VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

BUSINESS INTELLIGENCE IN AGRICULTURE

COURSE NAME : **BUSINESS INTELLIGENCE**

COURSE CODE : **CSI 3017**

SLOT : **C1+TC1+TCC1**

TEAM MEMBERS :

LOKESH R	**19MID0003**
PRATHIBAN V	**19MID0010**
ARUN SREERAM R	**19MID0041**

GUIDED BY :

Professor DR. DEEPA K

"Science and technology coupled with improved human capital have been powerful drivers of positive change in the performance and evolution of small holder systems"

-Food and agriculture organization of the United Nations

BUSINESS INTELLIGENCE IN AGRICULTURE

***Abstract*-The practise of producing plants and cattle is known as agriculture. Agriculture was a significant factor in the rise of sedentary human civilization, as it enabled humans to live in cities by creating food surpluses from tamed species. India leads the globe in net cropped area, followed by the United States and China. As per 2018, agriculture employed more than 50% of the Indian work force and contributed 17–18% to country's GDP. The total agriculture commodities export was US$3.50 billion in March - June 2020. The goal of Business Intelligence is to figure out how farmers can use the information they've gathered to better manage the resources they have, boost productivity and sustainability, and cut expenses. BI systems can track the financial side of your business and help you increase productivity to guarantee you're on track to meet your goals. Agricultural entrepreneurs can utilise Business Analytics tools to make easier and better decisions based on data. When properly designed, BI systems can monitor every part of an agricultural operation and gather and analyse data that farmers may utilise to enhance their operations. BI is on the lookout for new ways to save costs while making few or no sacrifices and it wants to optimise earnings. Farms can make greater use of their data by implementing BI systems. Making data-driven decisions, gaining a competitive advantage, and improving forecasts are all things that can be done using data. It helps to create an agri-tech network. BI may use data from prior years to improve planning and prevent having to start from scratch with each new production cycle. The ability of information and knowledge development for assisting decision making has been proved by Business Intelligence (BI) technologies with the Extract, Transform, and Loading process, Data Warehouse, and Power BI. In this project, we are going to find insights and**

patterns using BI tools and crop prediction using machine learning

- ***KEYWORDS: GDP, Agri-tech, Data Warehouse, Power BI***

I. INTRODUCTION

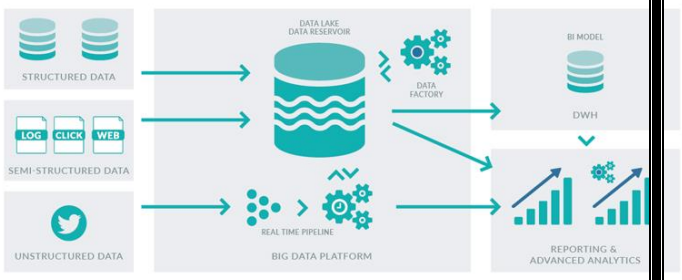
Agriculture is the raising and rearing of animals and plants for the purpose of producing food, materials, medicinal herbs, and other items in order to maintain and improve living conditions. Farming was a key factor in the development of sedentary human society, as it allowed people to sleep in cities by producing food surpluses from tamed species. Rural science is the study of agriculture.

II. HISTORY

People began collecting wild cereals at least 105,000 years ago and began planting them around 11,500 years ago before they became tamed. Around 10,000 years ago, pigs, sheep, and cattle were domesticated. Crops can be found in at least 11 different places of the world. Despite the fact that industrialized farming based on large-scale monoculture has come to dominate agricultural production in the previous century, about 2 billion people still rely on subsistence farming.

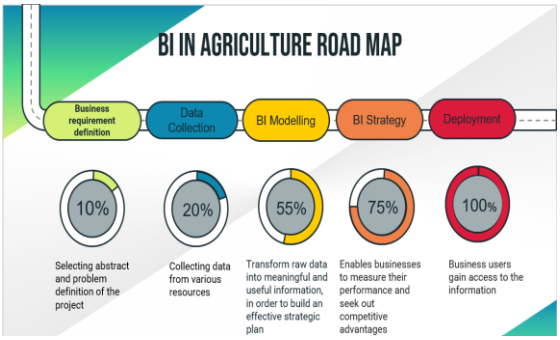
III. AGRICULTURE AND TECHNOLOGY

First and foremost, business intelligence aids in the making of more informed decisions. Farming's mission of providing a life-sustaining service allows minimal space for error, and farms require reliable data. Farmers may better comprehend the impact of their decisions on specific regions and the company as a whole by using BI tools, which deliver real-time data throughout the whole business operation.



IV. USES OF BI IN AGRICULTURE

- BI links numerous sources of data in order to have a better understanding of your farming operations.
- BI systems can track the financial side of your business and help you increase productivity to guarantee you're on track to meet your goals.
- Batch numbers, harvest dates, storage conditions, sell dates, and other crucial data that can assist you limit the consequences of a recall are all kept track of by BI.
- BI systems are comprehensive in nature and can assist farms in identifying waste areas throughout their operations. This enables you to plug money leaks before they cause financial devastation.



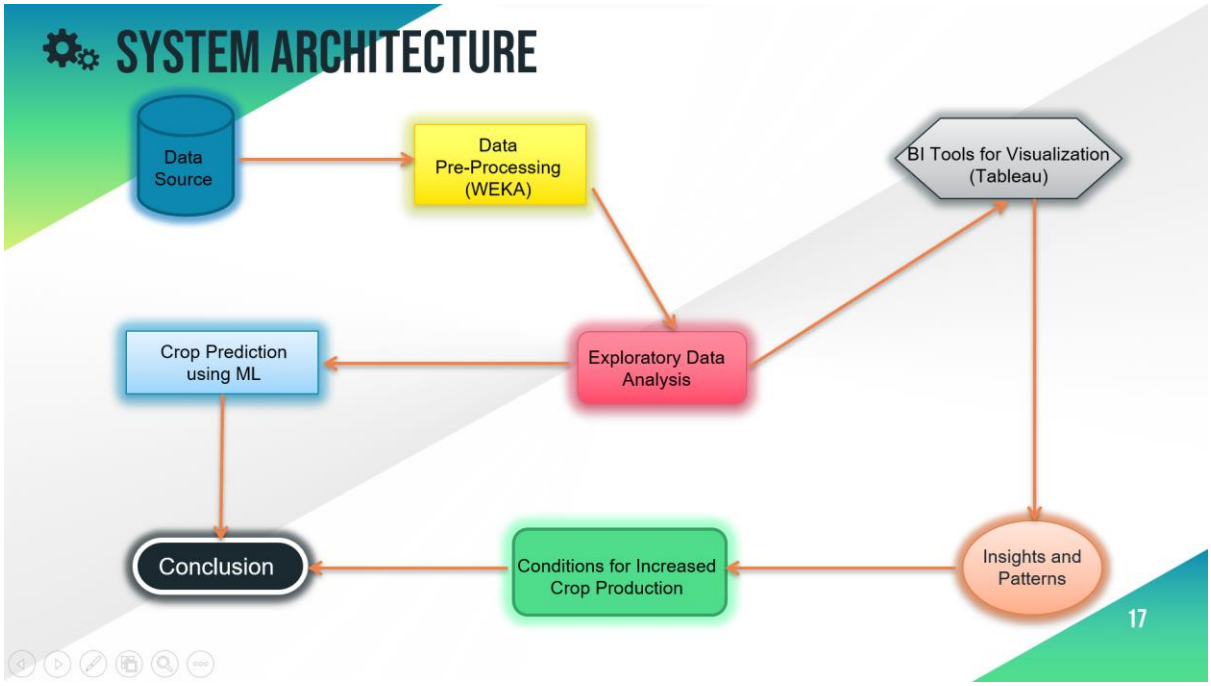
V. LITERARURE REVIEW

S.No	NAME	Journal and Year	Findings	Pros	Cons
1	Using Business intelligence for data analysis and decision support in agriculture	Conference on Biotechnology with International Participation, 2017	Modern technologies in the field of sensors, electronics and computing provide relatively inexpensive data integration solutions from all business processes and create preconditions for efficient analysis and reporting in management support functions.	Contribute to increasing the production potential and technical efficiency of agricultural enterprises and farms due to effective management support, analytical and planning activities.	The application of BI still not at a satisfactory level when it comes to small agricultural producers, as the level of application of these technologies are at a lower level, reasons for this this technologies and systems still inaccessible to small manufacturers due to high cost.
2	Business Intelligence and Business and analytics applied to the management of agricultural resources	IEEE Explore, 2021	Through the use of the IoT system and detection sensors pest control and field monitoring the agricultural sector constantly generates considerable amounts of data that need to be aggregated in order to be analysed.	The combination of these techniques results in a increase in productivity along with an agricultural practice more sustainable way in which resources are applied in a more sustainable way. More efficient, also contributing to the reduction of costs	Two factors that make it difficult this process of modernization, they are age and the low schooling level. It is then necessary to increase the incentives for the emergence of more young farmers and with better levels of training.

