PDS PROJECT REPORT

Bhanu Pranaswi Sai P
S20210020309

Data Driven Crop Analysis and Prediction

NUMERICAL DATA ANALYSIS

Abstract:

The Crop Recommendation System represents a pioneering approach at the intersection of agricultural tradition and technological innovation, redefining crop selection methodologies for farmers. This data-driven project harnesses the power of advanced analytics to provide precise crop recommendations by considering critical soil mineral compositions, including potassium, nitrogen, and phosphorous, alongside environmental variables such as humidity, temperature, and rainfall. Through comprehensive analysis and integration of these factors, the system empowers farmers with tailored guidance, facilitating informed decision-making in crop selection. This paper delves into the significance of these essential elements and their transformative impact on agricultural practices.

Index Terms—Exploratory Data Analysis, Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Random Forest, Decision Trees, Data preprocessing, Principal Component Analysis, Feature scaling, Crop recommendation.

Introduction:

The Crop Recommendation System represents a pioneering approach at the intersection of agricultural tradition and technological innovation, redefining crop selection methodologies for farmers. This data-driven project harnesses the power of advanced analytics to provide precise crop recommendations by considering critical soil mineral compositions, including potassium, nitrogen, and phosphorous, alongside environmental variables such as humidity, temperature, and rainfall. Through comprehensive analysis and integration of these factors, the system empowers farmers with tailored guidance, facilitating informed decision-making in crop selection. This paper delves into the significance of these essential elements and their transformative impact on agricultural practices.

Significance of crop recommendation system:

The significance of the Crop Recommendation System lies in its ability to address several critical challenges faced by farmers and agricultural professionals:

1. ***Improved Decision-Making***: By leveraging machine learning algorithms and analyzing a wide range of environmental and soil factors, the system provides farmers with data-driven insights to make informed decisions about crop selection. This helps optimize yields and maximize profitability.

2. ***Enhanced Productivity***: By recommending crops that are best suited to the local soil and climate conditions, the system enables farmers to improve productivity and

efficiency in their agricultural practices. This can lead to higher crop yields and better utilization of resources.

3. ***Sustainability***: By promoting the cultivation of crops that are well-adapted to local environmental conditions, the Crop Recommendation System supports sustainable agriculture practices. This can help reduce the use of chemical inputs, minimize environmental impact, and preserve natural resources.

4***. Risk Mitigation***: By considering various environmental factors such as temperature, humidity, and rainfall, the system helps farmers mitigate risks associated with adverse weather conditions or climate change. By diversifying crop portfolios and recommending resilient crop varieties, farmers can better cope with fluctuations in weather patterns.

5. ***Access to Expertise***: The Crop Recommendation System serves as a valuable tool for farmers who may not have access to agronomic expertise or agricultural extension services. By providing personalized recommendations based on data analysis, the system democratizes access to agricultural knowledge and expertise.

Objectives and Methodology:

This project aims to develop a comprehensive Crop Recommendation System by integrating exploratory data analysis (EDA) and various machine learning techniques. Objectives include conducting in-depth EDA to identify key predictors such as soil composition, climate factors, and historical crop performance. The methodology involves preprocessing the data to handle missing values and outliers, performing thorough EDA to gain insights into dataset characteristics, and developing predictive models using techniques such as decision trees, random forest, and gradient boosting. Challenges such as class imbalance and feature selection will be addressed through appropriate techniques. Advanced preprocessing methods like outlier handling and feature scaling will be employed to ensure model robustness.

This project aims to generate actionable insights for farmers and agricultural professionals to optimize crop selection, improve productivity, and promote sustainable agricultural practices. Through systematic analysis and model development, this research contributes to the advancement of crop recommendation systems and enhances our understanding of data-driven agriculture.

