,318
	Wheat	(no value)	(no value)	158,647,000	(no value)	(no value)	(no value)	158,647,000
	other oilseeds	(no value)	(no value)	(no value)	(no value)	(no value)	(no value)	(no value)
	Summary	**(no value)**	**88,593,481**	**173,272,098**	**(no value)**	**119,408,311**	**(no value)**	**381,273,890**
	Arhar/Tur	(no value)	591	(no value)	(no value)	(no value)	(no value)	591

5) States with crop production:



7.3 FEATURE 3:

CREATING THE DASHBOARD:



CODING:

- https://colab.research.google.com/drive/14py8wcogYJRqZNE1vGeyyIa2XAUI0H8H

8. TESTING

8.1 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 system meets 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 software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

8.2 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. TESTING

9.1 Performance Metrics:

Comparing Algorithm with prediction in the form of best accuracy result:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test.

In the example below 2 different algorithms are compared:

- Logistic Regression
- Random Forest

- dimensions of new features in a numpy array called ‘n’ and it want to predict the species of this

features and to do using the predict method which takes this array as input and spits out predicted target value as output.

- So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

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.

Used Python Packages:

Sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

tkinter:

- Standard python interface to the GUI toolkit.
- Accessible to everybody and reusable

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

- Predicting productivity of crop in various climatic conditions can **help farmer and other partners in essential basic leadership as far as agronomy and product decision**.
- This model can be used to select the most excellent crops for the region and also its yield thereby improving the values and gain of farming also.
- This will help the policy makers of the state to determine the budget.
- If the production of a crop observes a declining trend then, they can plan to implement the schemes at an early stage. This in return will save the state from shortage of the product.
- Monitors the growth of healthy crops.
- Helps the government to frame the government policies.
- Yield data helps the farmer to determine how much they should plant next year.
- Helps the farmer in Seed Selection, Pest Management, Irrigation Scheduling, etc,…
-

CHALLENGES:

Challenges are the major basis which imminent the negative impacts on current project. Some of the challenges faced during crop yield prediction are:

- Choosing appropriate dataset, after choosing dataset tuning of the parameters which makes project more efficient to get the desired results.
- Model must be trained by taking consideration of less computational efficiency and power.
- Increase of error rate due to dynamically changing the environment.

11. CONCLUSION

Our project will make policy maker of the state to determine the budget. If the production of a crop observes a declining then, they can plan to implement the schemes at an early stage. This in return will save the state from shortage of the product. Monitors the growth of healthy crops. Helps the government to frame the government policies. 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. The analysis of agricultural productivity and the uncovering of hidden patterns utilizing data sets related to seasons and crop yields have been noted in the literature. Using IBM Cognos, we have observed and conducted analysis regarding various crops grown, areas, and productions in various states and districts. **The scope of the project is to determine the crop yield of an area by considering dataset with some features which are important or related to crop production such as temperature, moisture, rainfall, and production of thecrop in previous years. To predict a continuous value, regression models are used.**

12. FUTURE SCOPE

Our future scope is to add many more geographical features and predict using those features.

13. APPENDIX

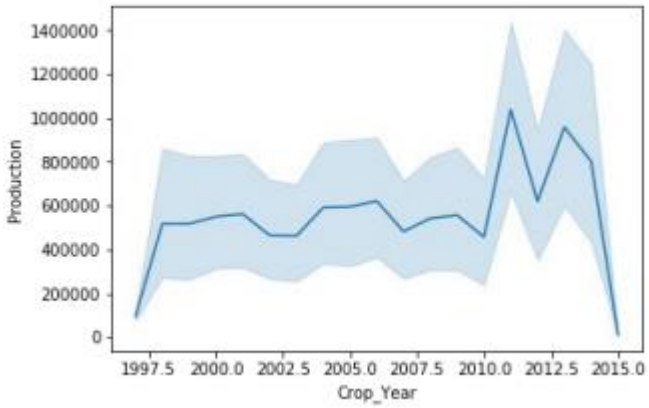
SOURCE CODE:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
df = pd.read_csv("../input/crop-production-in-india/crop_production.csv")
df[:5]

df.isnull().sum()