S.No	Name	Journal and Year	Findings	Pros	Cons
3	Artificial Intelligence Measuring, automatic Control and Expert Systems in Agriculture	IFAC Proceedings Volumes,2008	The presence of specific agricultural technological and bionic-industry processes, an expert system model for intelligent measuring processing in precision agriculture was developed	At a higher rate would correspond to perspective tasks and aims both of AI systems and accurate physical magnitude measurements	Error relativity and world outlook spatial-temporal fundamental problems of artificial intelligence
4	Impact of information in agriculture sector	International Journal of Food, Agriculture and vertinary sciences, 2014	This technology allows developing and underdeveloped countries to develop plans and compete with developed countries	IT has the potential to play a significant role in assisting rural India's development in order to solve these problems and close the rapidly widening digital divide	Only 38% of households in India are digitally literate

S.No	NAME	Journal and year	Findings	Pros	Cons
5	Intelligent Technologies for the Conformity Assessment in the Chain of Agricultural Production	Scientific Journal of Latvia University of Agriculture, 2007	Intelligent technology for assessing conformance in agricultural production applications.	The food product testing sector as a whole is steadily expanding.	When there is dispute over what is truly required to maintain safety, problems develop.
6	Development of intelligent technologies and systems in agriculture	Scientific Journal of Latvia University of Agriculture, 2008	Intelligent sensors and measurement tools for agricultural usage are being developed.	Intelligent technology can be successfully applied to the development of energy saving solutions to improve agricultural energy efficient	Danger can never be completely removed.

SYSTEM ARCHITECTURE:

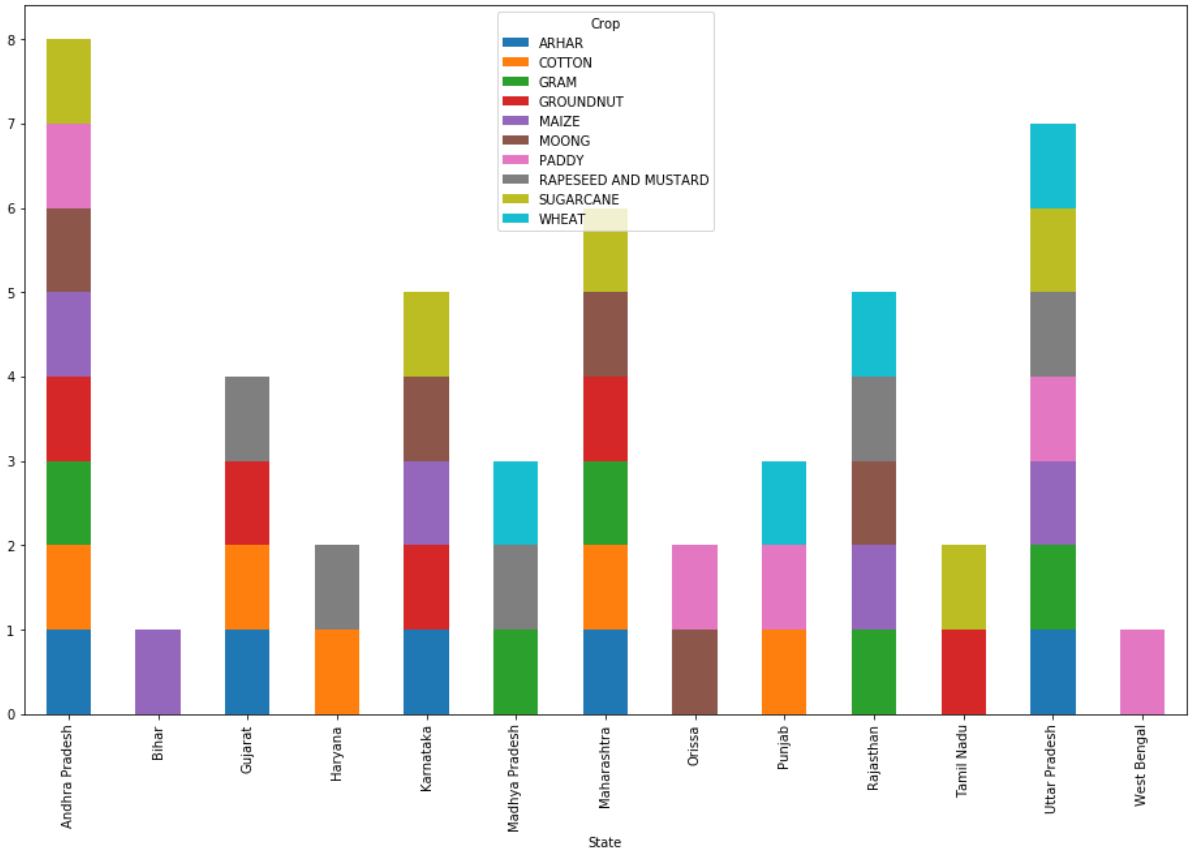


STEPS INVOLVED IN SYSTEM ARCHITECTURE:

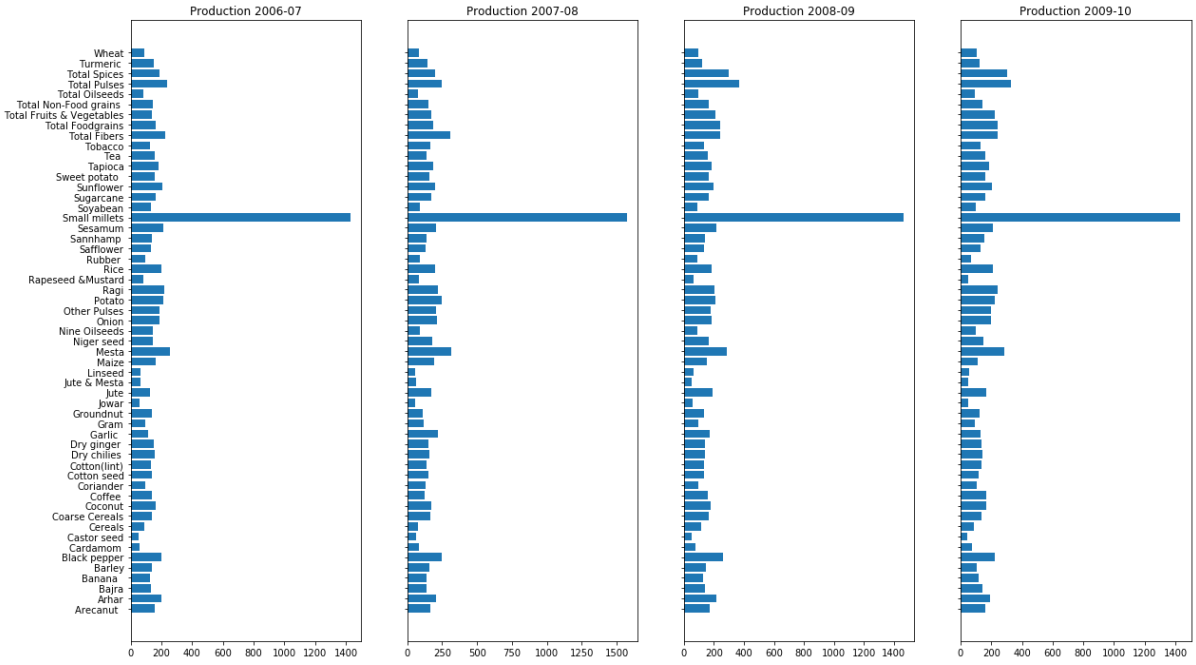
- First we collect our data from data sources, majorly from internet
- Then we apply data pre-processing techniques using a tool called “WEKA”, from which we can remove NULL values data, remove outliers and extreme values and we can also normalize our data
- After data pre-processing techniques we prepare our data for Exploratory data analysis where we visualize our data
- We perform EDA using a tool called tableau
- From EDA we can derive useful insights and patterns
- From which we can find out the conditions for increased crop production
- We also predict what crop needs to be cultivated inorder to get good profits based opn soil conditions such as ph value, acidity, rainfall, temperature etc. using machine learning algorithms

