A Thesis/Project/Dissertation Report

on

Crop Prediction Using Machine Learning

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology

Under The Supervision of

Professor

Dr. Saurabh Raj Sangwaan

Submitted By

Mohd Farhan 21SCSE1011184

Mohit Kumar Srivastawa 21SCSE1010684

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING /
DEPARTMENT OF COMPUTERAPPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

OCT, 2024

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"Crop Prediction Using Machine Learning"** in partial fulfillment of the requirements for the award of the **Bachelor Of Technology** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of may, 2024 to aug, 2024 under the supervision of Dr. Saurabh Raj Sangwaan , Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Mohd Farhan 21SCSE1011184

Mohit Kumar Srivastawa 21SCSE1010684

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Saurabh Raj Sangwaan

Professor

<u>CERTIFICATE</u>

The MTE Thesis/Project/ Dissertation Viva-Voice examination of **Mohd Farhan (21SCSE1011184) and Mohit Kumar Srivastawa (21SCSE1010684)** has been held on 24 Oct and his/her work is recommended for the award of Bachelor Of Technology.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Program Chair **Signature of Dean**

Date:

Place: Greater Noida

Abstract

Paragraph-1 Existing Problem- 5-8 Lines
Paragraph-2 Proposed Solution- 6-8 Lines
Paragraph-3 Tools and Technology Used- 3-5 lines
Paragraph -4 Results and output- 3-5 lines
Paragraph-5 Conclusion and Future Scope – 2-4 Lines

Table of Contents

List of Table

List of Figures

Abstract

The Final Thesis/Project/ Dissertation Viva-Voce examination of Agriculture holds immense importance in India as a widely practiced occupation and plays a crucial role in the country's development. With approximately 60 percent of the total land dedicated to agriculture, India meets the needs of its 1.2 billion population through this sector. Enhancing crop production is considered a significant aspect of agriculture. To effectively utilize a piece of land, it is essential to determine the suitable crops based on the specific area and its corresponding soil properties. Various factors, such as soil type, temperature, and humidity, influence crop cultivation. However, the agricultural sector faces challenges due to unnatural climate changes, leading to decreased food production and making it difficult for farmers to forecast future yields. Providing farmers with advance knowledge of suitable crops before sowing could greatly benefit them and enable informed decisions regarding storage and business operations.

To address these issues, our proposed project focuses on monitoring agricultural areas based on soil properties and recommending the most appropriate crops to farmers. This recommendation system utilizes machine learning techniques such as unsupervised learning, predictive modeling, K-Means clustering, and Logistic Regression. By considering input soil parameters, including temperature, humidity, and moisture content, the system suggests suitable crops. We collect relevant data on crop seeds, ensuring optimal conditions for successful growth. The implementation of this system aims to minimize financial losses incurred by farmers from planting unsuitable crops and enables them to explore new types of crops suitable for their specific regions.

Keywords- Machine Learning, Unsupervised learning, Predictive Modelling, K Means Clustering, Logistics Regression.

CHAPTER-1

Introduction

Agriculture plays a vital role in society, having been practiced in India since ancient times. In the past, people cultivated crops on their own lands to meet their needs. Over time, advancements in technology have supported the agricultural sector, which is often referred to as India's backbone. As the population continues to grow rapidly, innovative approaches are necessary to fulfill the increasing demands of the people.

In recent years, India has experienced significant fluctuations in onion prices. Observing the substantial increase in onion prices, many farmers opted to cultivate onions in their fields with the hope of making substantial profits. However, this resulted in an excess supply of onions in some regions, while others suffered from failed crop production, causing significant financial losses for the farmers. Unfavorable conditions hindered onion growth, leading to a persistent shortage and adversely affecting the lives of the common people. The middle-class population, in particular, faced challenges due to the soaring onion prices, as onions are a frequently used commodity in their households.

The aforementioned example serves as the basis for identifying our problem statement.

Problem Statement- The farmer's decision on which crop to cultivate on their land is often influenced by intuition and various factors such as the desire to generate significant profits quickly, limited awareness of market demand, and overestimating a soil's suitability for a specific crop. However, making a wrong decision can impose significant financial strain on the farmer and their family, leading to substantial losses. Therefore, choosing the right crop to grow on the land can be a stressful and critical decision for farmers.

Solution- Now, the key objective is to develop a recommendation system that assists farmers in determining the suitable crop for a specific plot of land. To achieve this goal, we propose the creation of a system that takes into account soil parameters such as Nitrogen (N), Phosphorus (P), Potassium (K), and pH values. By analyzing this data, the system will predict the most appropriate crop for the given region. The dataset already includes NPK and pH values for the soil, along with the corresponding recommended crops.

