Crop recommendation system

- Abstract:

In the realm of modern agriculture, data-driven approaches are rapidly transforming traditional farming practices. This project centers around the creation and deployment of a Crop Recommendation System, leveraging the rich dataset provided by Atharva Ingle on Kaggle. The system harnesses the potential of machine learning and data analysis to empower farmers with informed decisions for optimal crop selection, tailored to their specific geographical and environmental conditions. The cornerstone of this project is the Atharva Ingle Dataset, which encompasses an array of influential factors including soil properties, climatic variables, and historical crop yields. By ingesting this dataset, the Crop Recommendation System gains the ability to discern intricate patterns and relationships that govern successful crop growth. Machine learning algorithms are employed to create predictive models that anticipate crop performance based on the amalgamation of input factors. Key features of the Crop Recommendation System include real-time integration of climatic and soil data, intuitive interfaces enabling farmers to input their field characteristics, and a robust recommendation engine proficient in generating crop suggestions. These recommendations are grounded in the analysis of historical data and the correlation between different crops and environmental conditions. Moreover, the system factors in economic viability, thereby providing a comprehensive framework for decision-making.

- Introduction:

Agriculture, a cornerstone of human civilization, faces a pivotal juncture in its evolution. The challenges of a burgeoning global population, shifting climatic patterns, and the imperative for sustainable practices have converged to necessitate innovative solutions. The intersection of data science and agriculture presents a transformative opportunity, enabling the development of systems that optimize crop selection, resource allocation, and yield. This project delves into the creation and deployment of a Crop Recommendation System, a technological marvel designed to aid farmers in making informed decisions about the crops they cultivate. This recommendation system harnesses the potential of advanced machine learning techniques and the Atharva Ingle Dataset, an expansive repository of agricultural variables, to provide tailored crop suggestions based on prevailing environmental conditions. In the traditional agrarian landscape, farmers often rely on generational knowledge and experience to select crops. However, the intricate interplay between soil characteristics, climatic parameters, and crop requirements demands a more comprehensive and data-driven approach. This project recognizes the inherent power of data analytics to decode these intricate relationships and provide accurate insights, thereby enhancing agricultural productivity and sustainability. The Atharva Ingle Dataset, sourced from Kaggle, offers a comprehensive collection of data points encompassing soil attributes, temperature, humidity, rainfall, and crop yields. By feeding this dataset into the Crop Recommendation System, the project aims to unlock hidden patterns and correlations that are beyond the scope of conventional human analysis. Machine learning algorithms trained on this dataset have the capacity to unravel the complexities of successful crop cultivation in diverse conditions. At its core, this project strives to offer an intuitive and accessible tool for farmers. The Crop Recommendation System provides an interface for farmers to input their specific field characteristics and receive personalized recommendations. By considering historical data and employing predictive models, the system generates suggestions that take into account not only crop suitability but also economic viability.

- In the context of agriculture and crop cultivation, nitrogen, phosphorus, and potassium are three essential nutrients commonly referred to as NPK. These nutrients are crucial for the healthy growth, development, and overall productivity of plants. In your crop recommendation project, the use of nitrogen, phosphorus, and potassium likely plays a pivotal role in determining optimal crop choices. Here's an explanation of their significance:

Nitrogen (N):
Nitrogen is a fundamental nutrient that plants need for various processes, including photosynthesis, protein synthesis, and overall growth. It is a primary component of chlorophyll, the pigment responsible for capturing light energy during photosynthesis. Adequate nitrogen levels promote vigorous foliage growth, leading to healthier plants with vibrant green leaves. Nitrogen deficiency can result in stunted growth, pale leaves, and decreased crop yield. Excessive nitrogen application, however, can lead to imbalanced growth and environmental issues, such as water pollution.

Phosphorus (P):
Phosphorus is essential for energy transfer within plants, aiding in processes like root development, flower formation, and seed production. It is a critical component of ATP (adenosine triphosphate), which serves as the primary energy currency in cells. Phosphorus deficiency can lead to poor root growth, delayed flowering, and reduced fruiting. Incorporating sufficient phosphorus in soil supports healthy plant establishment and enhances overall crop quality. Balancing phosphorus application is important, as excess phosphorus can negatively impact water bodies and aquatic ecosystems.

Potassium (K):
Potassium is vital for various physiological functions in plants, including water uptake, enzyme activation, and overall stress resistance. It helps regulate water balance within cells, enhances disease resistance, and improves overall plant resilience to environmental stressors. Adequate potassium levels contribute to stronger stems, improved root growth, and increased crop yield. A lack of potassium can lead to weakened plant structure, reduced disease resistance, and decreased tolerance to extreme conditions.

- Visualizations:

- Boxplots for Outlier Detection and Training Data Analysis

In the pursuit of understanding and enhancing the performance of machine learning models, data visualization plays a crucial role. Boxplots, also known as box-and-whisker plots, stand as powerful tools for analyzing data distribution, identifying outliers, and gaining insights into the central tendencies of the dataset. In this project, boxplots were employed to assess and address outliers within the dataset, particularly during the phases of checking and training.

- Boxplot Components:

A boxplot consists of several components that collectively provide a comprehensive view of the data distribution:

Box: The box in the center of the plot represents the interquartile range (IQR), which encompasses the middle 50% of the data. The box is divided into the lower quartile (Q1) and the upper quartile (Q3), with the median marked by a line inside the box.

Whiskers: The whiskers extend from the box to the minimum and maximum values within a specified range, often 1.5 times the IQR. Data points beyond the whiskers are considered potential outliers.