# Droping Nan Values
```

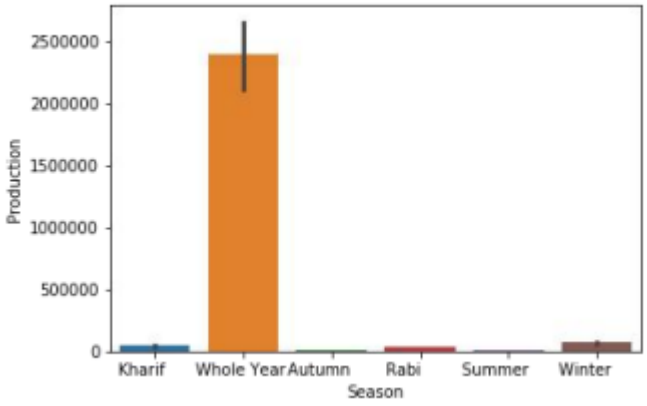
```
data = df.dropna()
print(data.shape)
test = df[~df["Production"].notna()].drop("Production",axis=1)
print(test.shape)
sum_maxp = data["Production"].sum()
data["percent_of_production"] = data["Production"].map(lambda x:(x/sum_maxp)*100)
data[:5]
```

Data Visualization

```
sns.lineplot(data["Crop_Year"],data["Production"])
```




```
plt.figure(figsize=(25,10))

sns.barplot(data["State_Name"],data["Production"])
plt.xticks(rotation=90)

sns.barplot(data["Season"],data["Production"])
```



Insights:

```
coc_df = data[data["Crop"]=="Coconut "]
print(coc_df.shape)
coc_df[:3]

sns.barplot("Season","Production",data=coc_df)
plt.figure(figsize=(13,10))
sns.barplot("State_Name","Production",data=coc_df)
plt.xticks(rotation=90)
plt.show()
top_coc_pro_dis =
coc_df.groupby("District_Name")["Production"].sum().reset_index().sort_values(
by='Production',ascending=False)
top_coc_pro_dis[:5]
sum_max = top_coc_pro_dis["Production"].sum()
top_coc_pro_dis["precent_of_pro"] = top_coc_pro_dis["Production"].map(lambda
x:(x/sum_max)*100)
top_coc_pro_dis[:5]
plt.figure(figsize=(18,12))
sns.barplot("District_Name","Production",data=top_coc_pro_dis)
plt.xticks(rotation=90)
plt.show()
plt.figure(figsize=(15,10))
sns.barplot("Crop_Year","Production",data=coc_df)
plt.xticks(rotation=45)
#plt.legend(rice_df['State_Name'].unique())
plt.show()
```

Insight from Coconut Production

```
sug_df = data[data["Crop"]=="Sugarcane"]
print(sug_df.shape)
sug_df[:3]

plt.figure(figsize=(13,8))
```

```
sns.barplot("State_Name","Production",data=sug_df)
plt.xticks(rotation=90)
plt.show()
```

Feature Selection

```
data1 = data.drop(["District_Name","Crop_Year"],axis=1)
data_dum = pd.get_dummies(data1)
data_dum[:5]
```

Test Train Split

```
x = data_dum.drop("Production",axis=1)
y = data_dum[["Production"]]
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.33, random_state=42)
print("x_train :",x_train.shape)
print("x_test :",x_test.shape)
print("y_train :",y_train.shape)
print("y_test :",y_test.shape)
```

Model -1: Random Forest

```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor()
model.fit(x_train,y_train)
preds = model.predict(x_test)
rom sklearn.metrics import r2_score
r = r2_score(y_test,preds)
print("R2score when we predict using Randomn forest is ",r)
```

Model -2: Linear Regression

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train,y_train)
preds = model.predict(x_test)
from sklearn.metrics import mean_squared_error, r2_score
mean_squared_error(y_test,preds)
r2_score(y_test,preds)
```

Model 3: XGBRegresor

```
import xgboost as xgb
xgbr = xgb.XGBRegressor(verbosity=0)
xgbr.fit(x_train,y_train)
preds = xgbr.predict(x_test)
mean_squared_error(y_test,preds)
r2_score(y_test,preds)
```

Model 4: Decision Tree

```
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor(random_state=42)
regressor.fit(x_train,y_train)
preds = regressor.predict(x_test)
mean_squared_error(y_test,preds)
r2_score(y_test,preds)
```

GIT REPO LINK:

https://github.com/IBM-EPBL/IBM-Project-21618-1659785900

DEMO LINK:

https://youtu.be/wQ2qlGawB24