SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

"Sustainable and Efficient Crop Production: Leveraging Data Analytics to Address the Challenges of Climate Change"

INT 500 – INTERNSHIP 4

Final Review

YESWANTH R
E0119031
B.TECH CSE AIML

Supervisor
DR. G SHIYAMALA GOWRI
ASSISTANT PROFESSOR
AIML
SRIHER(DU)

Table of Contents

Introduction

- DA, (called digital agriculture/farming or smart farming), is a modern approach that uses digital and intelligent devices in conjunction with Data analytics to improve productivity and optimize the use of resources.

- The agriculture sector in India is vulnerable to climate change. Higher temperatures tend to reduce crop yields and favor weed and pest proliferation. Climate change can negatively affect irrigated crop yields across agroecological regions due to temperature rise and changes in water availability.

- Taking the various parameters for a crop to grow well and increase productivity and finding out the best crop to grow can be done for any area with the right variables.

- This project focuses on the efficiency and maximization of yield produced considering various climatic variables using AIML and Data analytics.

Literature Survey

- Clustering is considered an unsupervised classification process. A large number of clustering algorithms have been developed for different purposes. This can be categorized into Partitioning clustering, Hierarchical clustering, Density-based methods, Grid-based methods, and Model-based clustering methods.
- There are different forecasting methodologies developed and evaluated by researchers all over the world in the field of agriculture. Some of such studies are: Researchers like Ramesh and Vishnu Vardhan analyzed the agriculture data for the years 1965–2009 in the district East Godavari of Andhra Pradesh, India. The purpose of this work is to find suitable data models that achieve high accuracy and a high generality in terms of yield prediction capabilities.
- Bangladesh offers several varieties of rice which have different cropping seasons. For this, a prior study of climate (effect on temperature and rainfall) in Bangladesh and its effect on the agricultural production of rice has been done. Then this study was taken into regression analysis with temperature and rainfall. Temperature puts an adverse consequence on crop production.

Refrence	Technique	Accuracy
Analysis of agriculture data using data mining techniques: application of big data (2017)	Big Data, PAM, CLARA and DBSCAN	95%
Research Paper on Big Data Analytics for Agricultural Development in India (2019)	Data Mining, IOT,VRT	90%
Project Report on Data analysis and prediction for agricultural production (2020)	Matlab visualization	-
Data analytics for crop management: a big data view (2022)	CNN, LSTM	>90%
Assessment of agricultural drought based on multi-source remote sensing data (2023)	PCI, VCI, TCI	83%

Challenges

- The effects of climate change affect farmers' ability to grow the food we all need. Increasingly volatile weather and more extreme events – like floods and droughts – change growing seasons, limit the availability of water, allow weeds, pests, and fungi to thrive, and can reduce crop productivity.

- Climate change may require the selection of different crop varieties that are adapted to changing climatic conditions, which can pose challenges in terms of availability, access, and suitability.

- Increase the cost of production due to the need for more expensive inputs, such as irrigation, fertilizers, and pesticides.

- Addressing these challenges requires innovative and data-driven solutions that can help farmers adapt to changing climate conditions, optimize crop production, and ensure long-term sustainability.

Research Objective

- The research objective for using data analytics in crop production is to develop data-driven solutions that can help farmers optimize crop yield, increase efficiency, and mitigate the impact of climate change on crop production.

- The objective is to leverage data analytics to gain insights into key factors that increase production and visualize the trends over the years in this agricultural sector.

- By analyzing data from multiple sources to identify patterns, trends, and anomalies in crop management.

- Develop predictive models that can help farmers make informed decisions about crop selection, planting, irrigation, fertilization, and pest and disease management, and ensure sustainable and efficient crop production.

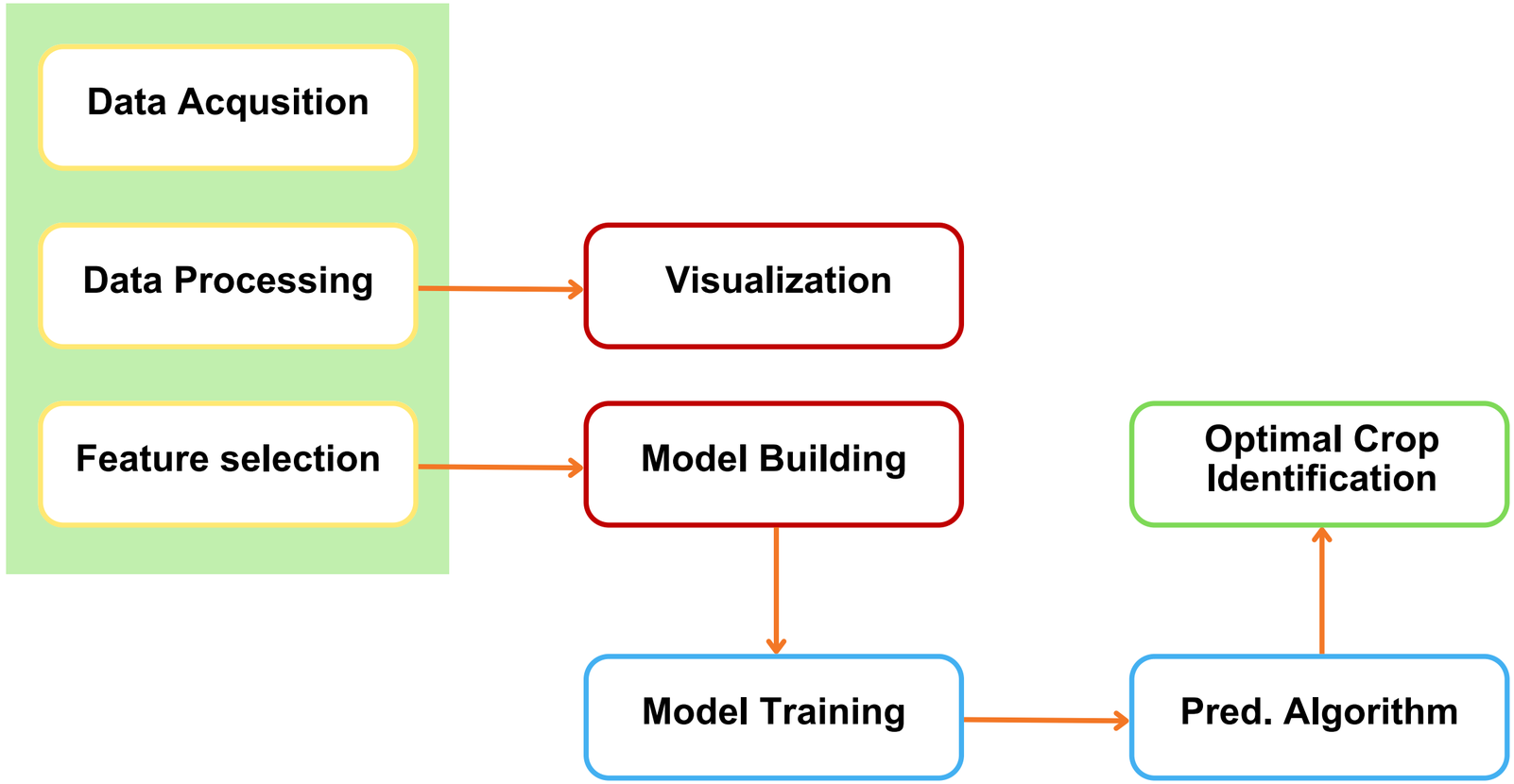
Tools and Techniques

Jupyter Notebook

R Studio

Python

Methodology - Architecture



Modules

Data Acquisition

- The goal of this step is to identify and obtain all data-related problems. In this step, we need to identify the different data sources, as data can be collected from various sources such as files and databases. The quantity and quality of the collected data will determine the efficiency of the output.

- The input dataset consists of data with the following parameters namely: Crop (cotton, groundnut, jowar, rice, and some fruit crops, etc.), Temperature, average rainfall (mm), soil, PH value, soil type, major fertilizers, nitrogen (kg/Ha), phosphorus (Kg/Ha), Potassium(Kg/Ha), minimum rainfall required, the minimum temperature required.

	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

Modules

Data Preparation

- Data preparation is a step used to understand the nature of data that we must work with. We need to understand the characteristics, format, and quality of data.