Outliers: Outliers are data points that fall significantly outside the whiskers' range. They are represented as individual points on the plot and could be indications of anomalies or errors in the data.

- Outlier Detection:

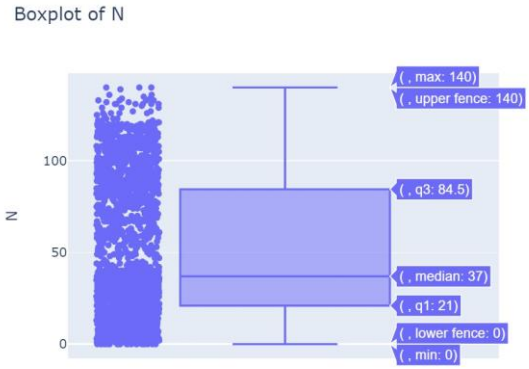
In the context of this project, boxplots were employed as a means of outlier detection. Outliers can significantly impact the performance of machine learning algorithms, leading to skewed models and erroneous predictions. By visualizing the distribution of data through boxplots, it becomes easier to identify these anomalies and assess their impact on the training process.

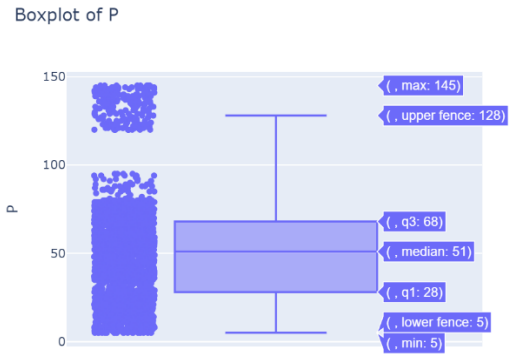
- Training Data Analysis:

Before training machine learning models, it is essential to gain a holistic understanding of the dataset's characteristics. Boxplots aid in this analysis by revealing the distribution of features and their potential variations. By visualizing the spread of data within each feature through boxplots, one can determine whether the dataset exhibits skewness, the presence of outliers, and the overall distribution pattern.

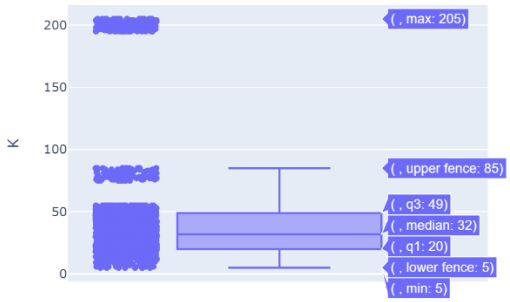
- Interpreting Results:

During the checking and training phases, the boxplots help in making informed decisions. If outliers are identified, it prompts a closer inspection to determine whether they are genuine data points or errors that need to be addressed. Outliers might necessitate data cleaning, transformation, or more advanced handling techniques. Additionally, analyzing boxplots for different features provides insights into the range of values and the relative magnitude of variations, which can guide feature engineering and model selection.

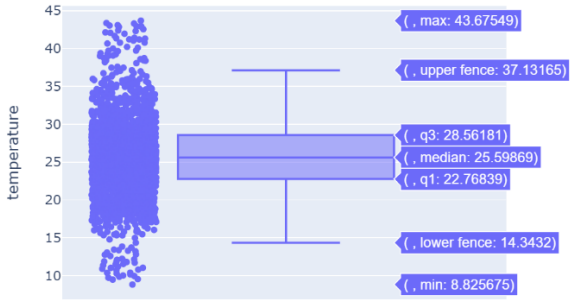




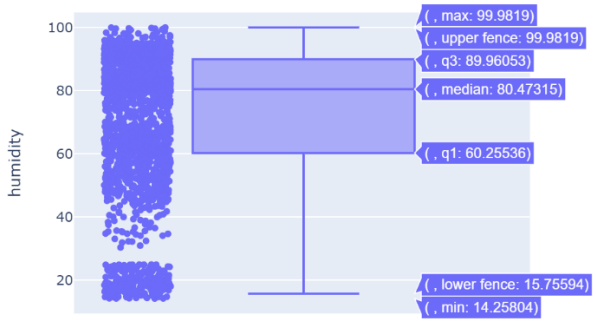
Boxplot of K



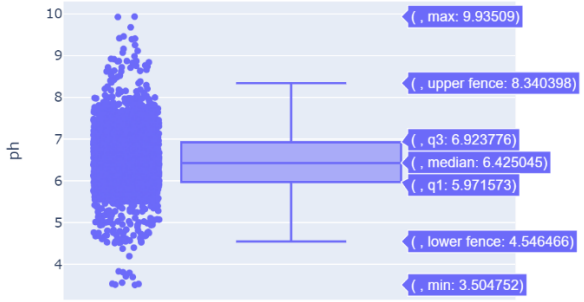
Boxplot of temperature



Boxplot of humidity



Boxplot of ph



Outliners is used because of the following reasons:

1. To detect the frauds in the dataset.
2. To detect the outliers in the dataset.
3. To detect the extreme values in the dataset.
    - Outliners means the values which are far away from the mean of the dataset.
    - Outliners can be detected by using the boxplot.
    - Outliners can be removed by using the IQR method.
    - Outliners can be removed by using the z-score method.
    - Outliners can be removed by using the standard deviation method.
    - Outliners can be removed by using the scatter plot method.
    - Outliners can be removed by using the histogram method.
    - Outliners can be removed by using the percentile method.
    - Outliners can be removed by using the quantile method.
    - Outliners can be removed by using the log method.
    - Outliners can be removed by using the square root method.