Our recommendation system utilizes machine learning algorithms. Learning, in this context, involves enhancing performance or actions based on gained knowledge and experience. By implementing this concept, we aim to build a system that can autonomously learn and improve over time. Machine learning operates on the principle that a computer program learns from experience (E) in performing certain tasks (T) and improves its performance (P). Essentially, machine learning allows the machine to learn on its own without the need for explicit programming. It is a subset of artificial intelligence that enables systems to learn and enhance their capabilities automatically. Programs are generated by integrating input and output, and there are methodological differences between machine learning and statistics. One distinction is that machine learning focuses on optimization and performance, while statistics emphasizes inference.

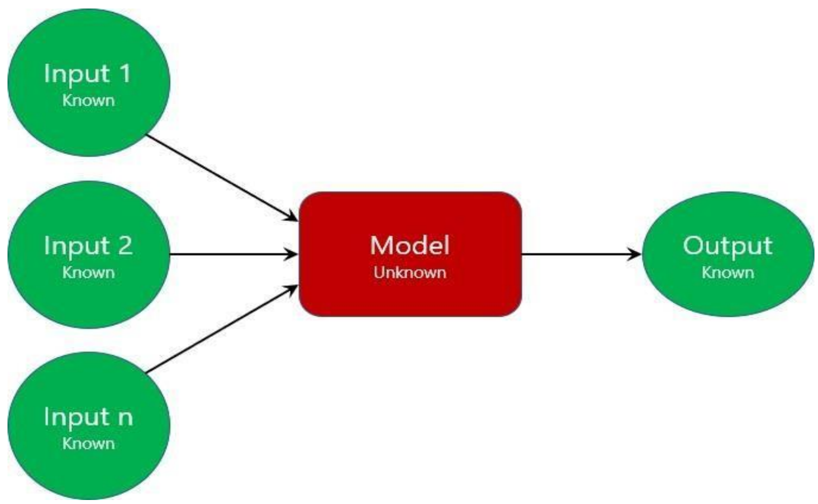


Fig. 1. General flow of Machine Learning

The diagram above illustrates the process for implementing a machine learning solution. In this approach, the computer is provided with the input data and the desired output or target solution. Based on this information, a program or model is developed to process new data or test values.

Machine learning encompasses three common techniques: supervised learning, unsupervised learning, and reinforcement learning. For this project, unsupervised learning clustering techniques are employed for prediction. In unsupervised learning, an AI system organizes unsorted data based on similarities and differences, even in the absence of predefined categories. Compared to supervised learning systems, unsupervised learning algorithms can handle more complex processing tasks. Employing unsupervised learning serves as a way to evaluate the capabilities of an AI system.

The goal of unsupervised learning is to allow algorithms to identify patterns within the training datasets and categorize input objects based on the identified patterns. By extracting useful information or features from the data, the algorithm analyzes the underlying structure of the datasets. Consequently, these algorithms are designed to generate specific outputs from unstructured inputs by identifying relationships between each sample or input object.