- Data wrangling is the process of cleaning and converting raw data into a usable format. It is the process of cleaning the data, selecting the variable to use, and transforming it into a proper format to make it more suitable for analysis in the next step.

```
crop.isna().sum()
```

```
N              0
P              0
K              0
temperature    0
humidity       0
ph             0
rainfall       0
label          0
dtype: int64
```

```
[7]  crop.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2200 entries, 0 to 2199
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   N            2200 non-null   int64  
 1   P            2200 non-null   int64  
 2   K            2200 non-null   int64  
 3   temperature  2200 non-null   float64
 4   humidity     2200 non-null   float64
 5   ph           2200 non-null   float64
 6   rainfall     2200 non-null   float64
 7   label        2200 non-null   object 
dtypes: float64(4), int64(3), object(1)
memory usage: 137.6+ KB
```

Modules

Data Preprocessing

- Study and understand the nature of data that was acquired in the previous step and know the quality of data. In this step, we will check for any null values and remove them as they may affect efficiency

- Identifying duplicates in the dataset and removing them and selecting the features and labels used for prediction

```
crop['label'].unique()
```

```
array(['rice', 'maize', 'chickpea', 'kidneybeans', 'pigeonpeas',
       'mothbeans', 'mungbean', 'blackgram', 'lentil', 'pomegranate',
       'banana', 'mango', 'grapes', 'watermelon', 'muskmelon', 'apple',
       'orange', 'papaya', 'coconut', 'cotton', 'jute', 'coffee'],
      dtype=object)
```

```
[11] crop['label'].nunique()
```

```
22
```

```
crop.drop_duplicates().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

```
crop.describe(include="all")
```

	N	P	K	temperature	humidity	ph	rainfall	label
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	22
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	rice
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	100
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655	NaN
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389	NaN

Modules

Feature Selection

- After pre-processing the acquired data, the next step is to identify the best features. The identified best features should be able to give high efficiency
- In this step, we process available features and select the final parameters that are influential in nature. The list of final parameters are.
- Plants that are nutritious and healthy while developing have good nitrogen content in them.
- Phosphorus also encourages the growth of roots, promotes flowering, and is necessary for DNA.
- Potassium provides strength to plants and can resist diseases.

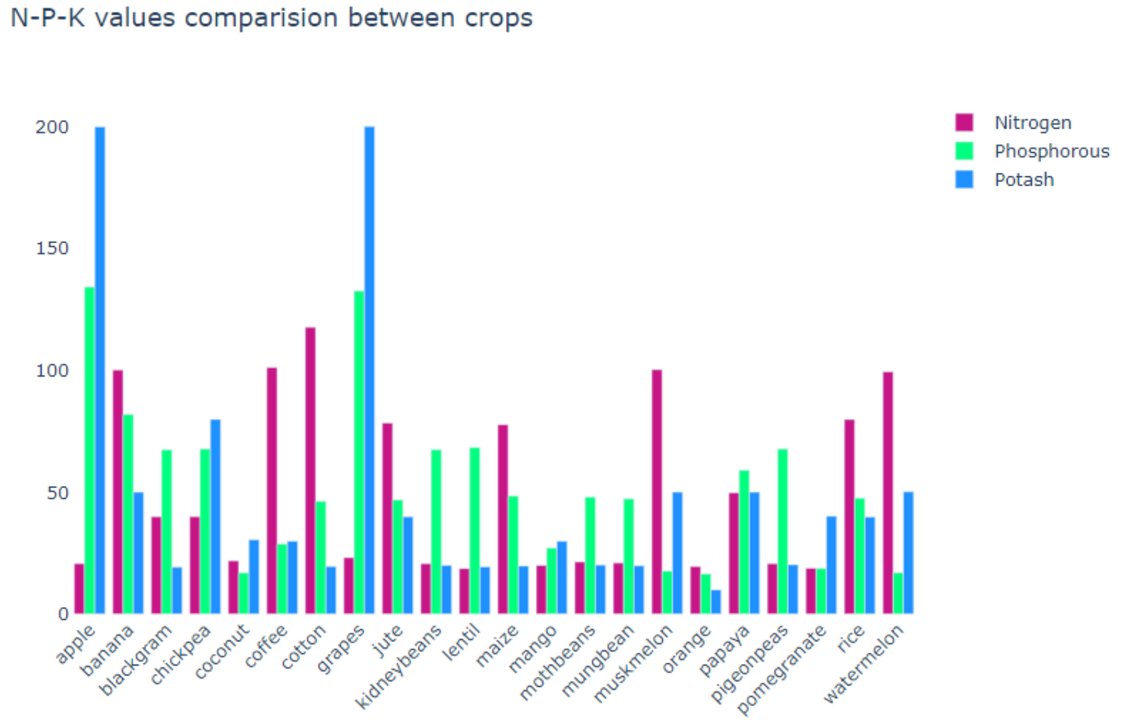
Crop	Name of the crop
pH Level	This describes the nature of the soil
Nitrogen	Amount of nitrogen present
Potassium	Amount of potassium present
Phosphorus	Amount of phosphorus present
Rainfall	Expected rainfall in millimeters
Temperature	Optimal temperature for the crop
Humidity	Water Vapour content available for the crop

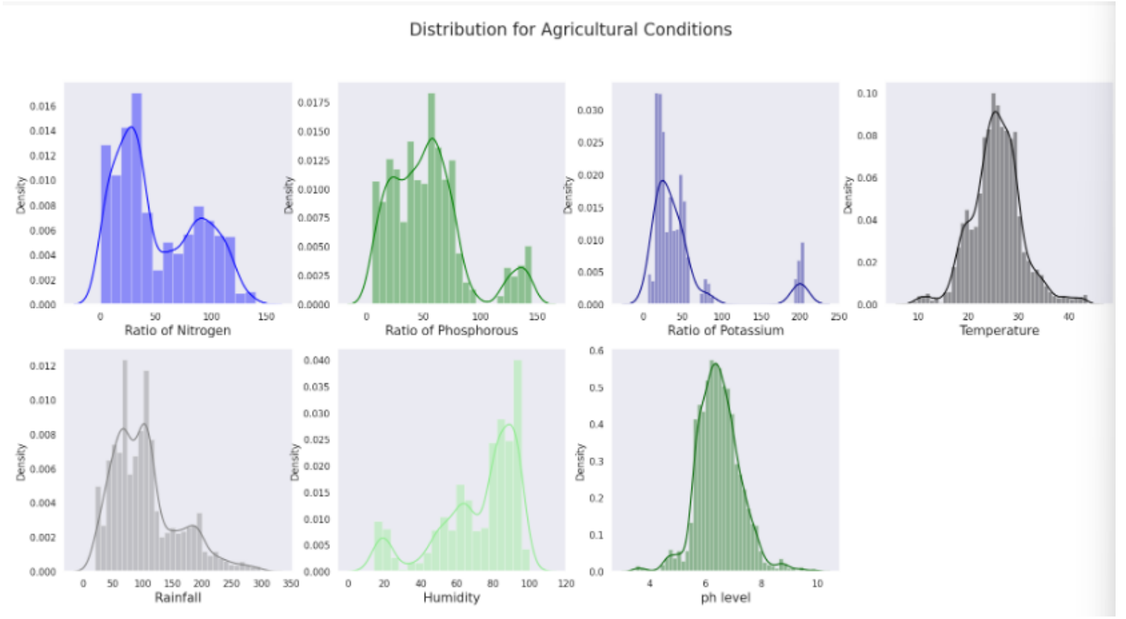
Modules

Visualization

- Data visualization is the representation of data through the use of common graphics, such as charts, plots, infographics, etc

- The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise





Modules

Model Building and Predictions

- Building a model in machine learning is creating a mathematical representation by generalizing and learning from training data. Then, the built machine learning model is applied to new data to make predictions and obtain results

- It is a crucial component of predictive analytics, a type of data analytics that uses current and historical data to forecast activity, behavior, and trends

	Real_class	Predicted_class
2121	coffee	coffee
960	pomegranate	pomegranate
952	pomegranate	pomegranate
1958	cotton	cotton
681	mungbean	mungbean

Experimental Results

Cross Validation Score - Evaluation Metrics

- Cross-validation may be a statistical procedure that is used to estimate the skill of machine learning models it gives a more accurate measure of model quality. Cross-validation results of Random Forest, Decision Tree, Logistic regression, and Gradient Boosting

```
# Cross validation score (Decision Tree)
score = cross_val_score(DecisionTree, features, target,cv=5)
score
```

```
array([0.93636364, 0.90909091, 0.91818182, 0.87045455, 0.93636364])
```

```
# Cross validation score (Random Forest)
score = cross_val_score(RF,features,target,cv=5)
score
```

```
array([0.99772727, 0.99545455, 0.99772727, 0.99318182, 0.98863636])
```

```
# Cross validation score (Logistic Regression)
score = cross_val_score(LogReg,features,target,cv=5)
score
```