IQR means Inter Quartile Range.
IQR = Q3 - Q1
Q1 = 25%
Q2 = 50%
Q3 = 75%
Q4 = 100%

Z-Score means the standard score.
Z-Score = (x - mean)/standard deviation
Z-Score = (x - mean)/sigma

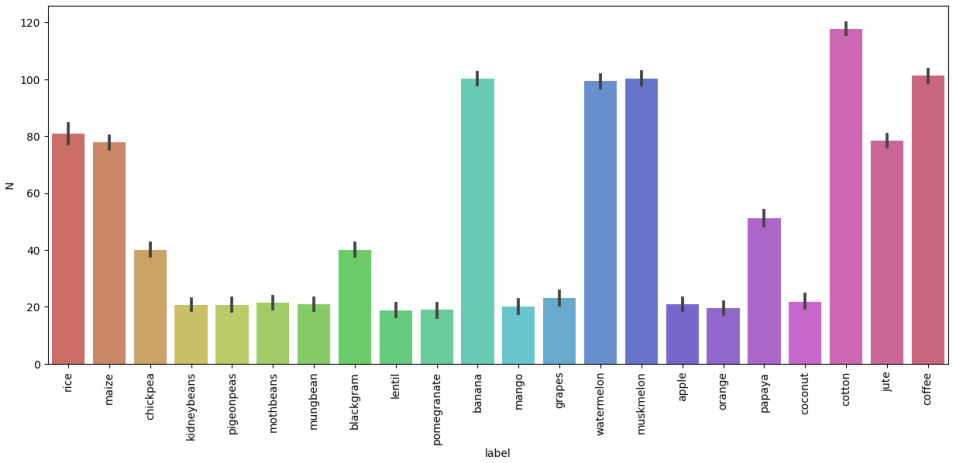
Standard Deviation means the square root of the variance.
Standard Deviation = sqrt(variance)
Standard Deviation = sqrt(sigma)

Scatter Plot means the plot between the two variables.
Scatter Plot means the plot between the two columns.

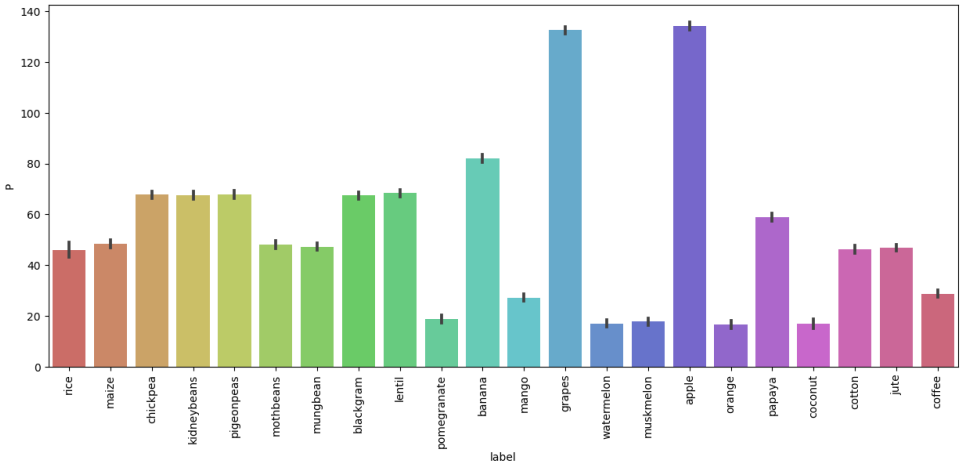
Histogram means the plot between the frequency and the variable.
Histogram means the plot between the frequency and the column.

Percentile means the percentage of the values.
Percentile means the percentage of the columns.

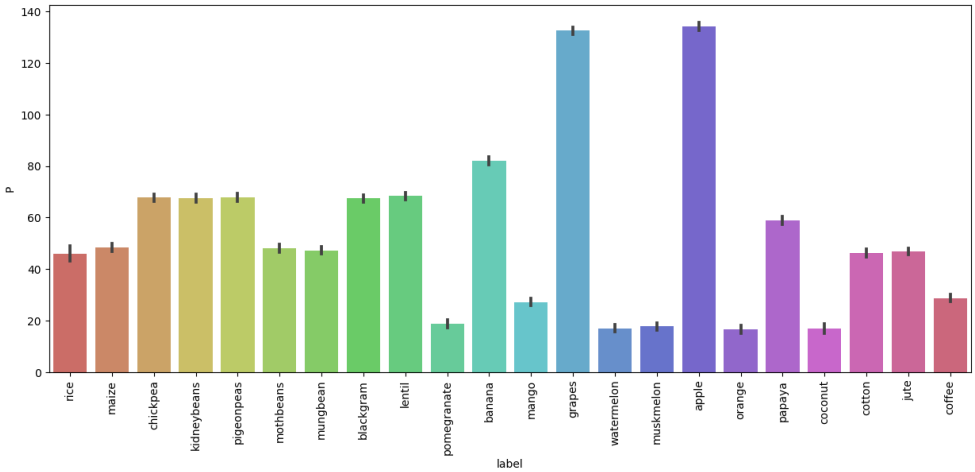
- This bar plot shows that the nitrogen content is highest in cotton and lowest in lentil.