EXPLORATORY DATA ANALYSIS:

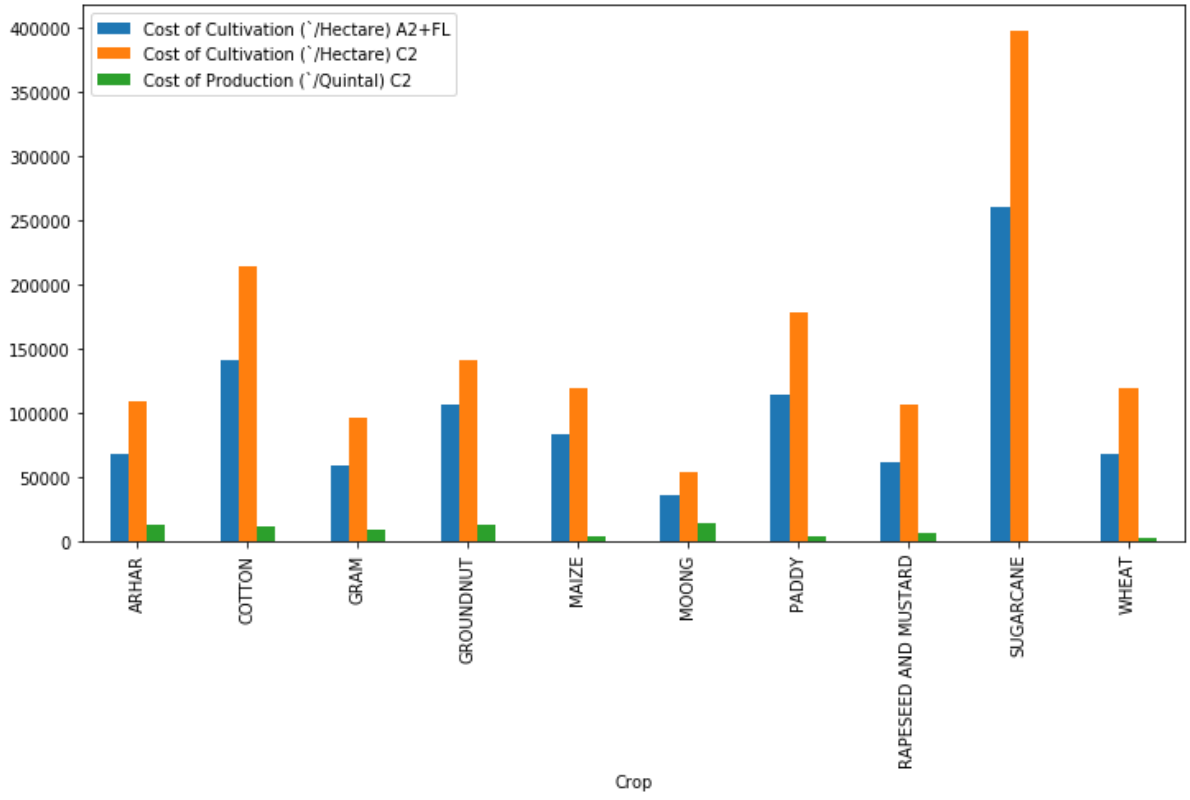
Stacked Bar Chart :



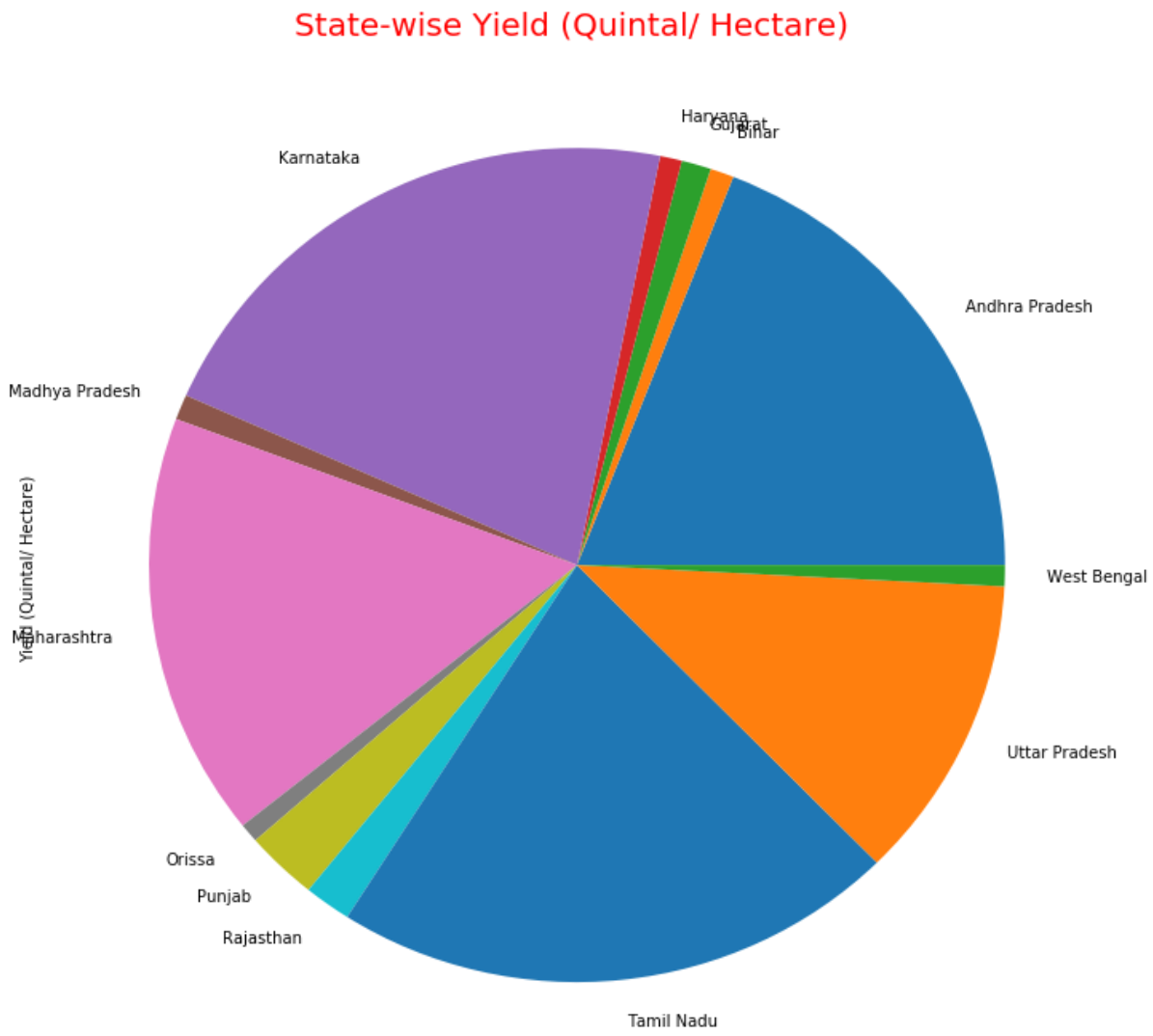
Varieties of crops have been visualized with respect to each state in India using stacked bar chart. In the above graph it has been seen that Andhra Pradesh tops and Bihar and West Bengal are least.