```
array([0.95      , 0.96590909, 0.94772727, 0.96590909, 0.94318182])
```

```
# Cross validation score (Gradient Boosting)
score = cross_val_score(grad,features,target,cv=5)
score
```

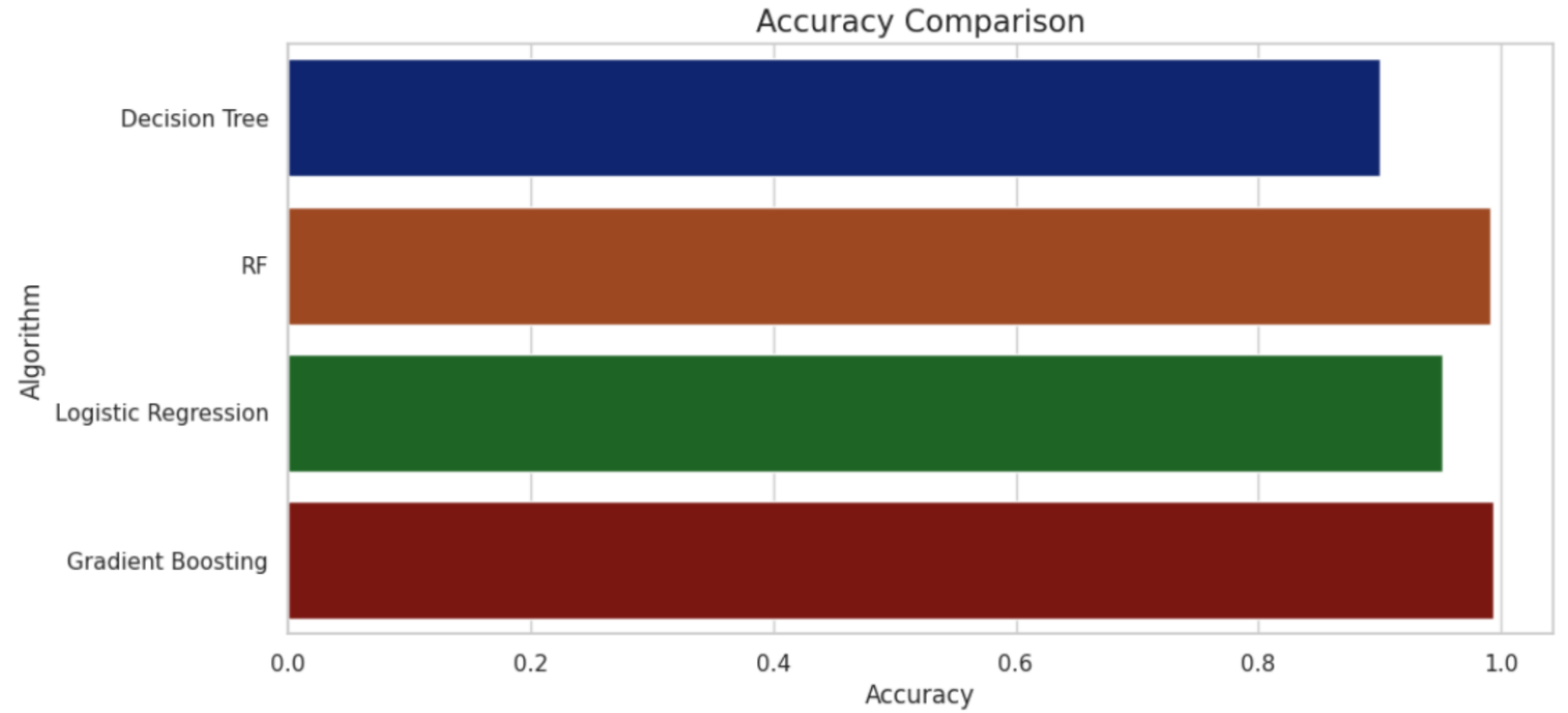
```
array([0.99090909, 0.98863636, 0.99318182, 0.99772727, 0.98636364])
```

Experimental Results

Accuracy Comparison

- Barplot displaying the accuracy comparison between the four models

<Axes: title={'center': 'Accuracy Comparison'}, xlabel='Accuracy', ylabel='Algorithm'>



	Model	Accuracy
0	Decision Tree	0.900000
1	RF	0.990909
2	Logistic Regression	0.952273
3	Gradient Boosting	0.993182

Experimental Results

Predictions

- The values for the parameters are entered and passed onto the Random forest and gradient-boosting model to get the predictions

Optimal Crop for given climatic variables

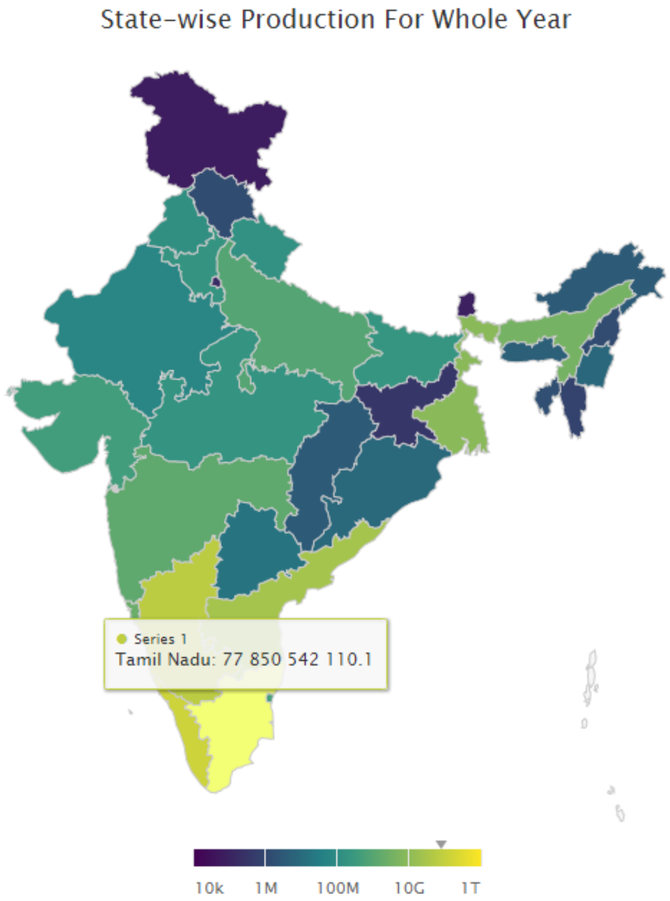
```
[70] data = np.array([[34,60,22,17.66148158,18.15302753,5.635231778,100.6711761]])
     prediction = RF.predict(data)
     print("Given Parameters best suit's for the crop : ",prediction)

     Given Parameters best suit's for the crop :  ['kidneybeans']
```

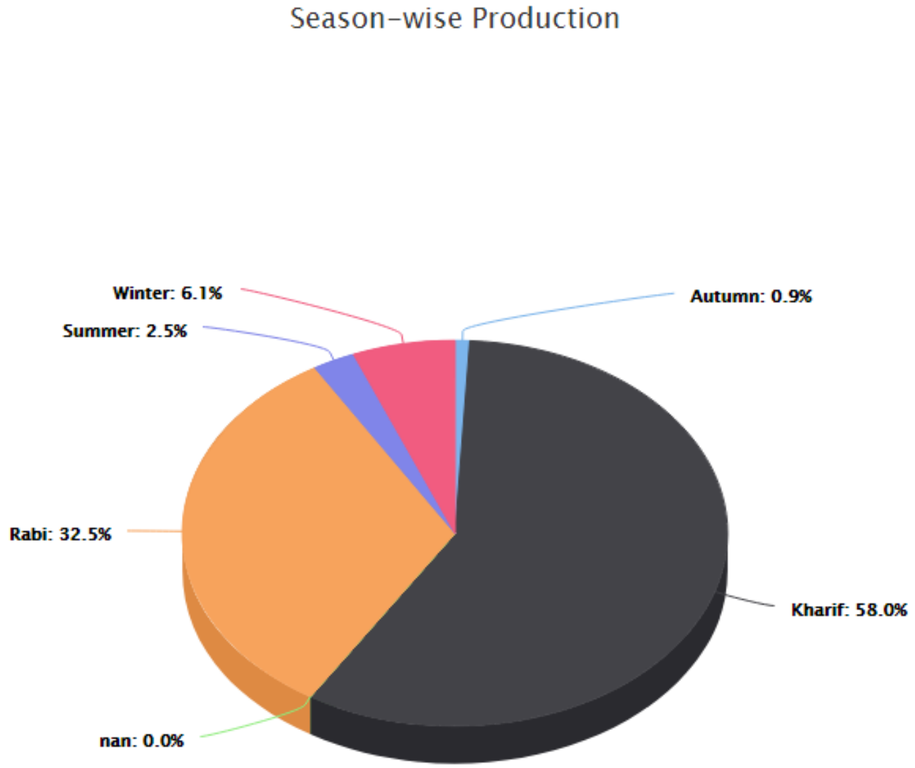
```
[94] data = np.array([[37,62,22,29.03,64.49,7.47,54.93]])
     prediction = grad.predict(data)
     print("Given Parameters best suit's for the crop : ",prediction)