Broader Implications:

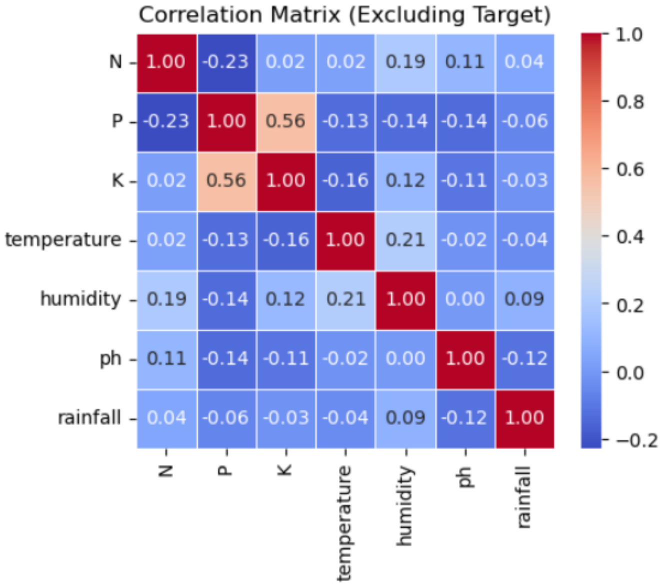
The broader implications of the Crop Recommendation System extend beyond individual farms or agricultural regions. By leveraging data-driven insights to optimize crop selection and cultivation practices, this system contributes to larger-scale agricultural sustainability and food security. Moreover, it empowers farmers with knowledge and tools to adapt to changing environmental conditions, mitigate risks associated with climate change, and enhance productivity. Ultimately, the widespread adoption of such technology has the potential to revolutionize global agriculture, fostering resilience, efficiency, and equitable access to nutritious food resources.

Concepts Used:

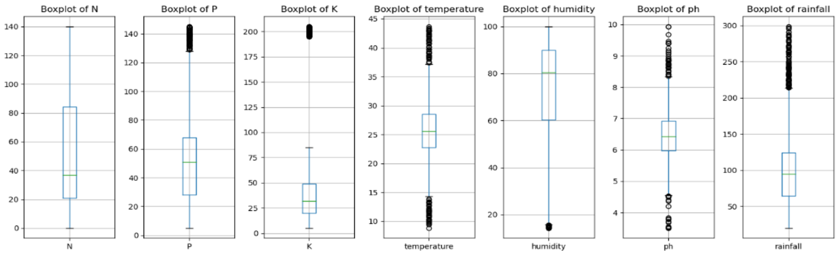
Exploratory Data Analysis:

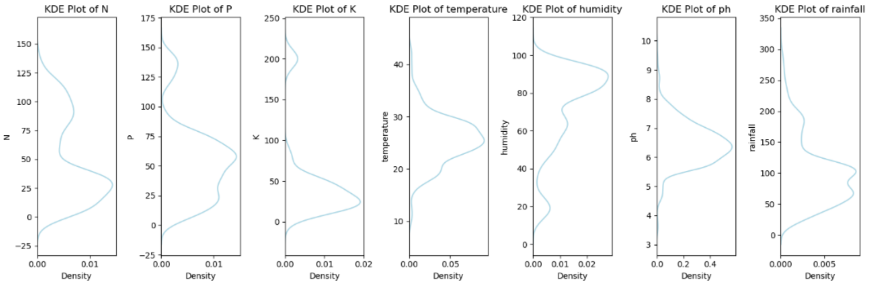
Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, aimed at gaining insights and understanding the underlying patterns and relationships within the dataset.

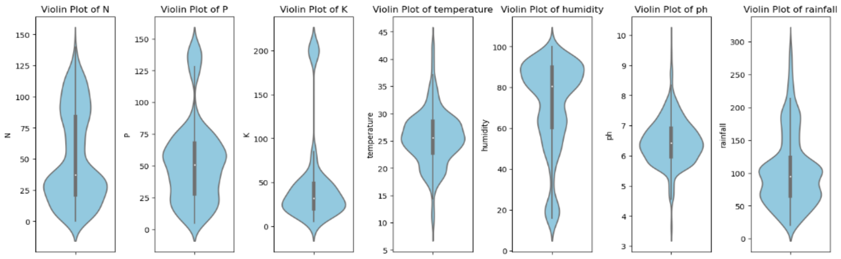
1. ***Identifying Patterns:*** Identified the patterns and trends in the dataset, such as correlations between environmental factors like temperature, humidity, and rainfall, and their impact on crop growth and yield.