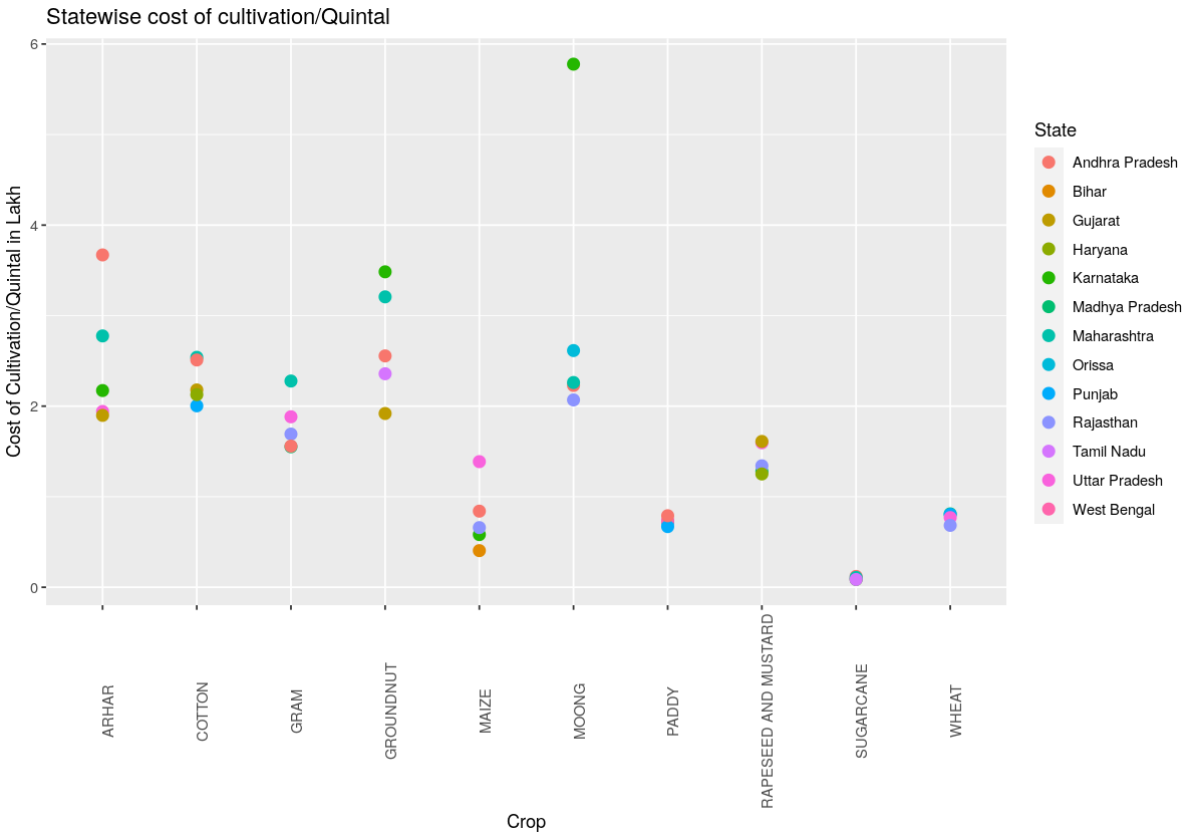
Production is compared with Years and the production rate of variety of crops using vertical bar chart. It has been seen that small millets has the highest production rate in al the 4 years.



Using grouped bar chart, visualizing the cost for cultivation and production with respect to variety of crops. In the above chart it has been seen that sugarcane needs more cost of production and cultivation.



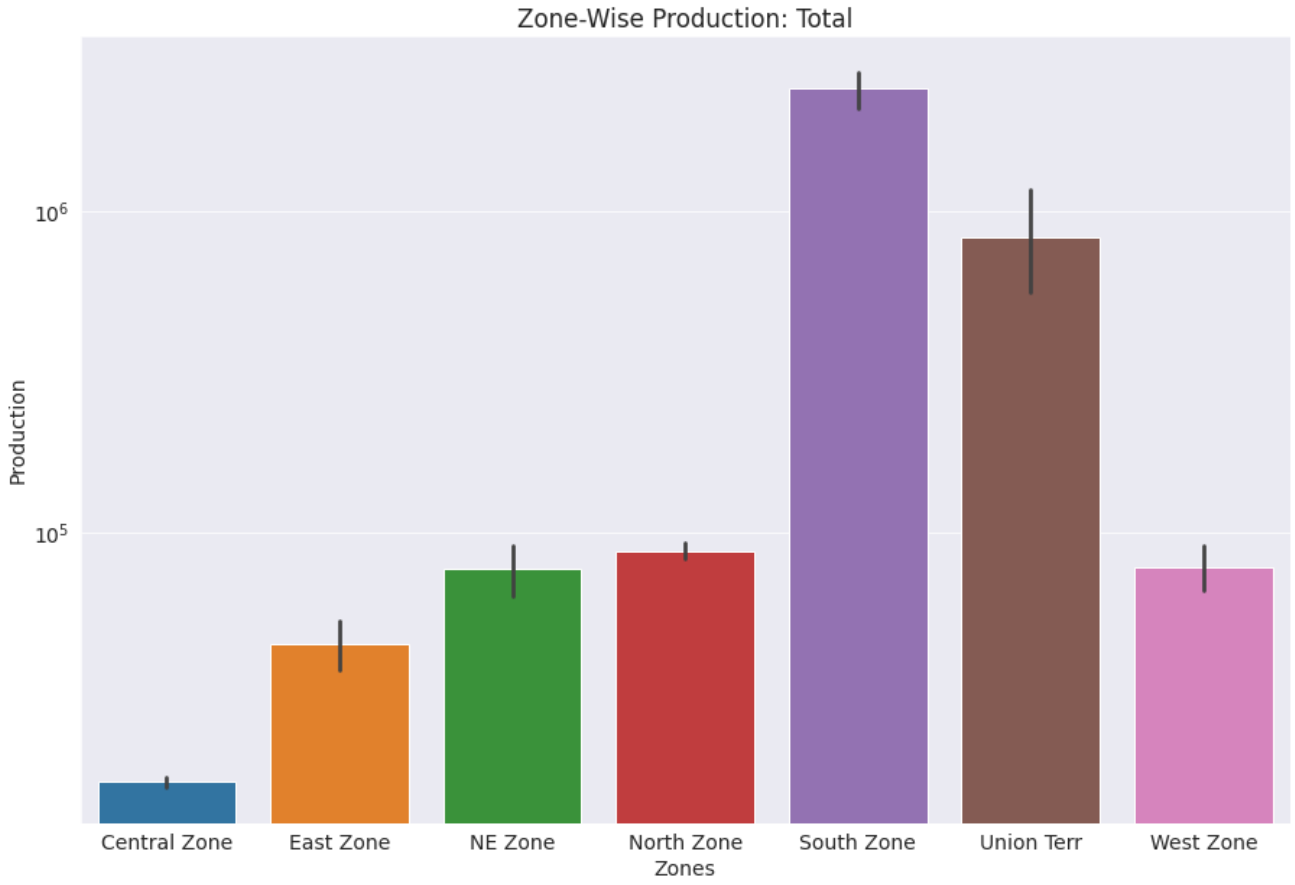
Using Pie chart State wise yield has been visualized. In the above chart it has been analyzed that Tamil Nadu has the highest yield and Orissa has the least yield.



Using Dot chart, state wise cost of cultivation is analyzed with respect to variety of crops in all states. By analyzing it has been seen that Cost of cultivation of the Moong crop in Andhra Pradesh has the highest cost of cultivation

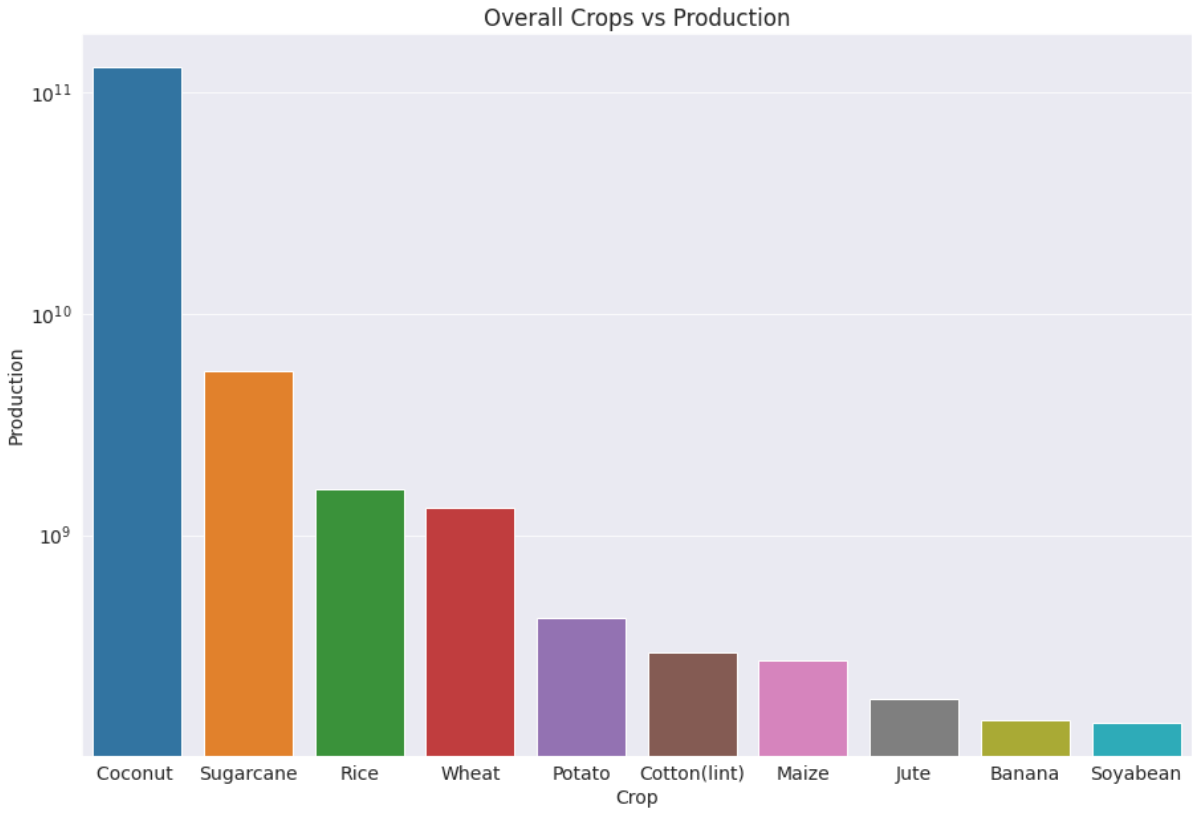
Zonal distribution of crops:

Production wise top zone is South India



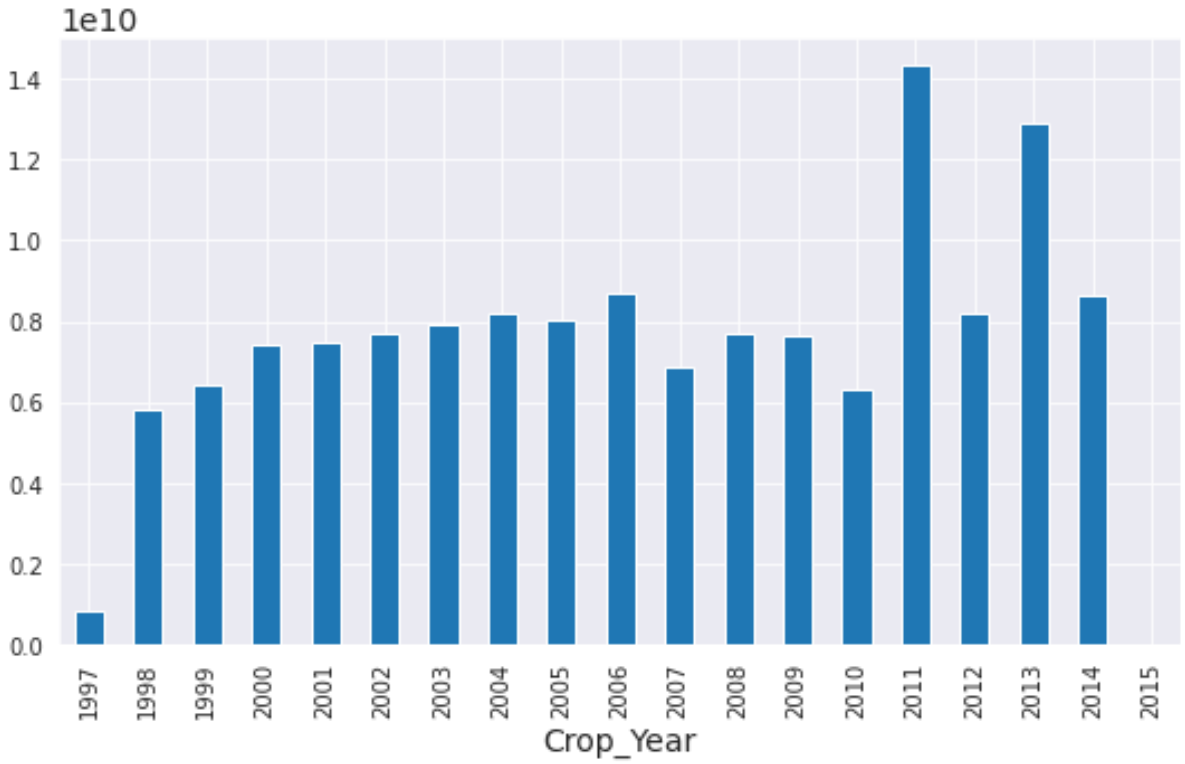
Crop wise Production status:

Top Crops Production wise are: Coconut, Sugarcane and Rice.



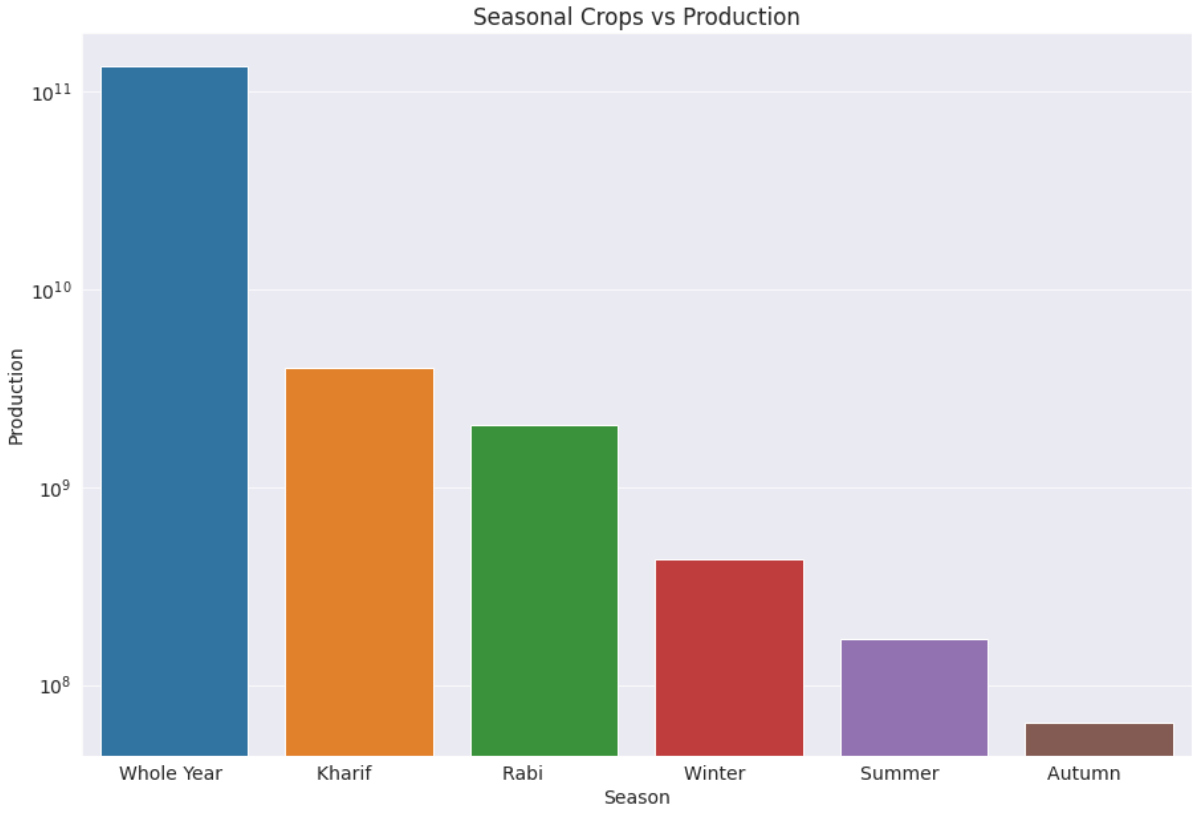
Yearwise Production Status:

Top production years are 2011, 2013 and 2014.