     Given Parameters best suit's for the crop :  ['lentil']
```

```
Average Ratio of Nitrogen in the Soil : 50.55
Average Ratio of Phosphorous in the Soil : 53.36
Average Ratio of Potassium in the Soil : 48.15
Average Tempature in Celsius : 25.62
Average Relative Humidity in % : 71.48
Average PH Value of the soil : 6.47
Average Rainfall in mm : 103.46
```

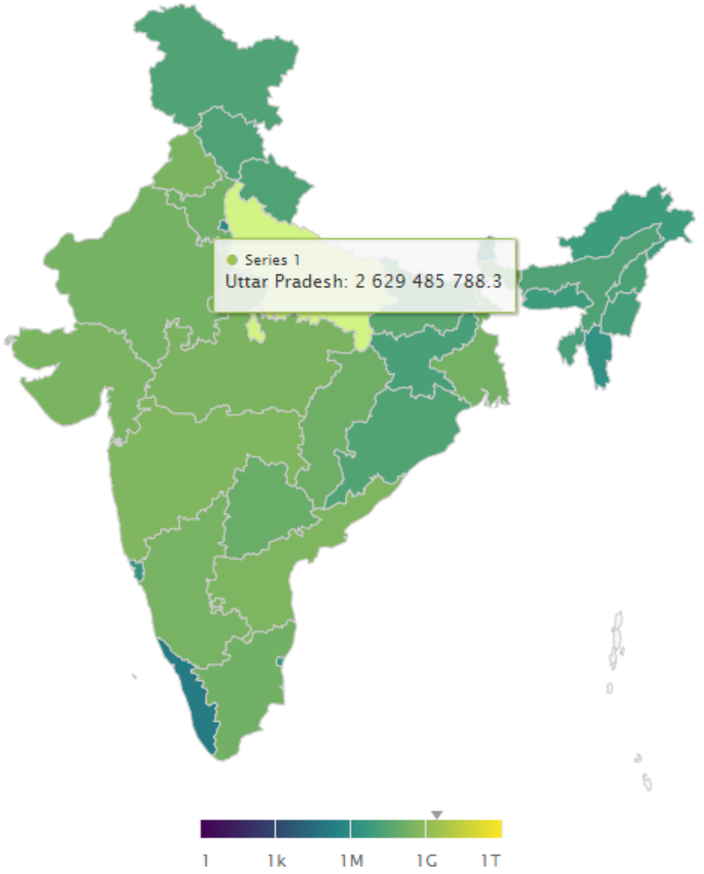
Visualizing Past Trends

- With the use of R Studio and Highcharter library and HighMaps visualizations of data are made.

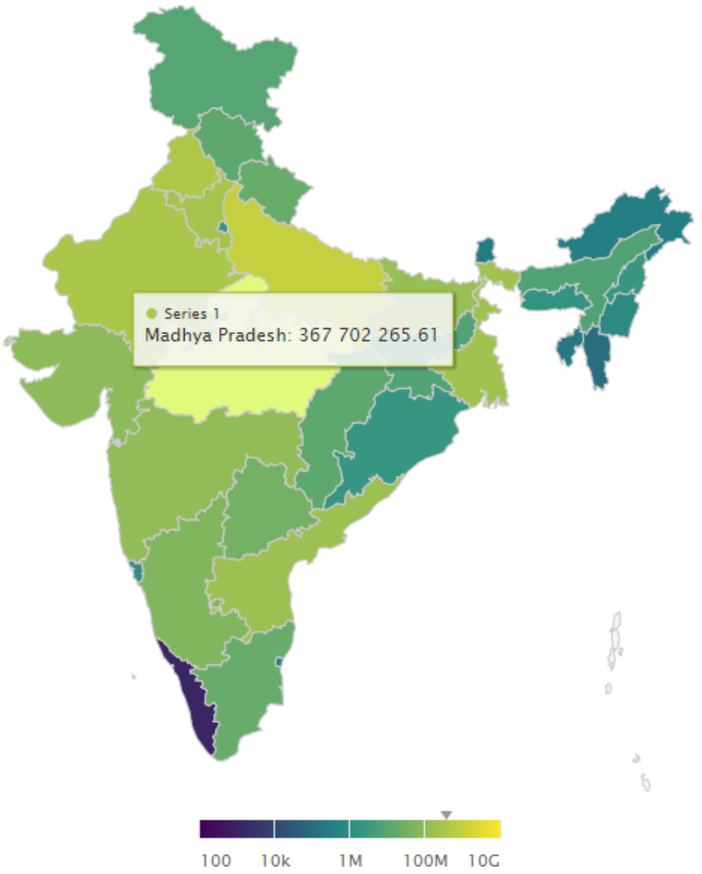




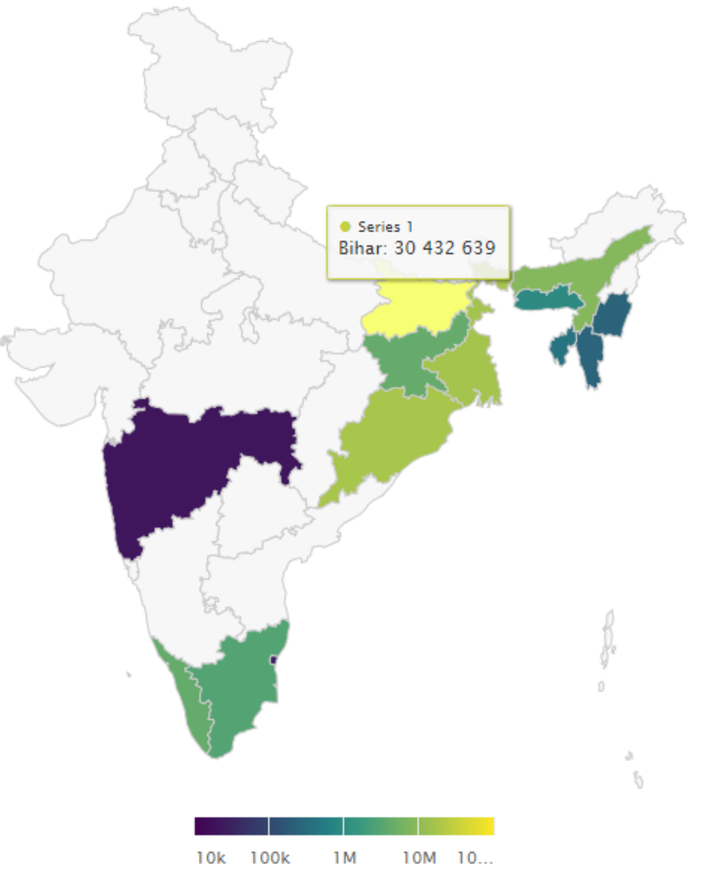