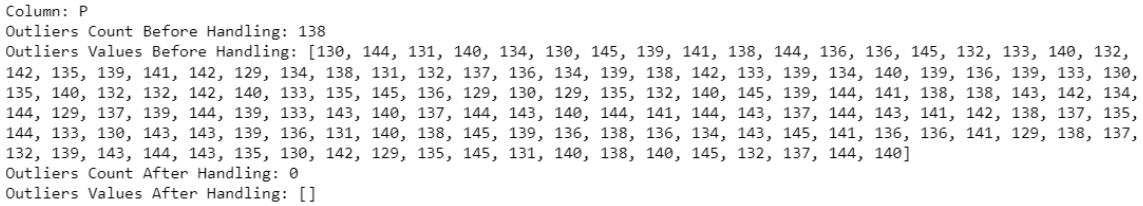
2. ***Detecting Outliers:*** Found out the outliers or anomalies in the data, which may indicate errors in data collection or provide valuable insights into unique environmental conditions that affect crop performance, hence use techniques to handle outliers.





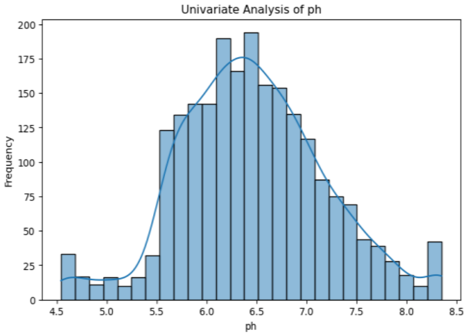


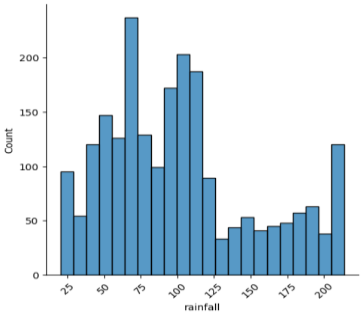
Handling outliers:

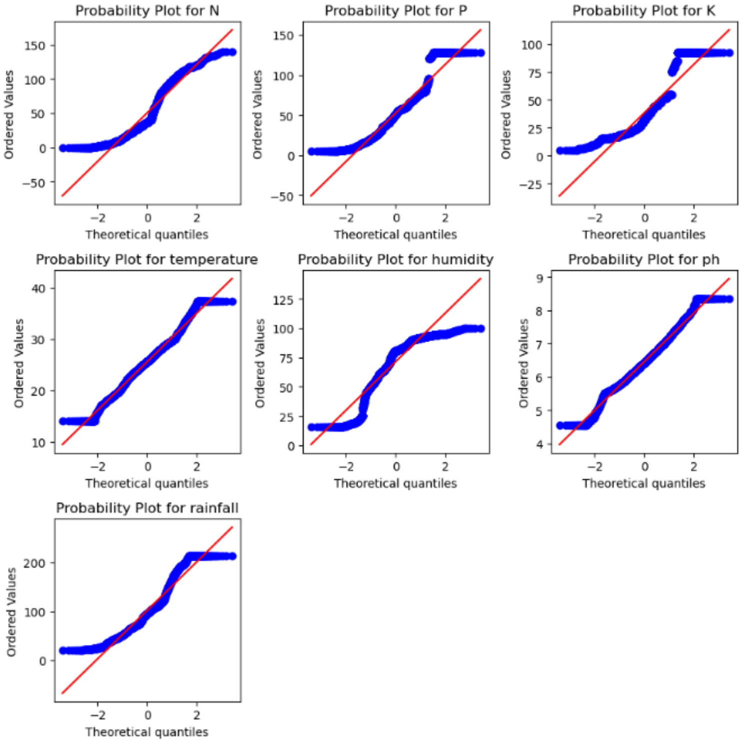


```
Column: P
Outliers Count Before Handling: 138
Outliers Values Before Handling: [130, 144, 131, 140, 134, 130, 145, 139, 141, 138, 144, 136, 136, 145, 132, 133, 140, 132,
142, 135, 139, 141, 142, 129, 134, 138, 131, 132, 137, 136, 134, 139, 138, 142, 133, 139, 134, 140, 139, 136, 139, 133, 130,
135, 140, 132, 132, 142, 140, 133, 135, 145, 136, 129, 130, 129, 135, 132, 140, 145, 139, 144, 141, 138, 138, 143, 142, 134,
144, 129, 137, 139, 144, 139, 133, 143, 140, 137, 144, 143, 140, 144, 141, 144, 143, 137, 144, 143, 141, 142, 138, 137, 135,
144, 133, 130, 143, 143, 139, 136, 131, 140, 138, 145, 139, 136, 138, 136, 134, 143, 145, 141, 136, 136, 141, 129, 138, 137,
132, 139, 143, 144, 143, 135, 130, 142, 129, 135, 145, 131, 140, 138, 140, 145, 132, 137, 144, 140]
Outliers Count After Handling: 0
Outliers Values After Handling: []
```