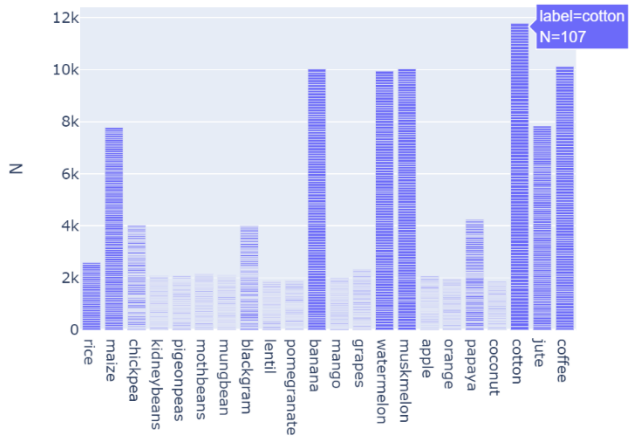
- This bar plot shows that the phosphorous content is highest in apple and lowest in watermelon.



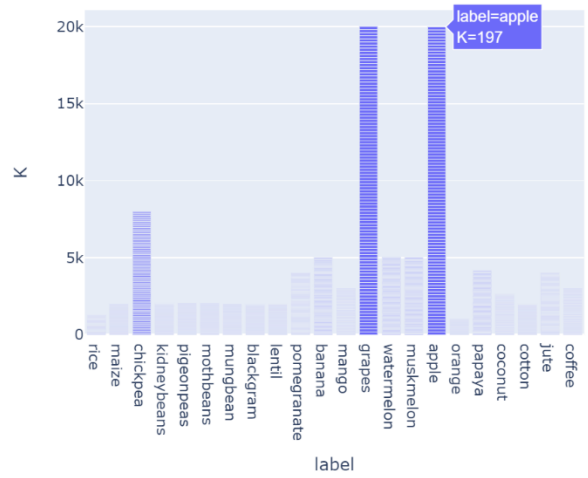
- This bar plot shows that the potassium content is highest in grapes and lowest in orange.



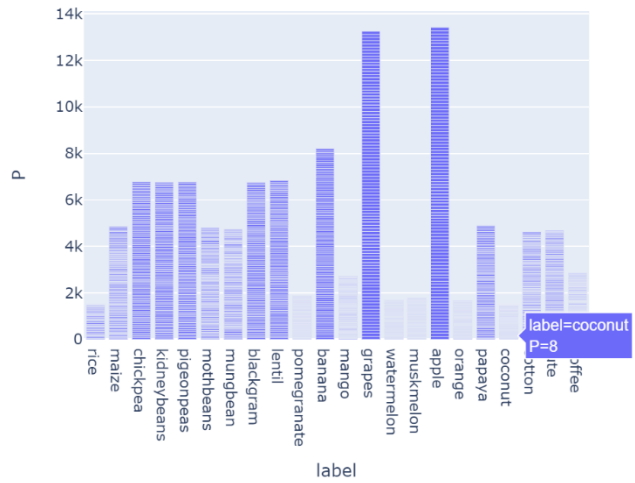
- This shows that the nitrogen content is highest in cotton and lowest in coconut.



- This shows that the crop which requires more nitrogen also requires more potassium and potassium content is highest in apple and grapes and lowest in orange.



- This shows that the crops which require more nitrogen also require more phosphorous and potassium and phosphorus is highest inapple and pomegranate, rice and coconut.



- Correlation:

Correlation refers to the statistical measure of the strength and direction of a relationship between two or more variables. In the context of your crop recommendation project, correlation analysis involves examining how different factors or variables within the dataset relate to each other. Specifically, you might be interested in understanding how variables like soil properties, weather conditions, and nutrient levels correlate with each other and potentially impact crop growth and yield.

Key aspects of correlation analysis include:

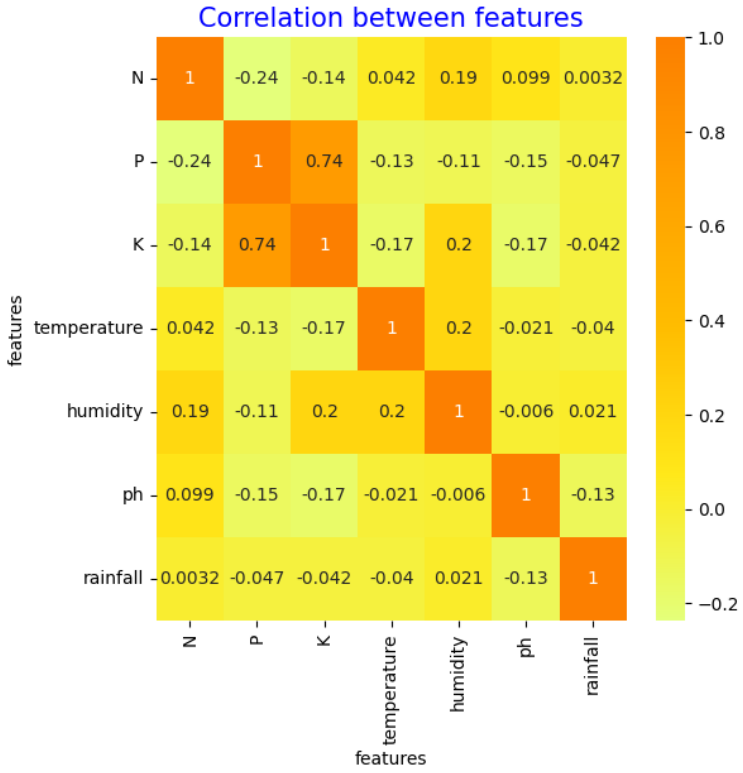
Strength of Correlation: The strength of correlation indicates how closely the variables are related. Correlation values range from -1 to 1:

A correlation coefficient of 1 indicates a perfect positive correlation, where as one variable increases, the other variable also increases proportionally.