Kharif Season Production (start in June and end in October)



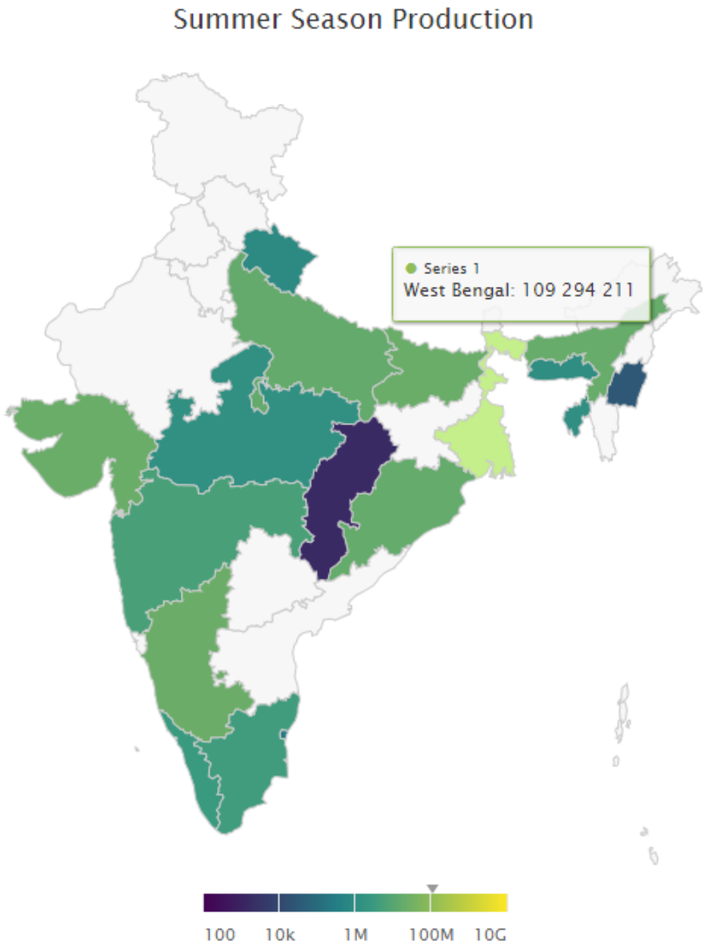
Rabi Season Production (October to March)

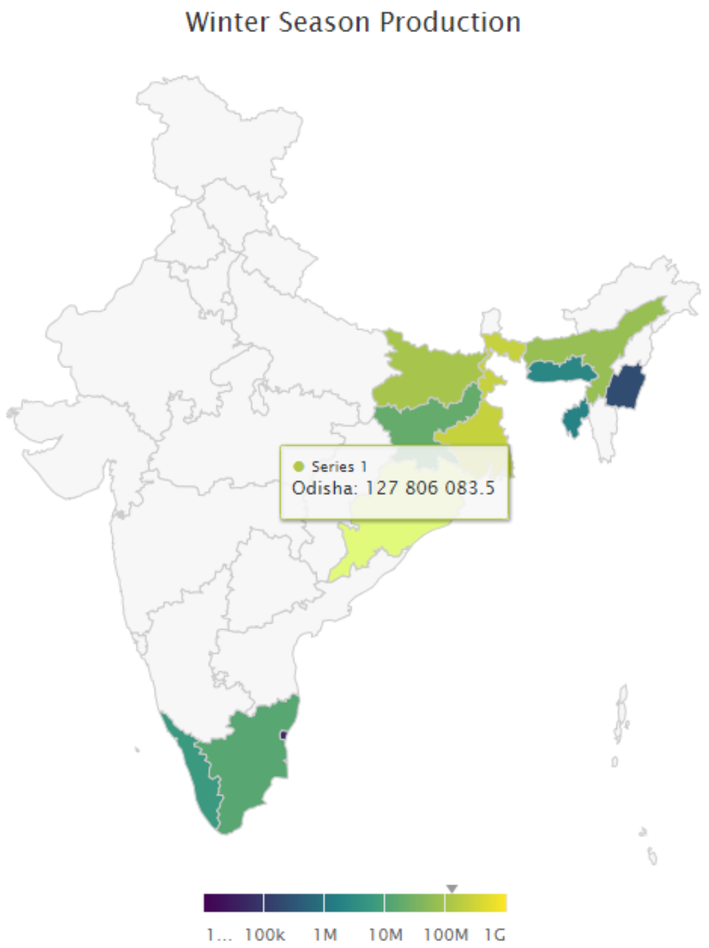


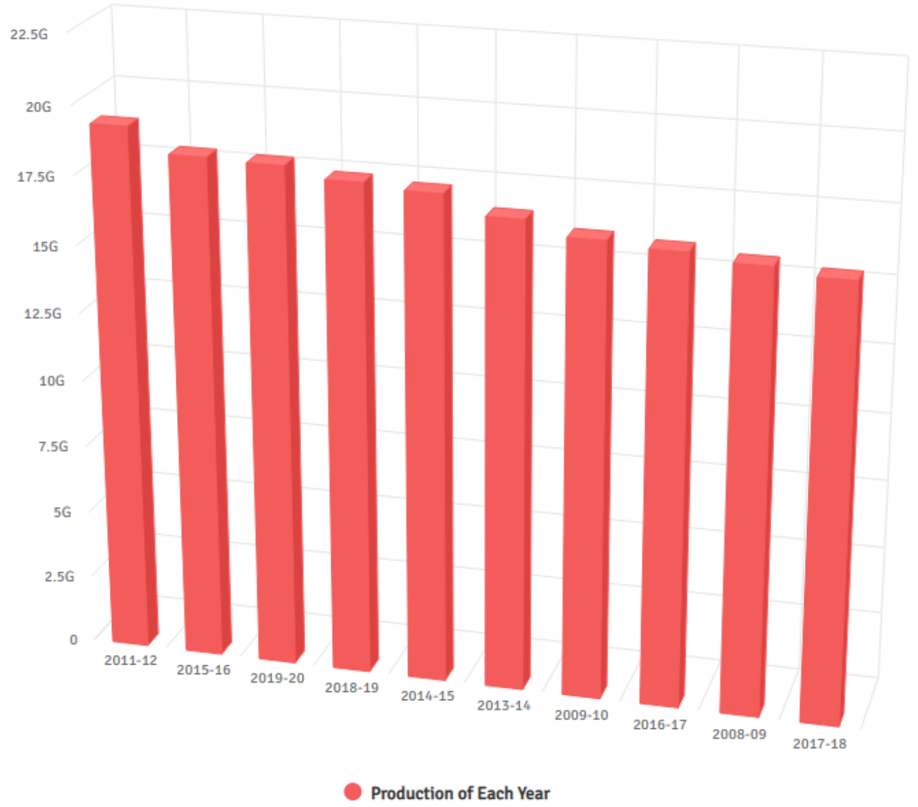
Autumn Season Production



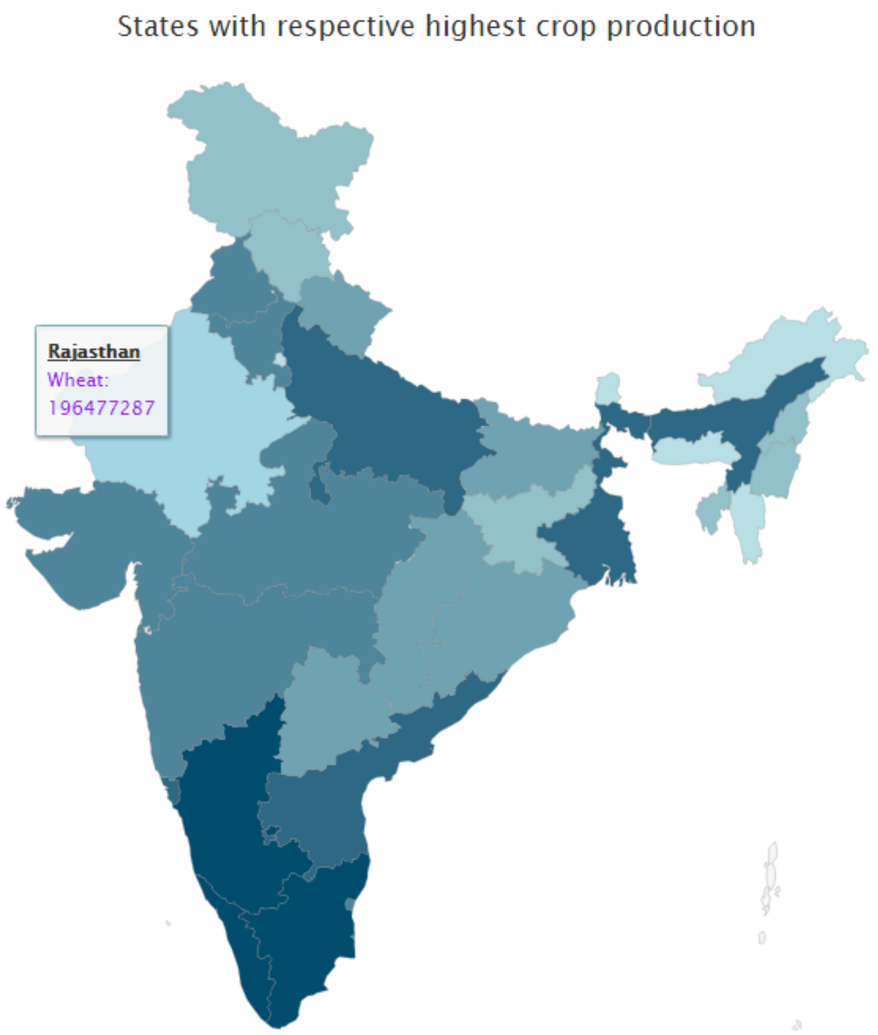
- Season-wise production and crop yield across all states in India.

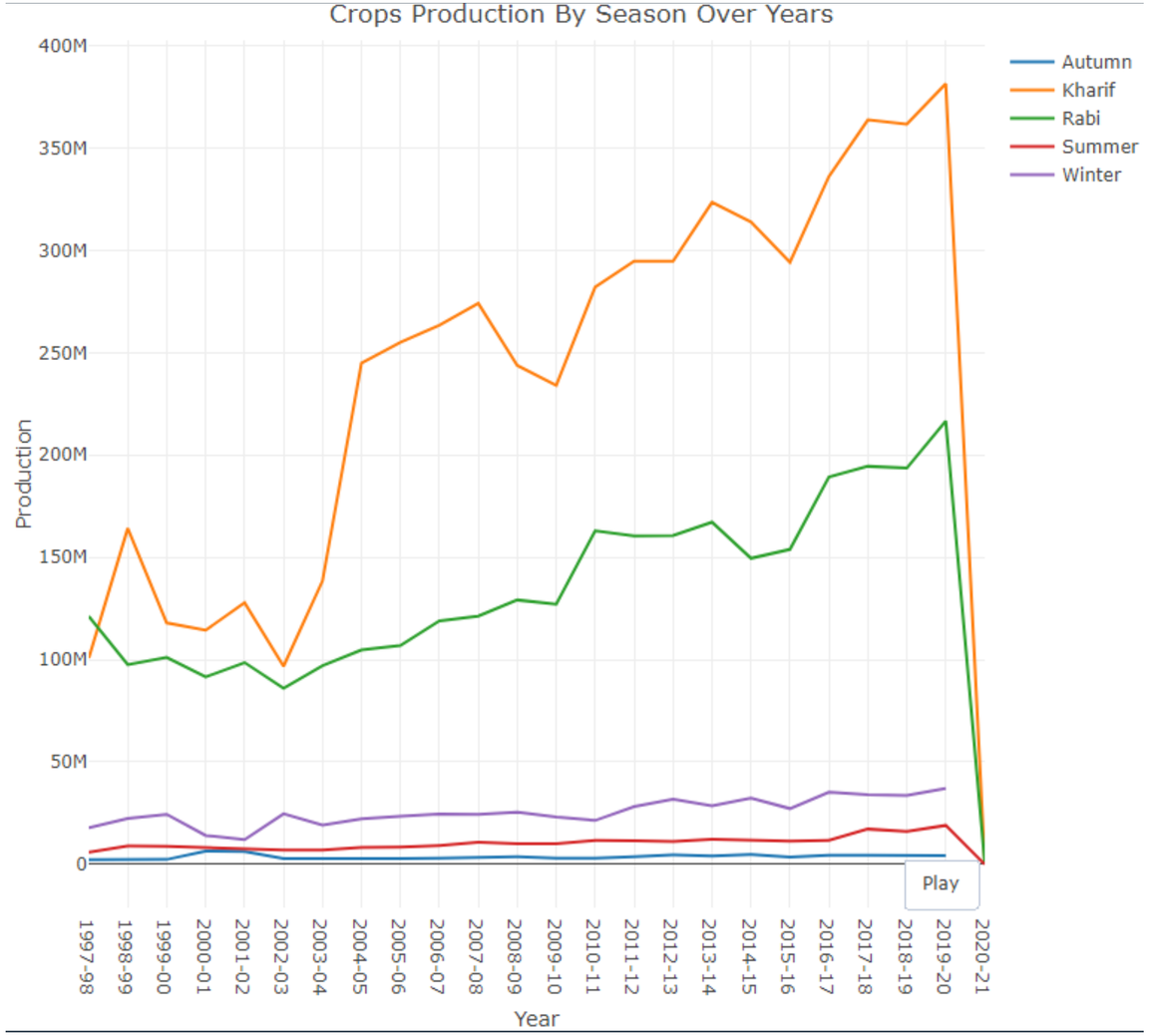






- Taking the Top ten years of production all over India and the summer and winter production

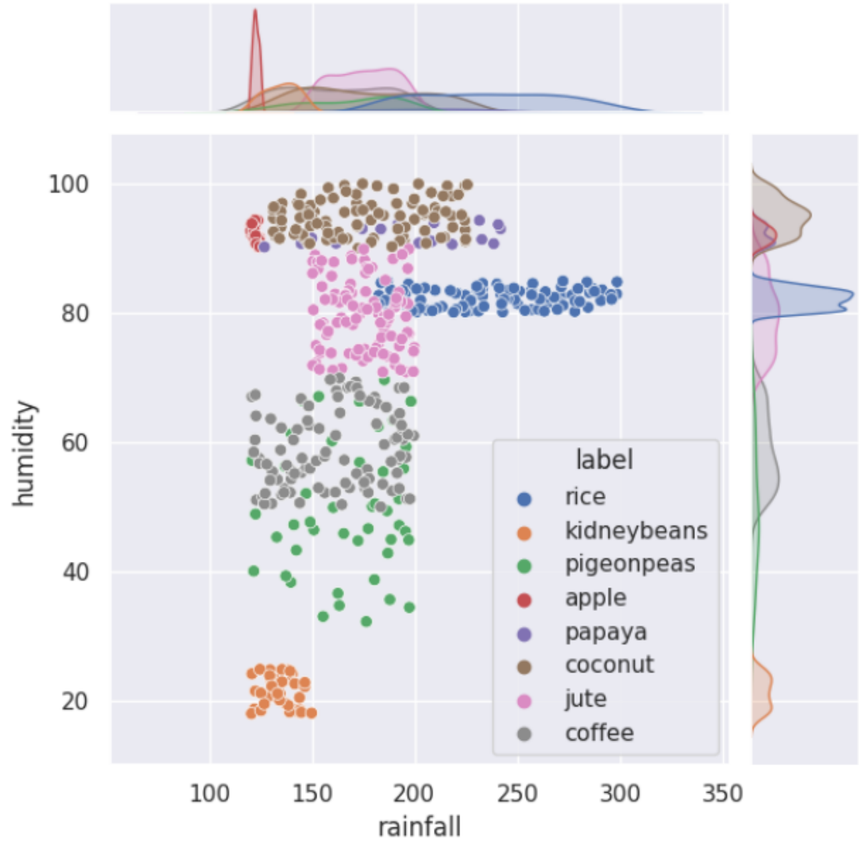




- States highest produced crop along with the production over the years across all seasons.