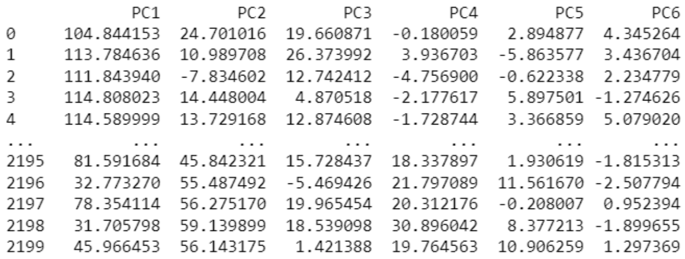
3. ***Understanding Data Distribution*:** Understood the distribution of data, allowing for a better understanding of the range and variability of key variables such as soil pH, nutrient levels, and crop yield.





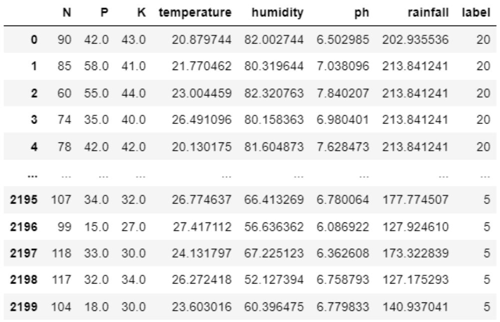


4. ***Feature Selection*****:** Selected features for modelling by identifying the most influential variables that contribute to crop recommendation. Used PCA which helped in prioritizing the factors that have a significant impact on crop suitability and productivity.



```
             PC1        PC2        PC3        PC4        PC5       PC6
0     104.844153  24.701016  19.660871  -0.180059   2.894877  4.345264
1     113.784636  10.989708  26.373992   3.936703  -5.863577  3.436704
2     111.843940  -7.834602  12.742412  -4.756900  -0.622338  2.234779
3     114.808023  14.448004   4.870518  -2.177617   5.897501 -1.274626
4     114.589999  13.729168  12.874608  -1.728744   3.366859  5.079020
...          ...        ...        ...        ...        ...       ...
2195   81.591684  45.842321  15.728437  18.337897   1.930619 -1.815313
2196   32.773270  55.487492  -5.469426  21.797089  11.561670 -2.507794
2197   78.354114  56.275170  19.965454  20.312176  -0.208007  0.952394
2198   31.705798  59.139899  18.539098  30.896042   8.377213 -1.899655
2199   45.966453  56.143175   1.421388  19.764563  10.906259  1.297369
```