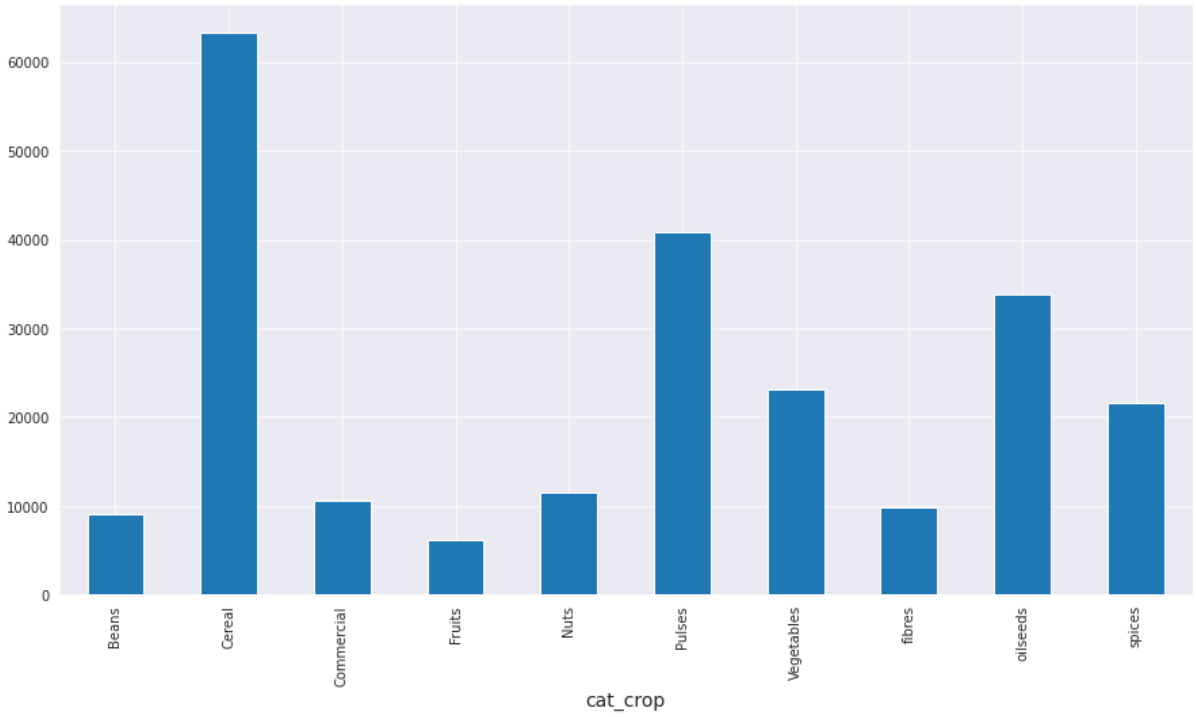
Season wise Production Status:

Top crop categories which shows high production values are Whole Year(Annual growing plants),Kharif and Rabi crops. It clearly shows these crops heavily dependent on seasonal monssons.



Crop wise Production plot describing production values for all crop types.

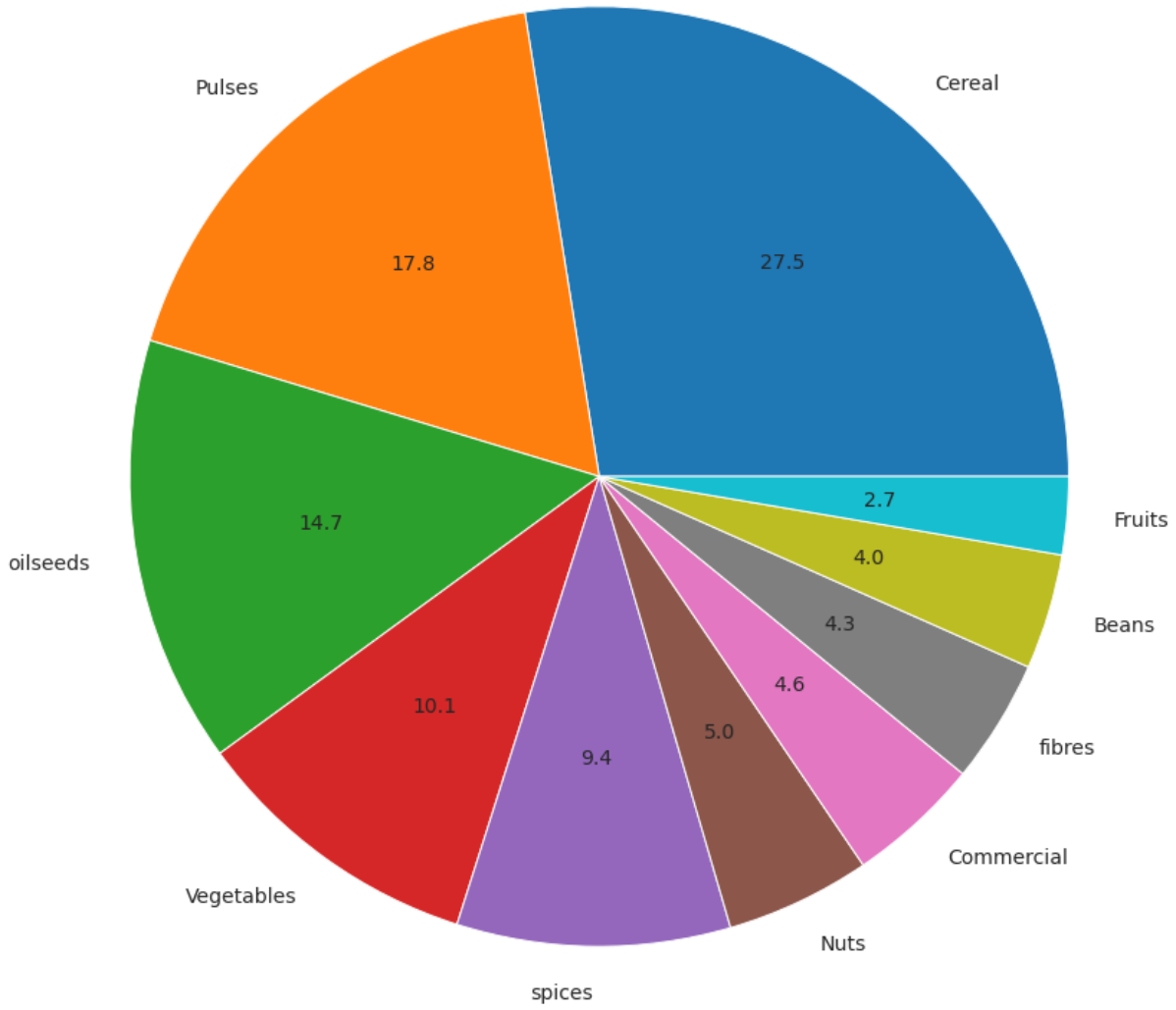
Top crop categories are Cereal, Pulses and Oilseeds.



State versus Crop Category versus Season plot:

Interesting facts: South zone: i. Top producing state Kerela shows a abundance of whole year seasonal crops, North Zone: ii. Top producing state Uttar Pradesh shows abundance of Kharif, Rabi and Summar crops

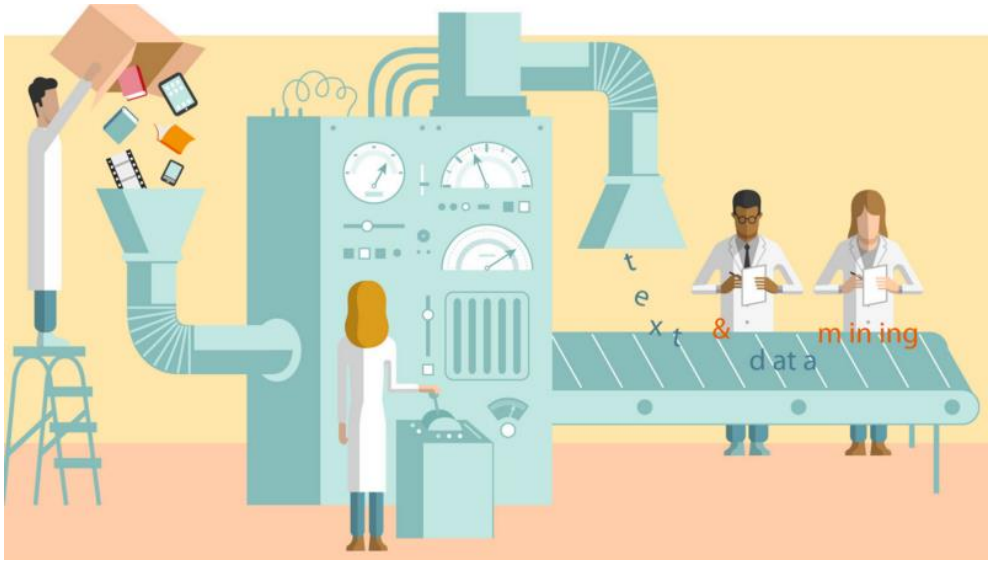
Different proportion of Crop Categories for India:



MACHINE LEARNING:-

Machine learning is an AI application that allows systems to learn and improve based on their own experiences without being explicitly programmed. Machine learning is concerned with creating computer programs that can access data and use it to learn on their own.