A correlation coefficient of -1 indicates a perfect negative correlation, where as one variable increases, the other variable decreases proportionally.
A correlation coefficient close to 0 indicates a weak or no linear correlation.
Direction of Correlation: The sign of the correlation coefficient (+ or -) indicates the direction of the relationship:

Positive correlation means that as one variable increases, the other variable tends to increase as well.
Negative correlation means that as one variable increases, the other variable tends to decrease.
Correlation Coefficient Calculation: Commonly used correlation coefficients include the Pearson correlation coefficient (for linear relationships), the Spearman rank correlation coefficient (for non-linear relationships), and the Kendall Tau rank correlation coefficient. These coefficients provide numerical values that quantify the strength and direction of the relationship between variables.
Correlation Matrix: A correlation matrix is a table that displays the correlation coefficients between multiple variables. It provides an overview of how each variable is correlated with others, helping identify patterns and relationships.
In this project, you might use correlation analysis to:

- Understand how different soil properties (e.g., pH, nutrient levels) correlate with each other.
- Determine if specific weather conditions (e.g., temperature, rainfall) are correlated with certain crop yields.
- Identify potential multicollinearity (high correlation) between features, which can impact the effectiveness of predictive models.
- By analyzing correlations, you can gain insights into which variables might be important for predicting crop outcomes and make informed decisions about feature selection, model training, and understanding the factors influencing crop performance.



- Models used in project:

In the pursuit of enhancing agricultural practices, machine learning algorithms have emerged as powerful tools capable of unravelling complex relationships within agricultural data. This project explores a diverse array of machine learning techniques, including Logistic Regression, Random Forest, Decision Tree, Hybrid Models, and Support Vector Machine (SVM), to create a comprehensive Crop Recommendation System. These algorithms collectively aim to decipher intricate patterns within the Atharva Ingle Dataset, leading to informed and optimal crop recommendations for farmers.

Logistic Regression:
Logistic Regression, though seemingly simple, is a fundamental algorithm for classification tasks. It models the probability of a binary outcome and is particularly suited for scenarios where the relationship between input variables and the outcome is linear. In the context of crop recommendation, Logistic Regression can play a role in delineating binary decisions, such as whether a specific crop is suitable or not based on prevailing conditions.

Random Forest:
Random Forest stands as a robust ensemble algorithm capable of handling complex relationships and avoiding overfitting. Comprising multiple decision trees, each trained on different subsets of the data, Random Forest aggregates their predictions to arrive at a more accurate and stable output. In our Crop Recommendation System, Random Forest can capture nonlinear relationships between diverse factors like soil type, weather conditions, and historical data, ultimately leading to precise crop suggestions.

Decision Tree:
Decision Trees offer a transparent way of mapping decisions based on input features. They recursively partition the data into subsets by selecting the best features at each node. Decision Trees are advantageous for their interpretability and can serve as an effective baseline model in our project. By employing Decision Trees, we aim to visually depict the thought process of crop selection while considering various factors.

Hybrid Models (Logistic Regression and Decision Tree):
Hybrid Models combine the strengths of different algorithms to create synergistic solutions. In this project, the combination of Logistic Regression and Decision Tree Classifier aims to harness the linear and nonlinear relationships simultaneously. This can provide a more nuanced approach to crop recommendation by considering both the simplicity of Logistic Regression and the complexity of Decision Trees.

Support Vector Machine (SVM):
Support Vector Machines are known for their prowess in classifying data by finding optimal hyperplanes that best separate different classes. In the context of crop recommendation, SVMs can excel in handling datasets with intricate decision boundaries. By mapping input variables to higher-dimensional spaces, SVMs can effectively capture subtleties in the data, leading to accurate crop predictions.

In summary, the utilization of these machine learning algorithms brings a multidimensional perspective to the realm of crop recommendation. By collectively leveraging the strengths of Logistic Regression, Random Forest, Decision Tree, Hybrid Models, and Support Vector Machine, our Crop Recommendation System aspires to provide farmers with comprehensive, accurate, and actionable insights for optimizing crop selection and resource allocation. This holistic approach underscores the potential of data science in revolutionizing modern agriculture.

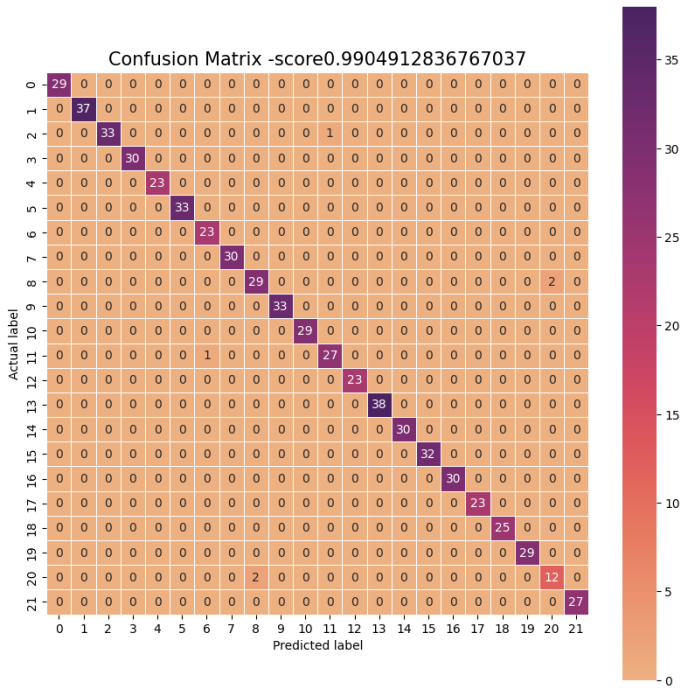
- Results:

1. DecisionTreeClassifier:

```
Accuracy: decision tree model accuracy score: 0.9905
```

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	37
blackgram	1.00	0.97	0.99	34
chickpea	1.00	1.00	1.00	30
coconut	1.00	1.00	1.00	23
coffee	1.00	1.00	1.00	33
cotton	0.96	1.00	0.98	23
grapes	1.00	1.00	1.00	30
jute	0.94	0.94	0.94	31
kidneybeans	1.00	1.00	1.00	33
lentil	1.00	1.00	1.00	29
maize	0.96	0.96	0.96	28
mango	1.00	1.00	1.00	23
mothbeans	1.00	1.00	1.00	38
mungbean	1.00	1.00	1.00	30
muskmelon	1.00	1.00	1.00	32
orange	1.00	1.00	1.00	30
papaya	1.00	1.00	1.00	23
pigeonpeas	1.00	1.00	1.00	25
pomegranate	1.00	1.00	1.00	29
rice	0.86	0.86	0.86	14
watermelon	1.00	1.00	1.00	27
accuracy			0.99	631
macro avg	0.99	0.99	0.99	631
weighted avg	0.99	0.99	0.99	631

Confusion Matrix:

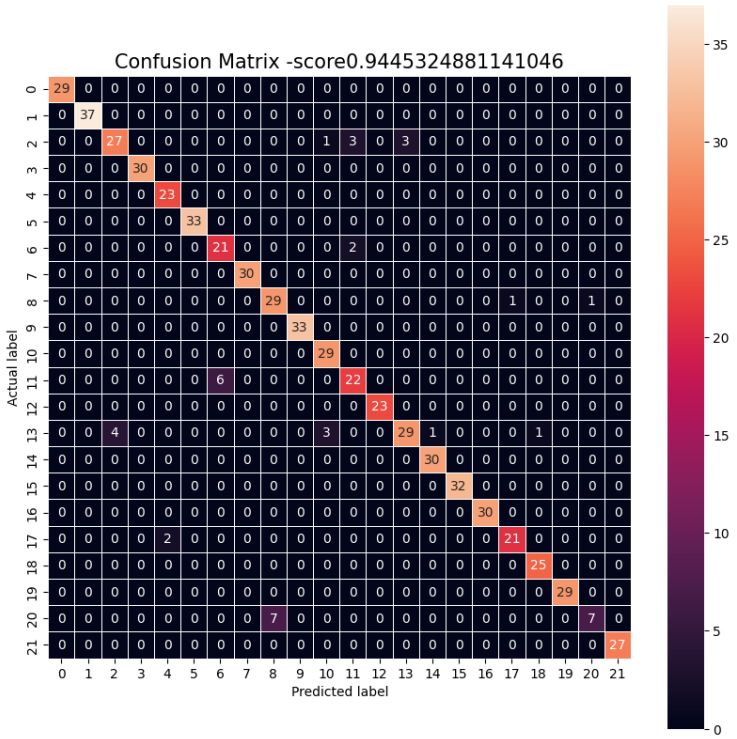


2. Logistic Regression:

Accuracy: Logistic Regression Model accuracy score: 0.9445

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	37
blackgram	0.87	0.79	0.83	34
chickpea	1.00	1.00	1.00	30
coconut	0.92	1.00	0.96	23
coffee	1.00	1.00	1.00	33
cotton	0.78	0.91	0.84	23
grapes	1.00	1.00	1.00	30
jute	0.81	0.94	0.87	31
kidneybeans	1.00	1.00	1.00	33
lentil	0.88	1.00	0.94	29
maize	0.81	0.79	0.80	28
mango	1.00	1.00	1.00	23
mothbeans	0.91	0.76	0.83	38
mungbean	0.97	1.00	0.98	30
muskmelon	1.00	1.00	1.00	32
orange	1.00	1.00	1.00	30
papaya	0.95	0.91	0.93	23
pigeonpeas	0.96	1.00	0.98	25
pomegranate	1.00	1.00	1.00	29
rice	0.88	0.50	0.64	14
watermelon	1.00	1.00	1.00	27
accuracy			0.94	631
macro avg	0.94	0.94	0.94	631
weighted avg	0.95	0.94	0.94	631

Confusion Matrix:

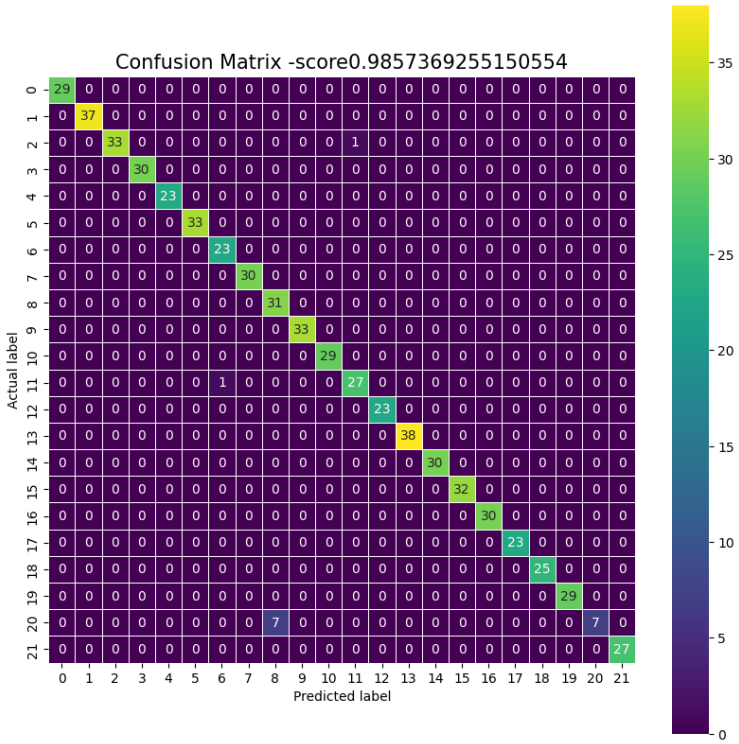


3. Random Forest Classifier:

```
Random Forest Model accuracy score: 0.9857
```

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	37
blackgram	1.00	0.97	0.99	34
chickpea	1.00	1.00	1.00	30
coconut	1.00	1.00	1.00	23
coffee	1.00	1.00	1.00	33
cotton	0.96	1.00	0.98	23
grapes	1.00	1.00	1.00	30
jute	0.82	1.00	0.90	31
kidneybeans	1.00	1.00	1.00	33
lentil	1.00	1.00	1.00	29
maize	0.96	0.96	0.96	28
mango	1.00	1.00	1.00	23
mothbeans	1.00	1.00	1.00	38
mungbean	1.00	1.00	1.00	30
muskmelon	1.00	1.00	1.00	32
orange	1.00	1.00	1.00	30
papaya	1.00	1.00	1.00	23
pigeonpeas	1.00	1.00	1.00	25
pomegranate	1.00	1.00	1.00	29
rice	1.00	0.50	0.67	14
watermelon	1.00	1.00	1.00	27
accuracy			0.99	631
macro avg	0.99	0.97	0.98	631
weighted avg	0.99	0.99	0.98	631

Confusion Matrix:

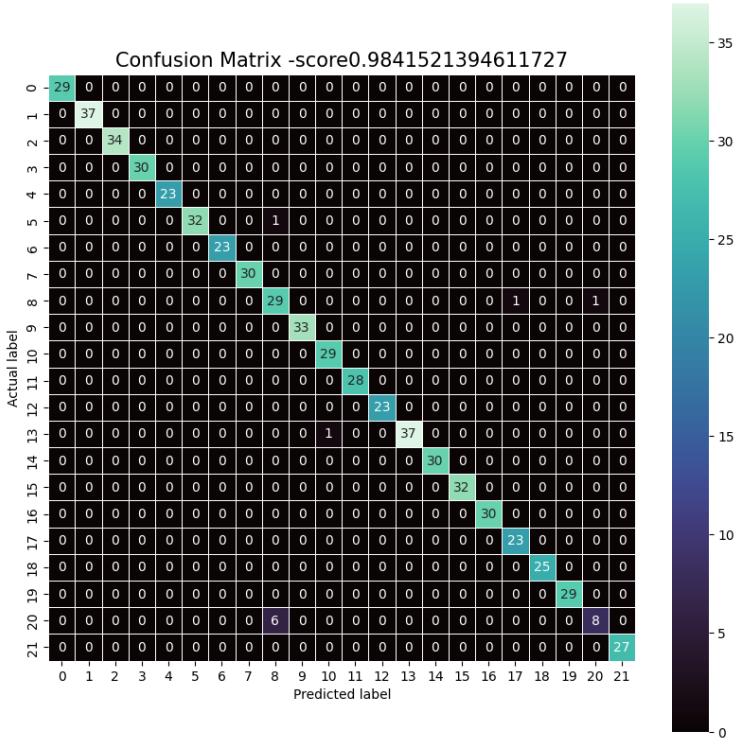


4. Server Vector Machine (SVM):

SVM Model accuracy score: 0.9842

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	37
blackgram	1.00	1.00	1.00	34
chickpea	1.00	1.00	1.00	30
coconut	1.00	1.00	1.00	23
coffee	1.00	0.97	0.98	33
cotton	1.00	1.00	1.00	23
grapes	1.00	1.00	1.00	30
jute	0.81	0.94	0.87	31
kidneybeans	1.00	1.00	1.00	33
lentil	0.97	1.00	0.98	29
maize	1.00	1.00	1.00	28
mango	1.00	1.00	1.00	23
mothbeans	1.00	0.97	0.99	38
mungbean	1.00	1.00	1.00	30
muskmelon	1.00	1.00	1.00	32
orange	1.00	1.00	1.00	30
papaya	0.96	1.00	0.98	23
pigeonpeas	1.00	1.00	1.00	25
pomegranate	1.00	1.00	1.00	29
rice	0.89	0.57	0.70	14
watermelon	1.00	1.00	1.00	27
accuracy			0.98	631
macro avg	0.98	0.98	0.98	631
weighted avg	0.98	0.98	0.98	631

Confusion Matrix:

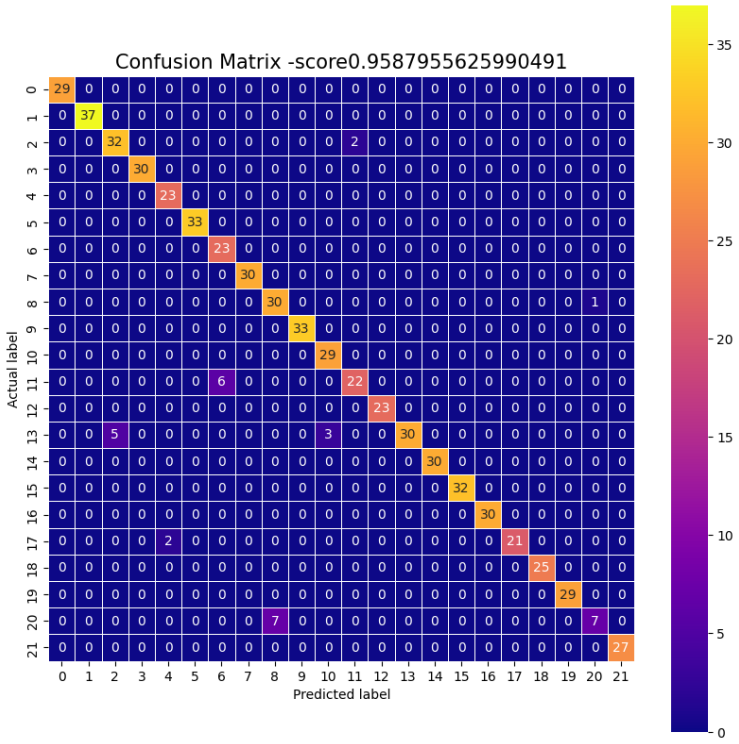


5. Hybrid Model (Logistic Regression + Decision Tree Classifier):

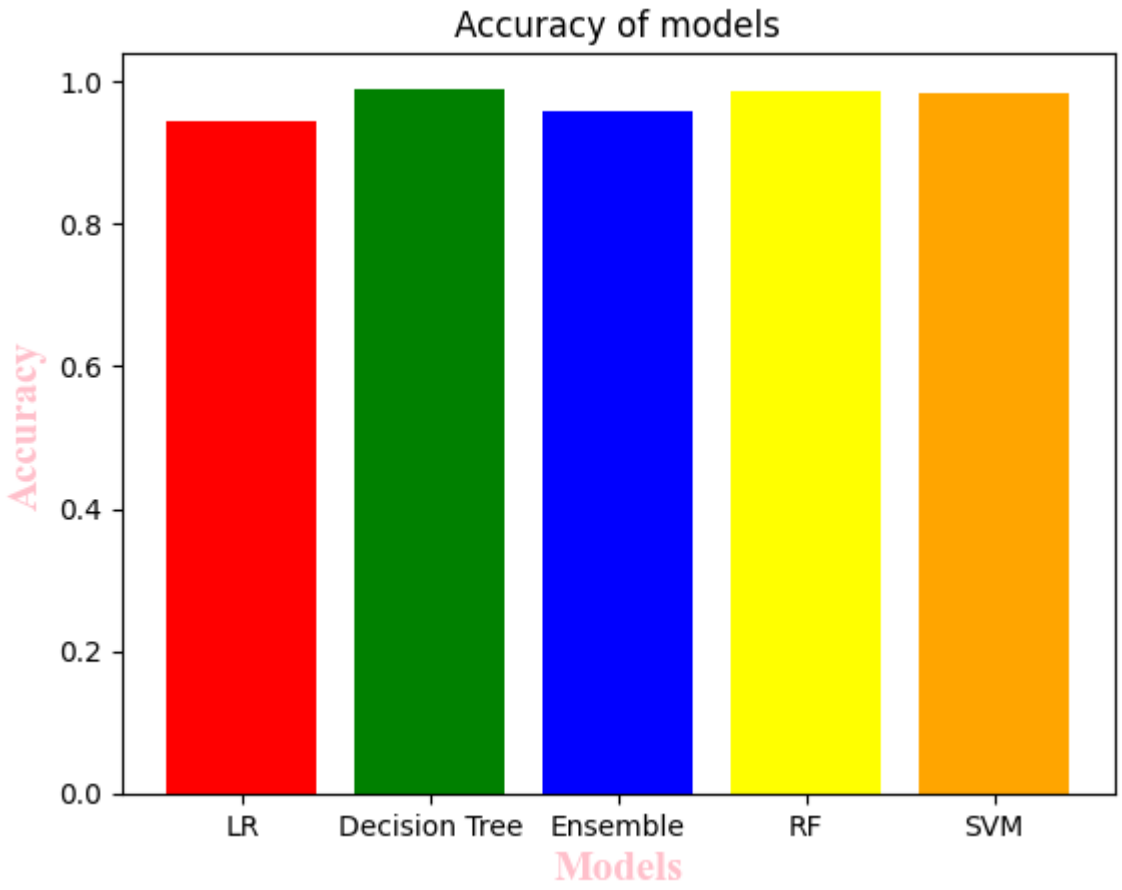
```
Accuracy score of ensemble model is: 0.9587955625990491
```

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	29
banana	1.00	1.00	1.00	37
blackgram	0.86	0.94	0.90	34
chickpea	1.00	1.00	1.00	30
coconut	0.92	1.00	0.96	23
coffee	1.00	1.00	1.00	33
cotton	0.79	1.00	0.88	23
grapes	1.00	1.00	1.00	30
jute	0.81	0.97	0.88	31
kidneybeans	1.00	1.00	1.00	33
lentil	0.91	1.00	0.95	29
maize	0.92	0.79	0.85	28
mango	1.00	1.00	1.00	23
mothbeans	1.00	0.79	0.88	38
mungbean	1.00	1.00	1.00	30
muskmelon	1.00	1.00	1.00	32
orange	1.00	1.00	1.00	30
papaya	1.00	0.91	0.95	23
pigeonpeas	1.00	1.00	1.00	25
pomegranate	1.00	1.00	1.00	29
rice	0.88	0.50	0.64	14
watermelon	1.00	1.00	1.00	27
accuracy			0.96	631
macro avg	0.96	0.95	0.95	631
weighted avg	0.96	0.96	0.96	631

Confusion Matrix:



- Accuracy of all models:



Final Result:

	N	P	K	temperature	humidity	ph	rainfall
1203	36	125	196	37.465668	80.659687	6.155261	66.838723

```
1203    grapes
   Name: label, dtype: object
```