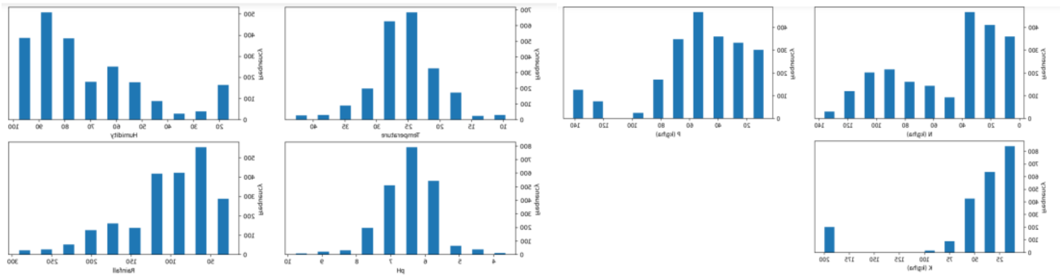
5. ***Data Preprocessing*****:** Performed data preprocessing steps such as handling missing values, dealing with categorical variables, and scaling numerical features. It ensures that the data is clean, standardized, and suitable for analysis and modelling.

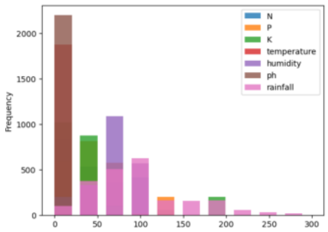


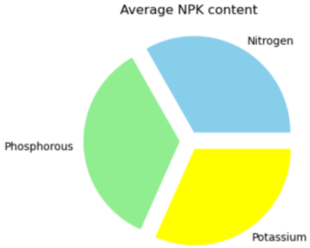
	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42.0	43.0	20.879744	82.002744	6.502985	202.935536	20
1	85	58.0	41.0	21.770462	80.319644	7.038096	213.841241	20
2	60	55.0	44.0	23.004459	82.320763	7.840207	213.841241	20
3	74	35.0	40.0	26.491096	80.158363	6.980401	213.841241	20
4	78	42.0	42.0	20.130175	81.604873	7.628473	213.841241	20
...	...	...	...	...	...	...	...	...
2195	107	34.0	32.0	26.774637	66.413269	6.780064	177.774507	5
2196	99	15.0	27.0	27.417112	56.636362	6.086922	127.924610	5
2197	118	33.0	30.0	24.131797	67.225123	6.362608	173.322839	5
2198	117	32.0	34.0	26.272418	52.127394	6.758793	127.175293	5
2199	104	18.0	30.0	23.603016	60.396475	6.779833	140.937041	5

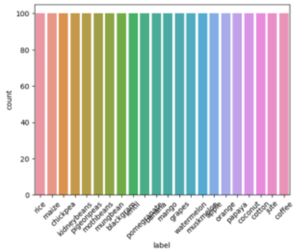
6. ***Visualizations*:** Plotted various visualization techniques such as histograms, scatter plots, box plots, relplot, probplot, displot, pie charts to visualize the distribution and relationships between different variables.