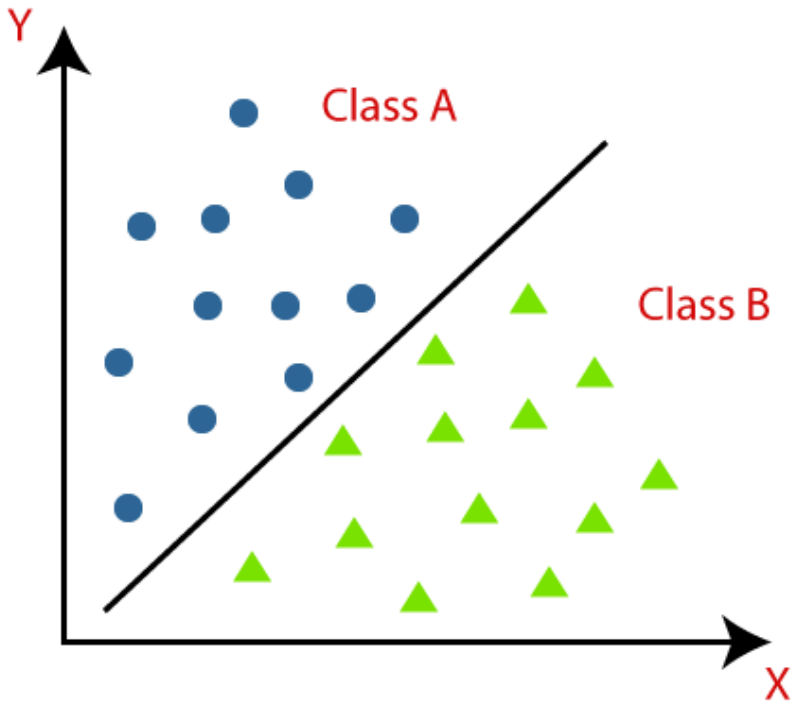
Machine learning can speed up data processing and analysis, making it a useful technology for predictive analytics programs. With minor changes in deployment, predictive analytics algorithms can train on even larger data sets and perform deeper analysis on multiple variables using machine learning.



CLASSIFICATION:-

Classification is the process of categorizing a given set of data into classes. It can be done with both structured and unstructured data. Predicting the class of given data points is the first step in the process. The classes are also known as the target, label, or categories.

Approximating the mapping function from discrete input variables to discrete output variables is the task of classification predictive modeling. The primary goal is to determine which class/category the new data will belong to.



STEPS IN CLASSIFICATION:-

I. Get data
II. Clean the data
III. Split the data into training and testing
IV. Apply various ML algorithm on the data
V. Select the ML algorithm with best accuracy
VI. Find the recommended crop with the selected algorithm

STEP(I):

Data collection is the process of acquiring, collecting, extracting, and storing large amounts of data, which can be structured or unstructured, such as text, video, audio, XML files, records, or other image files that will be used in later stages of data analysis.

"Data collection" is the first step in analysing patterns or useful information in data. The data to be analysed must be gathered from a variety of reliable sources. The data source is from government of India. The dataset contains temperature, humidity, ph, rainfall and name of the crop which is suitable for increased crop production. It is a verified dataset and data source.

```
# Importing libraries

from __future__ import print_function
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import classification_report
from sklearn import metrics
from sklearn import tree
import warnings
warnings.filterwarnings('ignore')
```

```
df = pd.read_csv('../Data-processed/crop-recommendation.csv')
```

```
df.head()
```

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

```
df.tail()
```

	N	P	K	temperature	humidity	ph	rainfall	label
2195	107	34	32	26.774637	66.413269	6.780064	177.774507	coffee
2196	99	15	27	27.417112	56.636362	6.086922	127.924610	coffee
2197	118	33	30	24.131797	67.225123	6.362608	173.322839	coffee
2198	117	32	34	26.272418	52.127394	6.758793	127.175293	coffee
2199	104	18	30	23.603016	60.396475	6.779833	140.937041	coffee

STEP(II):

The process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted is known as data cleaning. When it comes to data analysis, this data is usually not required or helpful because it can hinder the process or produce inaccurate results. We can clean the data using NULL option in python. We can also perform data cleaning with help of WEKA tool.

STEP(III):

Train/Test is a technique for determining the accuracy of your model. It is called Train/Test because the data set is divided into two parts: a training set and a testing set. Training accounts for 80% of the budget, while testing accounts for 20%. The training set is used to train the model. You use the testing set to put the model through its paces. Training the model entails creating the model. Testing the model entails determining the model's accuracy.

The reason for splitting data into training and testing set is that when the dataset is divided into train and test sets, the training dataset will not contain enough data for the model to learn an effective mapping of inputs to outputs. There will also be insufficient data in the test set to evaluate the model's performance effectively. We use in build function in python to split the data into training and testing.

Seperating features and target label

```
features = df[['N', 'P','K','temperature', 'humidity', 'ph', 'rainfall']]
target = df['label']
#features = df[['temperature', 'humidity', 'ph', 'rainfall']]
labels = df['label']
```

```
# Initializing empty lists to append all model's name and corresponding name
acc = []
model = []
```

```
# Splitting into train and test data

from sklearn.model_selection import train_test_split
Xtrain, Xtest, Ytrain, Ytest = train_test_split(features,target,test_size = 0.2,random_state =2)
```

STEP(IV):

We apply various ML algorithms on our dataset to find the algorithm with best accuracy and minimum error. Different algorithms have their own method for solving the problem so applying different algorithm helps us understand the problem to find the model with best accuracy. Algorithm we are going to use are

- Decision Tree
- Naive Bayes
- Support Vector Machine
- Logistic Regression
- Random Forest
- XGBoost

Decision Tree:

In the form of a tree structure, decision tree constructs classification or regression models. It incrementally divides a dataset into smaller and smaller subsets while also developing an associated decision tree. The end result is a tree with leaf nodes and decision nodes. A decision node (for example, Outlook) may have two or more branches (e.g., Sunny, Overcast and Rainy). A classification or decision is represented by a leaf node (for example, Play). The root node is the topmost decision node in a tree that corresponds to the best predictor. Both categorical and numerical data can be handled by decision trees.

Decision Tree

```
from sklearn.tree import DecisionTreeClassifier

DecisionTree = DecisionTreeClassifier(criterion="entropy",random_state=2,max_depth=5)

DecisionTree.fit(Xtrain,Ytrain)

predicted_values = DecisionTree.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Decision Tree')
print("DecisionTrees's Accuracy is: ", x*100)

print(classification_report(Ytest,predicted_values))
```

```
DecisionTrees's Accuracy is:  90.0
```

Naïve Bayes:

The Nave Bayes Classifier is a simple and effective Classification algorithm that aids in the development of fast machine learning models capable of making quick predictions. It is a probabilistic classifier, which means it predicts based on an object's probability.

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Guassian Naive Bayes

```
from sklearn.naive_bayes import GaussianNB

NaiveBayes = GaussianNB()

NaiveBayes.fit(Xtrain,Ytrain)

predicted_values = NaiveBayes.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Naive Bayes')
print("Naive Bayes's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

```
Naive Bayes's Accuracy is:  0.990909090909091
```

Support Vector Machine:

It transforms your data using a technique known as the kernel trick and then finds an optimal boundary between the possible outputs based on these transformations. Simply put, it does some extremely complex data transformations, then figures out how to seperate your data based on the labels or outputs you've defined.

Support Vector Machine (SVM)

```
from sklearn.svm import SVC
# data normalization with sklearn
from sklearn.preprocessing import MinMaxScaler
# fit scaler on training data
norm = MinMaxScaler().fit(Xtrain)
X_train_norm = norm.transform(Xtrain)
# transform testing dataabs
X_test_norm = norm.transform(Xtest)
SVM = SVC(kernel='poly', degree=3, C=1)
SVM.fit(X_train_norm,Ytrain)
predicted_values = SVM.predict(X_test_norm)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('SVM')
print("SVM's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

```
SVM's Accuracy is:  0.9795454545454545
```

Logisitic regression:

Logistic regression is a classification technique borrowed from statistics by machine learning. Logistic Regression is a statistical method for analysing a dataset in which one or more independent variables influence the outcome. The goal of logistic regression is to find the model that best describes the relationship between the dependent and independent variables.