```
sns.jointplot(x="rainfall",y="humidity",data=crop[(crop['temperature']<30) &
                                                  (crop['rainfall']>120)],hue="label")
```

```
<seaborn.axisgrid.JointGrid at 0x7f0fe7828a00>
```



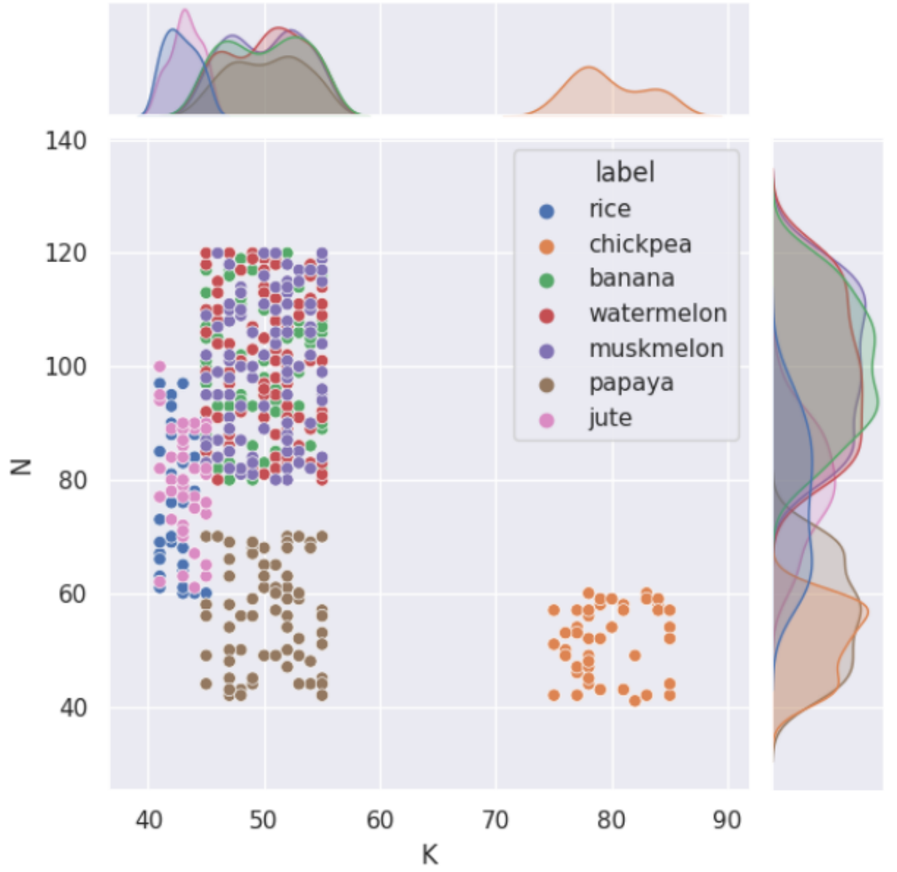
During rainy season, average rainfall is high (average 120 mm) and temperature is mildly chill (less than 30'C).

Rain affects soil moisture which affects ph of the soil. Here are the crops which are likely to be planted during this season.

- **Rice needs heavy rainfall (>200 mm) and a humidity above 80%. No wonder major rice production in India comes from East Coasts which has average of 220 mm rainfall every year!**
- **Coconut is a tropical crop and needs high humidity therefore explaining massive exports from coastal areas around the country.**

```
sns.jointplot(x="K",y="N",data=crop[(crop['N']>40)&(crop['K']>40)],hue="label")
```

```
<seaborn.axisgrid.JointGrid at 0x7f59ae3e5910>
```



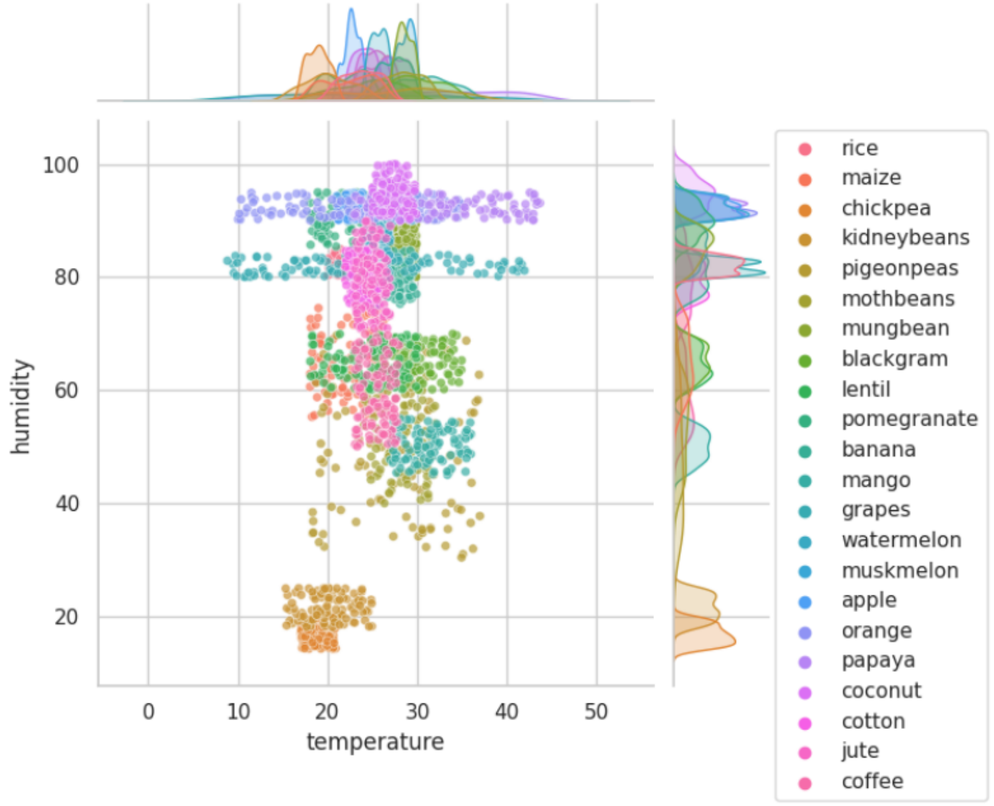
This graph correlates with average potassium (K) and average nitrogen (N) value (both>50).

These soil ingredients direcly affects nutrition value of the food.

Fruits which have high nutrients typically has consistent potassium values.