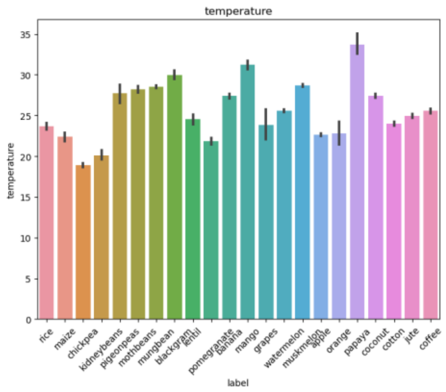
VISUALIZATIONS

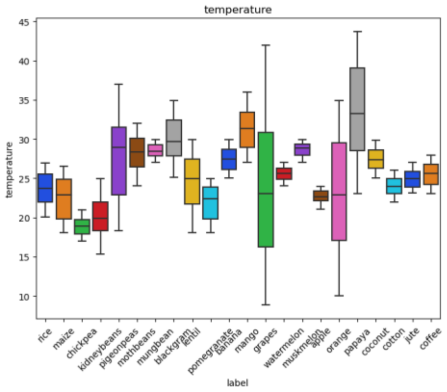


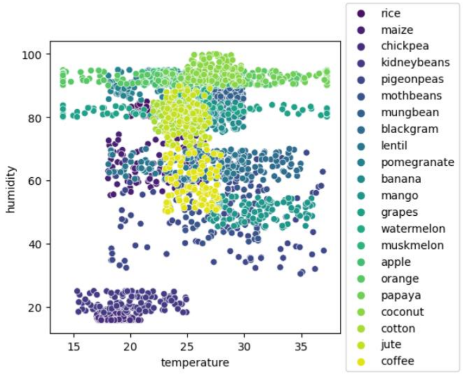


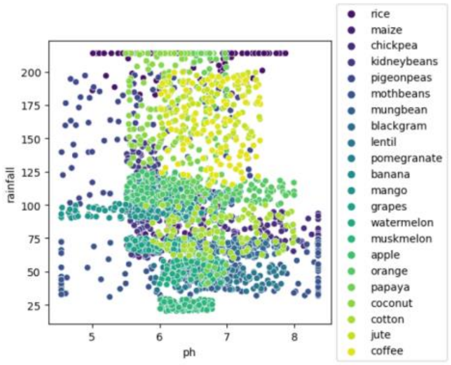


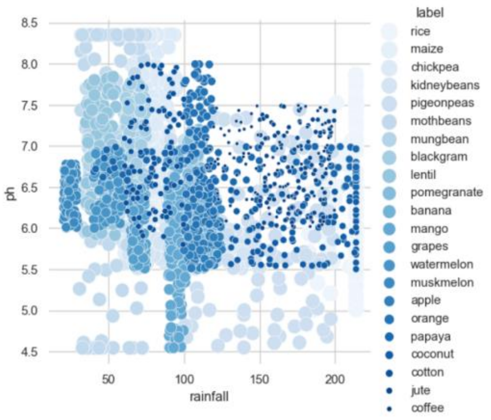


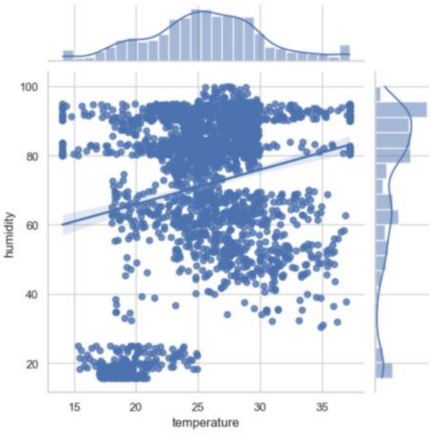

temperature
temperature
0
5
10
15
20
25
30
35
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee
label


temperature
temperature
10
15
20
25
30
35
40
45
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee
label


humidity
100
80
60
40
20
15
20
25
30
35
temperature
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee


rainfall
200
175
150
125
100
75
50
25
5
6
7
8
ph
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee


ph
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
50
100
150
200
rainfall
label
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee


humidity
100
80
60
40
20
15
20
25
30
35
temperature

Modelling:

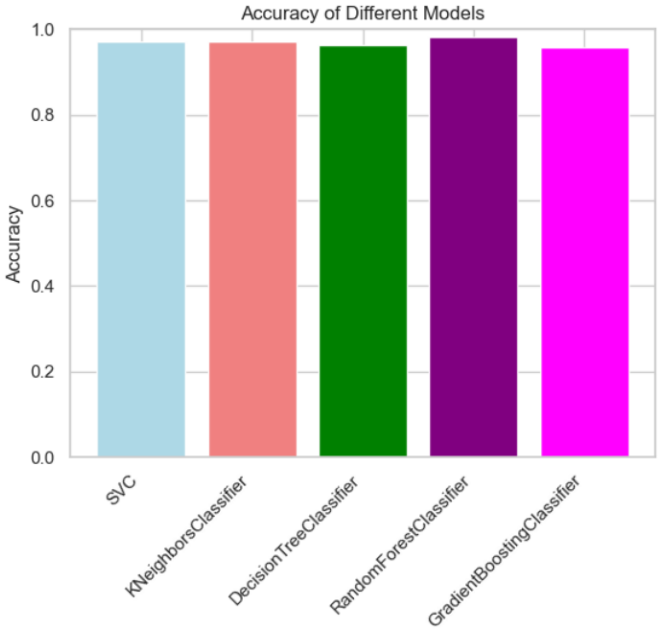
1. ***Random Forest Classifier***: This ensemble learning method is employed to classify crops based on various environmental and soil parameters. It leverages multiple decision trees to make predictions and provides robust recommendations for crop selection.