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

LogReg = LogisticRegression(random_state=2)

LogReg.fit(Xtrain,Ytrain)

predicted_values = LogReg.predict(Xtest)

x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Logistic Regression')
print("Logistic Regression's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

```
Logistic Regression's Accuracy is:  0.9522727272727273
```

Random Forest:

Random forest constructs multiple decision trees and merges them to produce a more accurate and stable prediction. Random forest has a significant advantage in that it can be used for both classification and regression problems, which comprise the majority of current machine learning systems.

Random Forest

```
from sklearn.ensemble import RandomForestClassifier

RF = RandomForestClassifier(n_estimators=20, random_state=0)
RF.fit(Xtrain,Ytrain)

predicted_values = RF.predict(Xtest)

x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('RF')
print("RF's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

```
RF's Accuracy is:  0.990909090909091
```

XG Boost:

XGBoost is a gradient boosting ensemble Machine Learning algorithm that is decision-tree based. Artificial neural networks outperform all other algorithms or frameworks in prediction problems involving unstructured data (images, text, etc.). FORMULA

F2(x) = σ (0+1*h1(x)+1*h2(x))) where the resulting value of F2(x) is considered as the prediction from XgBoost model

XGBoost

```
import xgboost as xgb
XB = xgb.XGBClassifier()
XB.fit(Xtrain,Ytrain)

predicted_values = XB.predict(Xtest)

x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('XGBoost')
print("XGBoost's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

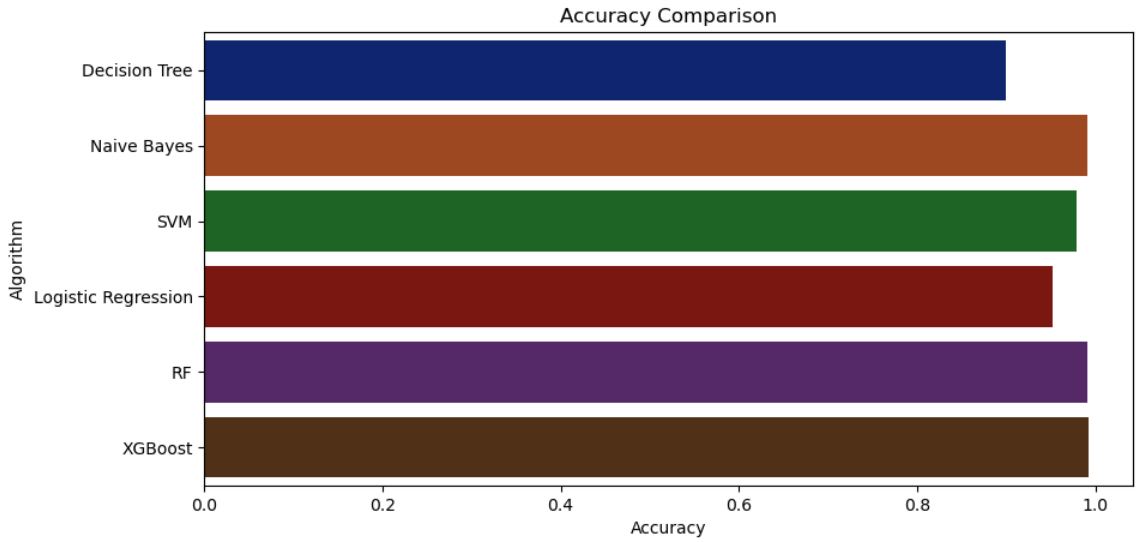
```
[14:16:03] WARNING: C:/Users/Administrator/workspace/xgboost-win6
metric used with the objective 'multi:softprob' was changed from
avior.
XGBoost's Accuracy is:  0.9931818181818182
```

STEP(V):

Select the ML algorithm with best accuracy, Choosing the best algorithm from list of available of list of algorithm is a essence of machine learning. Various machine learning algorithms look for various trends and patterns. For all data sets and application circumstances, one algorithm isn't the best. You'll need to run a lot of tests, analyze machine learning algorithms, and fine-tune their hyperparameters to find the optimum answer. To analyze your model and justify the choice of an algorithm, you must first identify, justify, and use a model performance indicator. The F1 score, true positive rate, and within cluster sum of squared error are examples of model performance measures. At least one deep-learning and one non-deep-learning algorithm should be used to implement your approach. Compare and document model performance after that. Apply at least one further iteration to the process model, at least one of which contains the feature generation task. Data normalization and principal component analysis, for example, have an impact on model performance (PCA). You may use specific technologies or frameworks to address your challenge depending on the algorithm class and data set size.

Accuracy Comparison

```
plt.figure(figsize=[10,5],dpi = 100)
plt.title('Accuracy Comparison')
plt.xlabel('Accuracy')
plt.ylabel('Algorithm')
sns.barplot(x = acc,y = model,palette='dark')
```




```
accuracy_models = dict(zip(model, acc))
for k, v in accuracy_models.items():
    print (k, '-->', v)
```

```
Decision Tree --> 0.9
Naive Bayes --> 0.990909090909091
SVM --> 0.9795454545454545
Logistic Regression --> 0.9522727272727273
RF --> 0.990909090909091
XGBoost --> 0.9931818181818182
```

Random forest has best accuracy we are going to choose random forest as our algorithm to find recommended crop.

STEP(VI):

Find the recommended crop with the selected algorithm, get the algorithm and put temperature, humidity, ph, rainfall in the algorithm to find the recommended crop.

Making a prediction

```
data = np.array([[104,18, 30, 23.603016, 60.3, 6.7, 140.91]])
prediction = RF.predict(data)
print(prediction)
```

```
['coffee']
```

```
data = np.array([[83, 45, 60, 28, 70.3, 7.0, 150.9]])
prediction = RF.predict(data)
print(prediction)
```

```
['jute']
```

Conclusion:

Uttar Pradesh is topping in producing more crop categories than any other Indian state and the stats are: Beans(1112), Cereal(9719),Commercial(1741), Fruits(269), Nuts(958), Pulses(6549), Vegetables(3734), Fibres(724), oilseeds(4028) and spices(2529). Rice is grown heavily when we look the frequency of crops in India Rice needs Winter for it mature Statewise Punjab dominates in rice production District wise its BARDHAMAN(2.13%), MEDINIPUR WEST(1.8%) and WEST GODAVARI(1.73%) which contributes to total rice production. Yearwise 2014 is the year when production reached the peak production

Correlation between Area and Production shows high production is directly proportional to Area under cultivation. Top cultivating states based on the Cultivation area are: Uttar Pradesh(4.33e+08), Madhya Pradesh(3.29e+08) and Maharashtra(3.22e+08) Yearwise Statues of these States: a.Uttar Pradesh: High Production was seen in 2005 and after that it's been reducing gradually. b. Madhya Pradesh:1998 showed a high production and then there was gradual reduction but it picked up and 2012 also showed a peak in Production c. Maharashtra:Production went down drastically in 2006 and again the levels went up and hit a high peak after 2007 d. Rajasthan: the production hit a all time low in the year 2002 and then picked up by 2010 e. West Bengal:the production hit a peak around 2006 but it has hit a low after 2007 and never recovered back.

Production wise top states of North zone are: Punjab(5.86e+08), Uttar Pradesh(3.23e+09), and Haryana(3.81e+08) Top crops of these states are: Sugarcane, Wheat and Rice Coconut cultivation is yearlong and doesn't get restricted to any particular seasons Top states involved in coconut production are: Kerala, Andhra Pradesh and Tamil Nadu Top districts featuring in coconut production is KOZHIKODE(11.75%), MALAPPURAM(11.16%) and THIRUVANANTHAPURAM(7.7%) Yearwise coconut cultivation is strong and its increasing healthy High coconut cultivation is directly proportional to area under cultivation.

We have successfully implemented machine learning model to find the right crop for the given conditions.

REFERENCES:

[1] Artificial Intelligence Measuring, automatic Control and Expert Systems in Agriculture,2008

[2] Impact of information in agriculture sector,2014

[3] Intelligent Technologies for the Conformity Assessment in the Chain of Agricultural Production,2007

[4] Business Intelligence and Business and analytics applied to the management of agricultural resources,2021

[5] Development of intelligent technologies and systems in agriculture,2008

[6] Using Business intelligence for data analysis and decision support in agriculture,2017