1. **Problem Statement**: Crop Recommendation System

2. Input Parameters:

N: The amount of nitrogen in the soil in kg/ha.

P: The amount of phosphorus in the soil in kg/ha.

K: The amount of potassium in the soil in kg/ha.

temperature: The temperature in Celsius (°C) at the time of crop cultivation.

humidity: The relative humidity in percentage (%) at the time of crop cultivation.

ph: The pH value of the soil.

rainfall: The amount of rainfall in mm during the crop cultivation period.

Output Parameters:

Label: The target variable which indicates the type of crop that is recommended based on the given environmental factors

3. DATASET INSIGHTS:

Dataset Used: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

The dataset includes various environmental factors that affect the growth of crops, such as soil nutrient levels, temperature, humidity, pH, and rainfall. The target variable is the type of crop that is recommended based on these environmental factors.

Here's some additional information about the crop recommendation system dataset:

- The dataset contains data for 22 different crops like rice, wheat, maize etc.
- There are a total of 2200 instances in the dataset, with 100 instances for each crop.
- The data is not normalized, meaning that the values for each feature are not on the same scale. This can cause issues when working with certain machine learning algorithms that require features to be on the same scale.
- The dataset may contain missing or invalid data, which may need to be addressed before using it for machine learning.
- The dataset may require further feature engineering, such as creating new features or combining existing features, to improve the performance of machine learning models.

	count	unique	top	freq
N	2200	137	22	44
P	2200	117	60	56
K	2200	73	17	90
temperature	2200.0	2200.0	8.825675	1.0
humidity	2200.0	2200.0	14.25804	1.0
ph	2200.0	2200.0	3.504752	1.0
rainfall	2200.0	2200.0	20.211267	1.0
label	2200	22	apple	100

4. PROPOSED SOLUTION

a. Data Preprocessing

i. Remove any missing or null attributes from the dataset
ii. Encode labels to digits

b. Possible Algorithms for Classification:

i. Logistic Regression
ii. Naïve Bayes
iii. Decision Trees
iv. Support Vector Machines (SVM)
v. Random Forest

c. Libraries Used

i. **Pandas**: A powerful data manipulation library in Python, widely used for data analysis and manipulation.

ii. **NumPy**: A fundamental package for scientific computing with Python, providing support for arrays, matrices, and mathematical functions.

iii. **seaborn**: A Python visualization library based on matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.

iv. **matplotlib.pyplot**: A collection of command-style functions that make matplotlib work like MATLAB for creating static, animated, and interactive visualizations in Python.

v. **MinMaxScaler**: A utility class in scikit-learn used for scaling features to a specified range (by default, between 0 and 1) using Min-Max scaling.

vi. **StandardScaler**: A utility class in scikit-learn used for standardizing features by removing the mean and scaling to unit variance.

vii. **Scikit-learn**: Simple and efficient tools for predictive data analysis

d. Implementation

https://colab.research.google.com/drive/13GxD3DlmdRddVQSfJ7V61gzr6QRRULSa?usp=sharing

e. Accuracy Measure

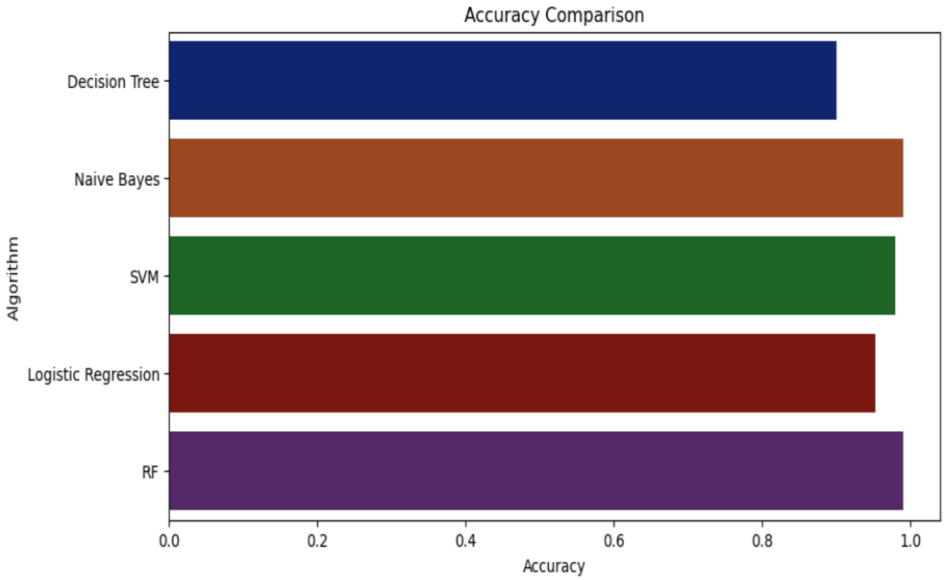
i. We track when the predicted class and actual class is the same and include that in our score, else the score does not increase.

Results:

Testing Accuracy:

```
Logistic Regression --> 0.9522727272727273

Decision Tree --> 0.9

RF --> 0.990909090909091

Naive Bayes --> 0.990909090909091

SVM --> 0.9795454545454545
```

Algorithm v/s Accuracy Score plot



➔ From the above accuracy scores, we can choose Random Forest algorithm.