```
sns.jointplot(x="temperature",y="humidity",data=crop,hue='label',s=25,alpha=0.7)
plt.legend(loc='upper left',bbox_to_anchor=(1.2, 1))
```

```
<matplotlib.legend.Legend at 0x7f0fdc8f9cd0>
```



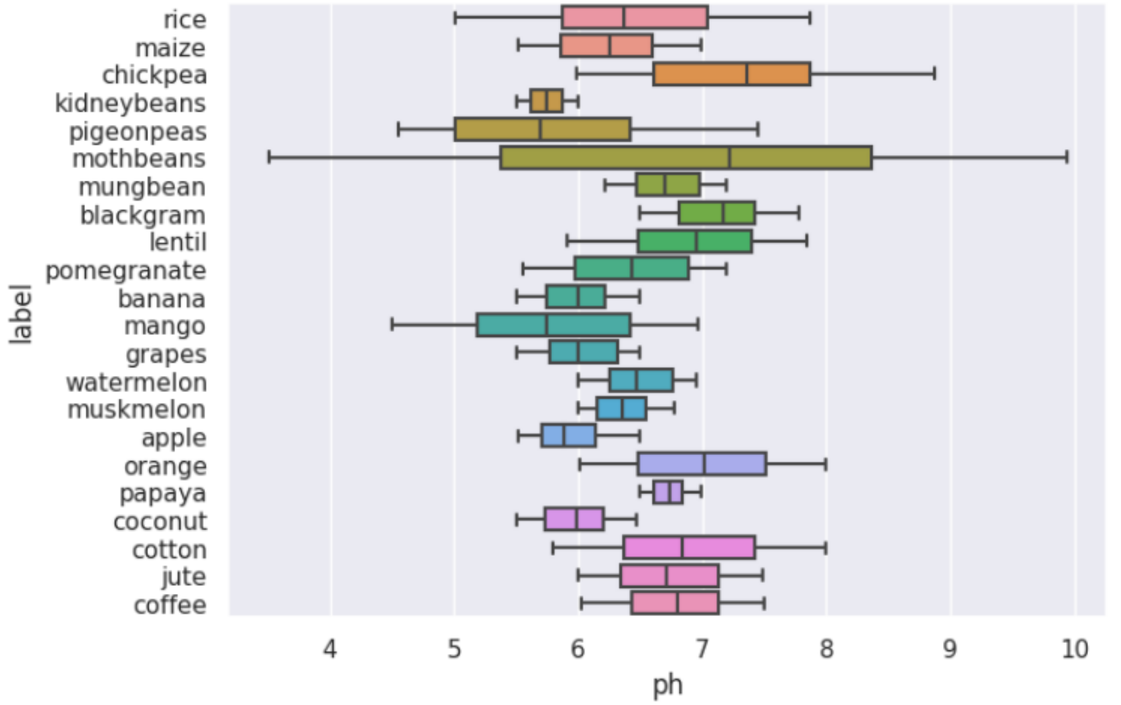
This graph correlates with temperature and humidity values.

Here the optimal temperature ranges from 10 to 40 for

most of the crops and the humidity ranges from 40 to 100

```
sns.boxplot(y='label',x='ph',data=crop)
```

```
<Axes: xlabel='ph', ylabel='label'>
```

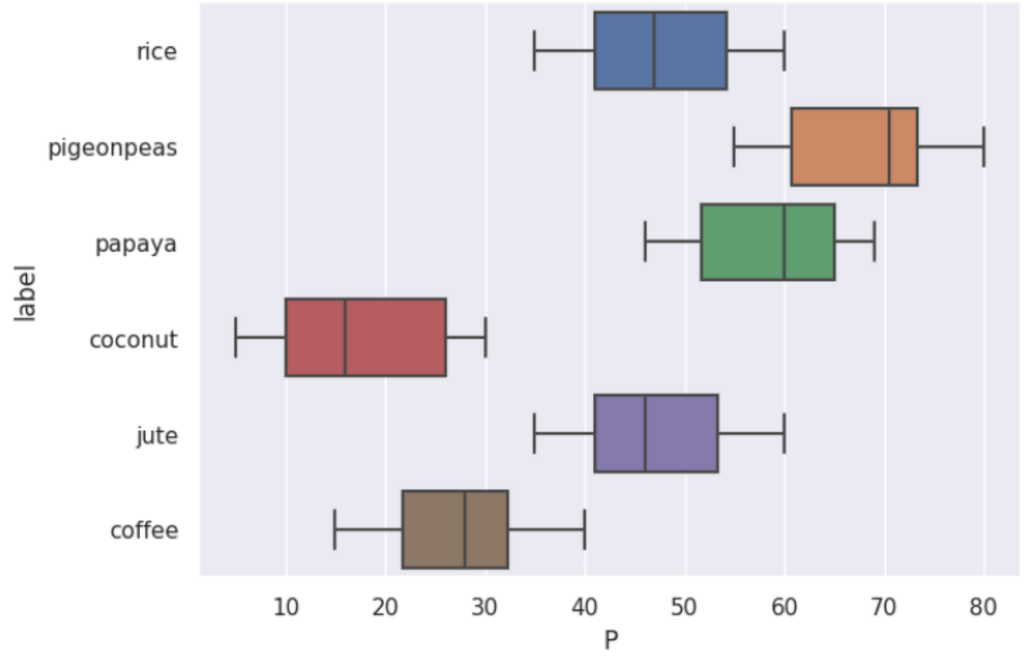


We can see ph values are critical when it comes to soil. A stability between 6 and 7 is preffered

- Analyzing Phosphorous levels comparing with humidity and rainfall

```
[23] sns.boxplot(y='label',x='P',data=crop[crop['rainfall']>150])
```

```
<Axes: xlabel='P', ylabel='label'>
```



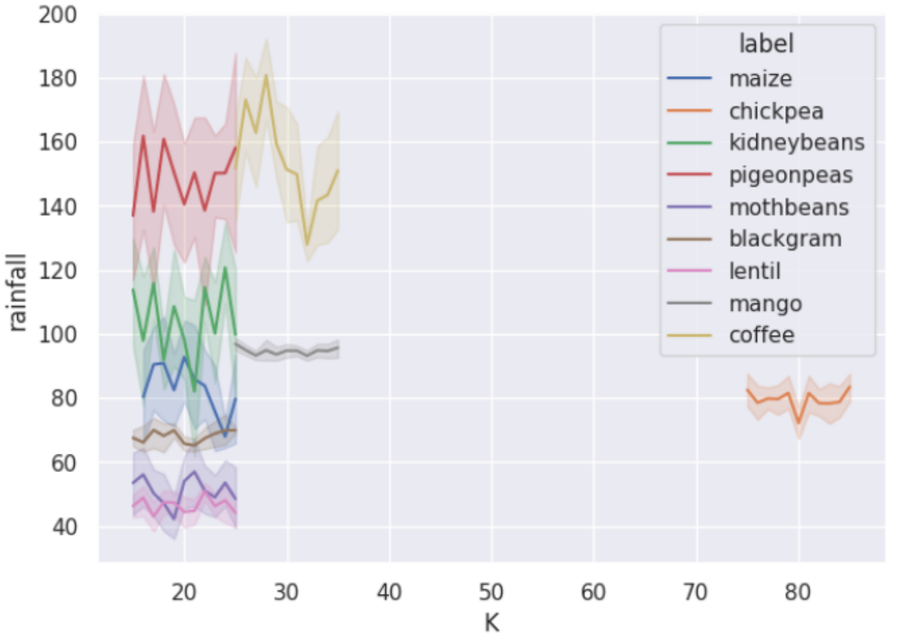
Another interesting analysis where Phosphorous levels are quite differentiable when it rains heavily (above 150 mm).

Further analyzing phosphorous levels.

When humidity is less than 65, almost same phosphor levels (approx 14 to 25)
are required for 6 crops which could be grown just based on the amount of rain expected over the next few weeks.

```
sns.lineplot(data = crop[(crop['humidity']<65)], x = "K", y = "rainfall",hue="label")
```

```
<Axes: xlabel='K', ylabel='rainfall'>
```



Conclusion

- Decision tree regression, Gradient Boosting, Random Forest, and Logistic Regression techniques are implemented on the input data to assess the best performance-yielding method.

- These methods are compared using performance metrics. According to the analyses of cross-evaluation metrics, all four algorithms work well, but gradient boosting gives a slightly better accuracy score on test data than the other four models.

- The proposed work can also be extended to analyze the climatic conditions and other factors for the crop and to increase crop production.

- With the help of the high charter library visualizations on maps are made easier and we can see the insights the visualizations provide helps in making informed decisions over the coming years.

Referrences

- https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80
- https://link.springer.com/chapter/10.1007/978-3-319-14142-8_1
- https://link.springer.com/chapter/10.1007/978-3-319-14142-8_6
- https://stackoverflow.com/questions/58983528/how-to-find-optimal-parametrs-for-dbscan
- https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd
- https://blog.exploratory.io/visualizing-k-means-clustering-results-to-understand-the-characteristics-of-clusters-better-b0226fb3d
- https://scikit-learn.org/stable/modules/model_evaluation.html

Appendix

Summary for all the crops