2. ***Support Vector Classifier (SVC):*** SVC is utilized to classify crops by creating hyperplanes in multidimensional space, effectively separating different crop categories. It's particularly effective in handling complex datasets and finding optimal decision boundaries.

3. ***K-Nearest Neighbors (KNN)***: KNN algorithm predicts the class of a sample by identifying the majority class among its nearest neighbors. It's employed to classify crops based on similarity to neighboring data points in the feature space.

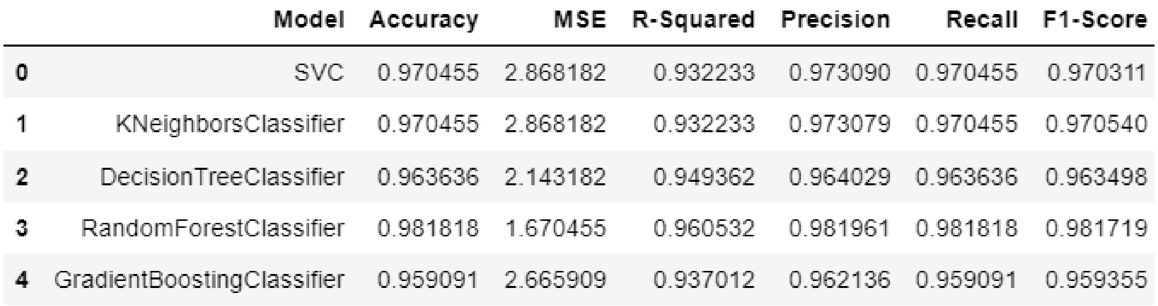
4. ***Decision Tree Classifier***: Decision trees are used to model the decision-making process based on input features. They partition the feature space into distinct regions, making them interpretable and effective for crop classification tasks.

5. ***Logistic Regression:*** Logistic regression is employed for binary classification tasks, predicting the probability that a crop belongs to a particular category. It's used to model the relationship between input features and crop categories.



Model Evaluation:

Various metrics such as accuracy, precision, recall, and F1-score are used to evaluate the performance of machine learning models. confusion matrices are employed to assess model generalization and identify potential areas for improvement.



	Model	Accuracy	MSE	R-Squared	Precision	Recall	F1-Score
0	SVC	0.970455	2.868182	0.932233	0.973090	0.970455	0.970311
1	KNeighborsClassifier	0.970455	2.868182	0.932233	0.973079	0.970455	0.970540
2	DecisionTreeClassifier	0.963636	2.143182	0.949362	0.964029	0.963636	0.963498
3	RandomForestClassifier	0.981818	1.670455	0.960532	0.981961	0.981818	0.981719
4	GradientBoostingClassifier	0.959091	2.665909	0.937012	0.962136	0.959091	0.959355

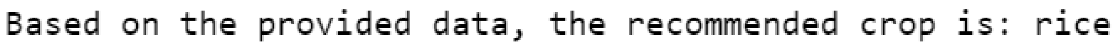
Hyper Parameter Tuning:

Hyperparameter tuning optimizes machine learning models by systematically searching for the best hyperparameter values. In the provided code, a RandomForestClassifier is trained using GridSearchCV, which explores various combinations of hyperparameters like n_estimators, max_depth, min_samples_split, and min_samples_leaf. The best combination of hyperparameters is selected based on cross-validated performance, enhancing model accuracy and generalization.

Final output:

For the input -

90 42 43 20.879744 82.002744 6.502985 202.935536



```
Based on the provided data, the recommended crop is: rice
```

CONCLUSION:

The Crop Recommendation System developed in this project harnesses the power of machine learning to provide personalized crop recommendations based on environmental and soil conditions. Through extensive exploratory data analysis, various machine learning models, and hyperparameter tuning, the system demonstrates robustness and accuracy in predicting suitable crops. By empowering farmers with data-driven insights, the system contributes to optimizing agricultural practices, maximizing yields, and ultimately fostering sustainable farming practices.