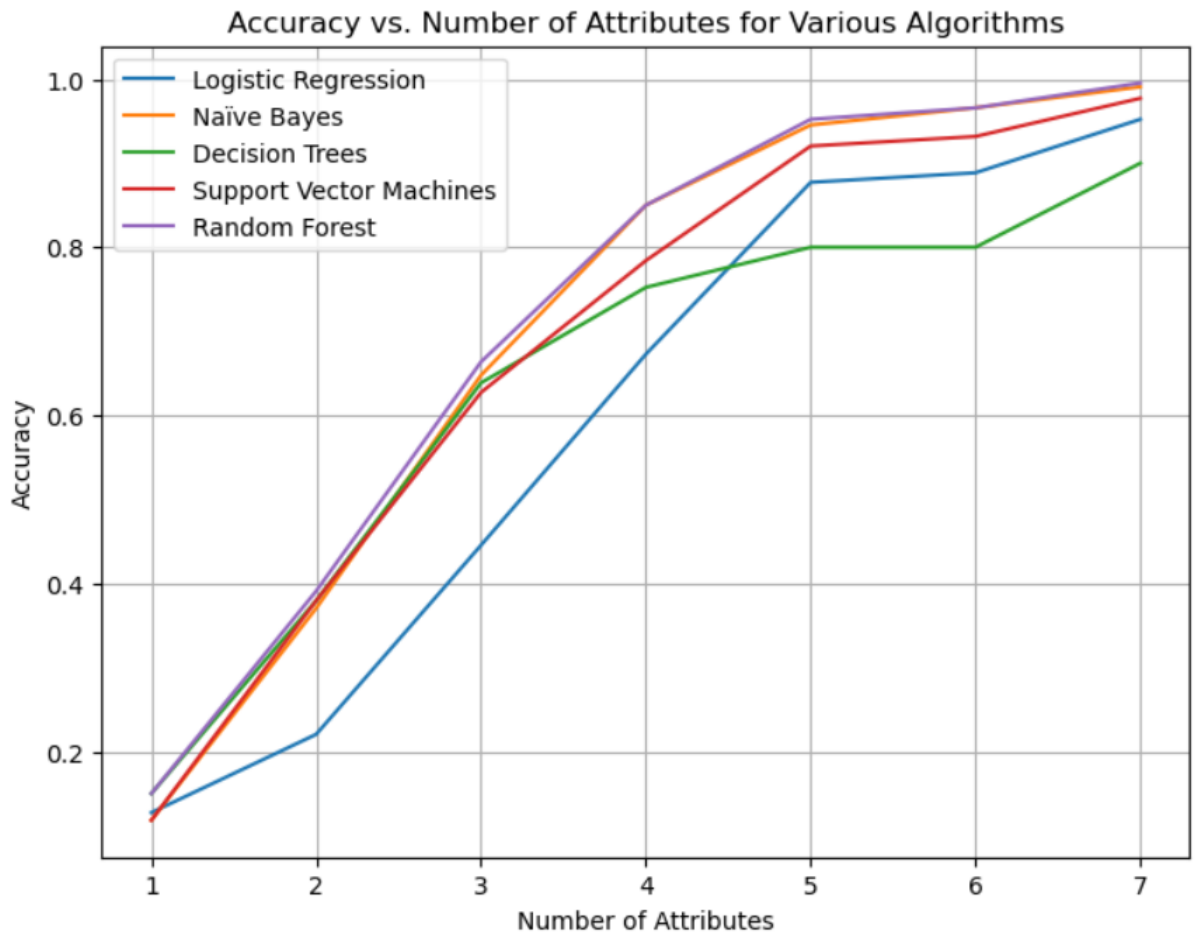
ii. **K-Fold Cross Validation**: The dataset is divided into k parts and then k-1 parts are used for training and the kth part is used for validation.

4-fold Cross validation:

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**
Logistic Regression	**0.95272727**	**0.94909091**	**0.96181818**	**0.94727273**
Decision Tree	**0.93272727**	**0.90909091**	**0.88363636**	**0.93272727**
Random Forest	**0.99454545**	**0.99090909**	**0.9963636**	**0.98181818**
Naïve Bayes	**0.99818182**	**0.99090909**	**0.99636364**	**0.99272727**
SVM	**0.98363636**	**0.98181818**	**0.98909091**	**0.98**

f. Feature Selection: Can we achieve similar accuracy with less features?

i. Chi-Square Analysis - Using chi-square analysis we can pick the top most independent attributes and only use them for classification and check the accuracy.



Here we can see that accuracy is increasing almost linearly with the number of features

g. Advantages and Disadvantages

Algorithm	Advantages	Disadvantages
Logistic Regression	Easy to interpret: You can understand how each feature affects the outcome	Not suitable for non-linear relationships: Assumes a linear relationship between features and target variable
Naive Bayes	Fast and efficient for large datasets: Makes calculations very quickly	Assumes independence of features: This assumption may not always hold true
Decision Trees	Easy to understand and visualize: You can see the decision-making process	Prone to overfitting: Can be too specific to the training data
Support Vector Machines (SVM)	Effective for high-dimensional data: Handles many features well	Can be computationally expensive for large datasets: Training can be slow
Random Forest	Handles high dimensionality well: Handles many features well	Can be a black box model: Difficult to interpret why it makes certain predictions