```
print("Average Ratio of Nitrogen in the Soil : {0:.2f}".format(crop['N'].mean()))
print("Average Ratio of Phosphorous in the Soil : {0:.2f}".format(crop['P'].mean()))
print("Average Ratio of Potassium in the Soil : {0:.2f}".format(crop['K'].mean()))
print("Average Tempature in Celsius : {0:.2f}".format(crop['temperature'].mean()))
print("Average Relative Humidity in % : {0:.2f}".format(crop['humidity'].mean()))
print("Average PH Value of the soil : {0:.2f}".format(crop['ph'].mean()))
print("Average Rainfall in mm : {0:.2f}".format(crop['rainfall'].mean()))
```

```
Average Ratio of Nitrogen in the Soil : 50.55
Average Ratio of Phosphorous in the Soil : 53.36
Average Ratio of Potassium in the Soil : 48.15
Average Tempature in Celsius : 25.62
Average Relative Humidity in % : 71.48
Average PH Value of the soil : 6.47
Average Rainfall in mm : 103.46
```

crops rice

```
----------------------------------------
Statistics for Nitrogen
Minimum Nitrigen required : 60
Average Nitrogen required : 79.89
Maximum Nitrogen required : 99
----------------------------------------
Statistics for Phosphorous
Minimum Phosphorous required : 35
Average Phosphorous required : 47.58
Maximum Phosphorous required : 60
----------------------------------------
Statistics for Potassium
Minimum Potassium required : 35
Average Potassium required :
 39.87
Maximum Potassium required : 45
----------------------------------------
Statistics for Temperature
Minimum Temperature required : 20.05
Average Temperature required : 23.69
Maximum Temperature required : 26.93
----------------------------------------
Statistics for Humidity
Minimum Humidity required : 80.12
Average Humidity required : 82.27
Maximum Humidity required : 84.97
----------------------------------------
Statistics for PH
Minimum PH required : 5.01
Average PH required : 6.43
Maximum PH required : 7.87
----------------------------------------
Statistics for Rainfall
Minimum Rainfall required : 182.56
Average Rainfall required : 236.18
Maximum Rainfall required : 298.56
```

Appendix

Average Requirement for each crops with average conditions

conditions N

```
Average Value for N is 50.55
----------------------------------------
Rice : 79.89
Black Grams : 40.02
Banana : 100.23
Jute : 78.40
Coconut : 21.98
Apple : 20.80
Papaya : 49.88
Muskmelon : 100.32
Grapes : 23.18
Watermelon : 99.42
Kidney Beans: 20.75
Mung Beans : 20.99
Oranges : 19.58
Chick Peas : 40.09
Lentils : 18.77
Cotton : 117.77
Maize : 77.76
Moth Beans : 21.44
Pigeon Peas : 20.73
Mango : 20.07
Pomegranate : 18.87
Coffee : 101.20
```

```
@interact
def compare(conditions = ['N','P','K','temperature','ph','humidity','rainfall']):
    print("Crops which require greater than average", conditions,'\n')
    print(crop[crop[conditions] > crop[conditions].mean()]['label'].unique())
    print("----------------------------------------------")
    print("Crops which require less than average", conditions,'\n')
    print(crop[crop[conditions] <= crop[conditions].mean()]['label'].unique())
```

conditions temperature

```
Crops which require greater than average temperature

['rice' 'maize' 'pigeonpeas' 'mothbeans' 'mungbean' 'blackgram' 'lentil'
 'banana' 'mango' 'grapes' 'watermelon' 'muskmelon' 'orange' 'papaya'
 'coconut' 'cotton' 'jute' 'coffee']
----------------------------------------------
Crops which require less than average temperature

['rice' 'maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'mothbeans'
 'blackgram' 'lentil' 'pomegranate' 'banana' 'grapes' 'watermelon' 'apple'
 'orange' 'papaya' 'coconut' 'cotton' 'jute' 'coffee']
```

Appendix

Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score
import sklearn.metrics as metrics

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
```

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*100)

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

```
Logistic Regression's Accuracy is:  95.22727272727273
```

Random Forest

```
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*100)

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

```
RF's Accuracy is:  99.0909090909091
```

Gradient Boosting

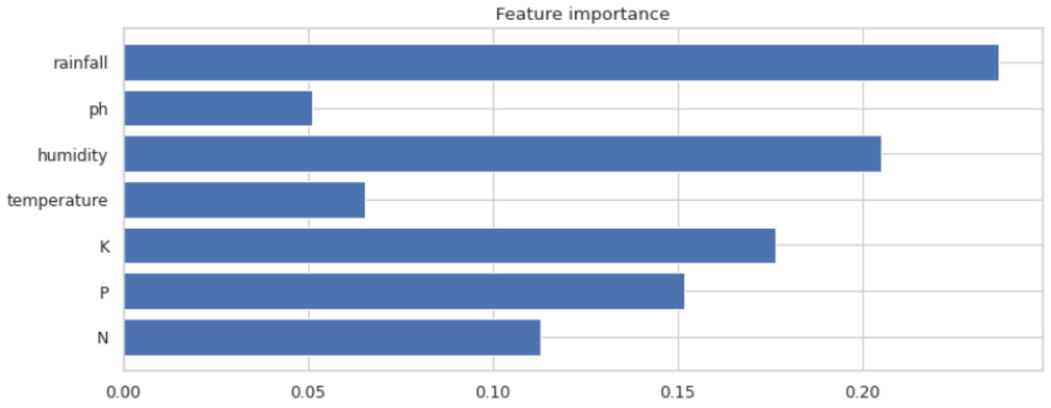
```
from sklearn.ensemble import GradientBoostingClassifier
grad = GradientBoostingClassifier().fit(Xtrain, Ytrain)

predicted_values = grad.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Gradient Boosting')
print("Gradient Boosting Accuracy is: ", x*100)

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

```
Gradient Boosting Accuracy is:  99.31818181818181
```

Appendix

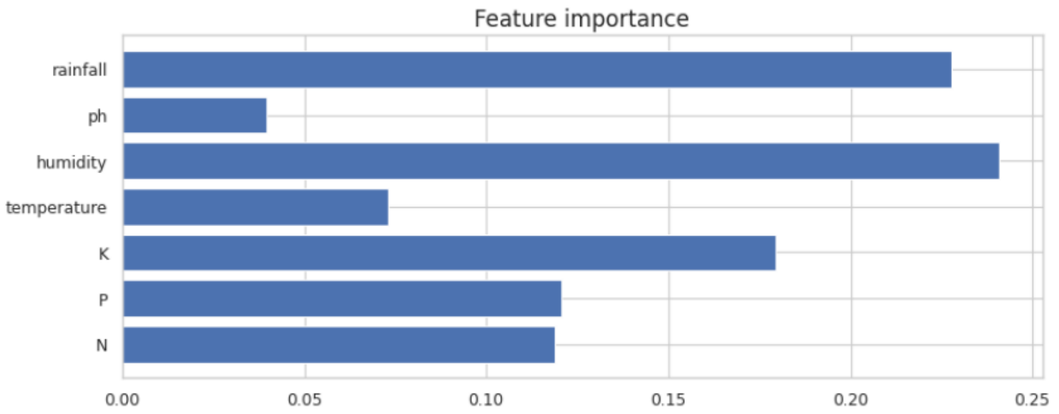
Feature Importance taken for Random forest model as we can see more importance is given to rainfall and humidity respectively

```
[65] plt.figure(figsize=(10,4), dpi=80)
     c_features = len(Xtrain.columns)
     plt.barh(range(c_features), RF.feature_importances_)
     plt.title("Feature importance")
     plt.yticks(np.arange(c_features), Xtrain.columns)
     plt.show()
```



Feature Importance taken for Gradient Boosting model as we can see more importance is given to humidity and then rainfall respectively

```
plt.figure(figsize=(10,4), dpi=80)
c_features = len(Xtrain.columns)
plt.barh(range(c_features), grad.feature_importances_)
plt.title("Feature importance",fontsize=15)
plt.yticks(np.arange(c_features), Xtrain.columns)
plt.show()
```



Worklog

Day	Date	Task done
Day 1	17-02-2023	Collecting Research Paper on Project Topic
Day 2	20-02-2023	Collecting Research Paper on Project Topic
Day 3	22-02-2023	Collecting Research Paper on Project Topic
Day 4	24-02-2023	Collecting Research Paper on Project Topic
Day 5	27-02-2023	Collecting Research Paper on Project Topic
Day 6	01-03-2023	Gathering data sources

Worklog

Day	Date	Task done
Day 7	03-03-2023	Understanding data mining techniques
Day 8	06-03-2023	Looking into PAM,CLARA, DBSCAN
Day 9	08-03-2023	Looking to improve problem statement
Day 10	10-03-2023	Understanding challenges in Climate changes
Day 11	17-03-2023	Finalizing Dataset
Day 12	22-03-2023	Data preparation

Worklog

Day	Date	Task done
Day 13	05-04-2023	Data preprocessing
Day 14	07-04-2023	Clustering Techniques
Day 15	12-04-2023	Visualization on dataset
Day 16	14-04-2023	Model Building
Day 17	15-04-2023	Model Predictions
Day 18	16-04-2023	Working with Tableau

Worklog

Day	Date	Task done
Day 19	17-04-2023	Working with Rstudio
Day 20	18-04-2023	Hcmap visualizations
Day 21	19-04-2023	ppt preparation
Day 22	20-04-2023	Finalizing visualizations on the data
Day 23	21-04-2023	ppt finalizing
Day 24	22-04-2023	ppt finalizing

Thank You