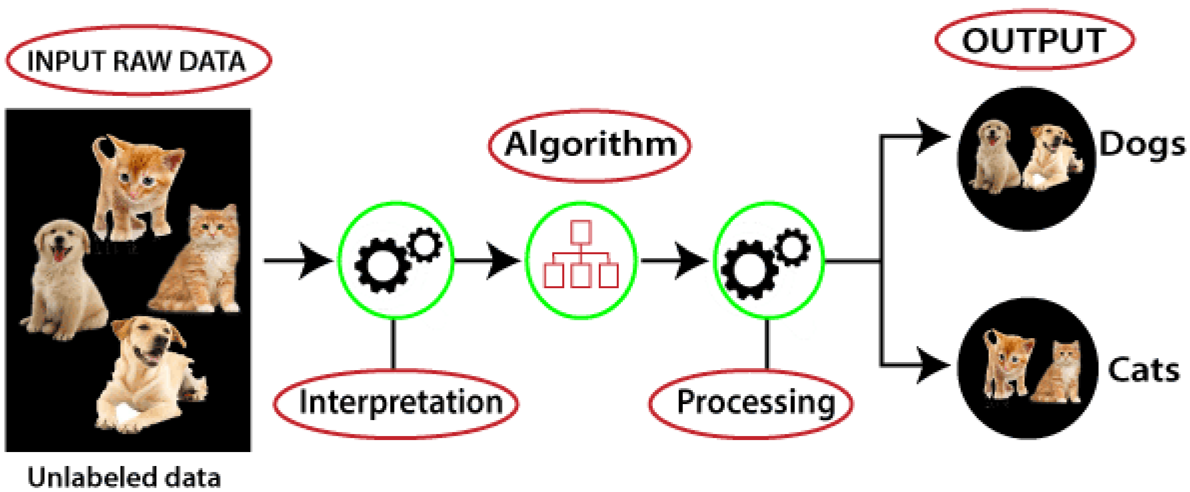


Fig. 2. Unsupervised Learning

In this context, we are working with unlabeled input data, meaning that the data is not pre-categorized and does not have corresponding outputs. The machine learning model is trained using this unlabeled input data. Initially, the model analyzes the raw data to identify hidden patterns within it. It then applies suitable algorithms such as k-means clustering, Logistic Regression, and others. By applying these algorithms, the model groups the data objects based on their similarities and differences.

Machine learning (ML) has significantly transformed the agricultural sector. As a subset of artificial intelligence, machine learning, in conjunction with big data technologies and high-performance computing, has created new opportunities for data-intensive research in the multi-disciplinary field of agri-technology. In agriculture, machine learning is not a mystical or magical concept; it comprises well-defined models that gather specific data and apply specific algorithms to achieve desired outcomes.

The objective of this project is to focus on precision farming, aiming to optimize productivity by understanding the requirements of crops in terms of climatic and soil conditions. The proposed system will assist farmers in dealing with the unpredictability of weather conditions.
The objective of this project is to focus on precision farming, aiming to optimize productivity by understanding the requirements of crops in terms of climatic and soil conditions. The proposed system will assist farmers in dealing with the unpredictability of weather conditions.

CHAPTER-2

Literature Survey

The paper proposes an analysis of exploratory data and discusses the design of various predictive models. A sample dataset is used, and different regression techniques are employed to analyze and understand each attribute. Algorithms like K Nearest Neighbors, Naïve Bayes, and KNN with Cross Validation are applied to the dataset to predict the most suitable crop for cultivation [1].

The developed system recommends the optimal crop for a specific plot of land based on weather parameters and soil characteristics, such as rainfall, temperature, humidity, and pH. This input data is utilized in machine learning predictive algorithms like Support Vector Machine (SVM) and Decision tree to identify patterns in the data and process it based on the input conditions. The system not only suggests the crop to the farmer but also provides recommendations on the required amount of nutrients for the predicted crop. Additionally, the system includes features such as displaying an estimated yield in q/acre, the required seed quantity for cultivation in kg/acre, and the market price of the crop [2].

This paper presents a solution for smart agriculture by monitoring agricultural fields, thereby assisting farmers in significantly increasing productivity. It also employs machine learning and prediction algorithms, such as Multiple Linear Regression, to identify patterns in the data and process it based on input conditions [3].

The objective of this research paper is to propose and implement a rule-based system for predicting crop yield production based on historical data. Association rule mining techniques are applied to agricultural data collected from 2000 to 2012 [4].

The project aims to create a prediction model that can estimate the maximum crop production rate prior to sowing. Information about the farmer's location (state, district), season, land area, and crop type is utilized to estimate crop production rates by applying a machine learning algorithm to the data [5].

Based on the literature survey:

- Approximately 60% of India's land is dedicated to agriculture to meet the needs of its 1.3 billion population. As the population continues to grow, there is a need for agricultural modernization to benefit farmers and address various challenges.
- The existing conventional system lacks technological connectivity and analysis for farmers.
- Farmers often rely on a "trial and error" method, conducting tests with different crops and water availability. After multiple attempts, they may achieve the desired crop production.
- Several papers have conducted surveys considering different parameters.
- While some systems aid in crop selection, their accuracy is limited.
- Certain papers utilize data mining techniques to predict crop yield based on climatic input parameters. However, climatic conditions alone are insufficient for accurate crop yield prediction.
- Various survey papers have analyzed different machine learning algorithms that can be used for crop prediction.
- Review papers have listed different algorithms for crop prediction, but no such system currently exists. Therefore, there is a need to implement such a system to benefit farmers.

DESIGN OF PROPOSED SYSTEM

The Proposed system will predict the most suitable crop for particular land based on soil contents and weather parameters such as Temperature, Humidity, soil PH and Rainfall.

TECHNOLOGY USED-

1. Machine Learning
2. Unsupervised Learning
3. K Means Clustering

TOOLS USED-

Python- Python is a popular interpreted, object-oriented, and high-level programming language known for its dynamic semantics. It offers built-in data structures, dynamic typing, and dynamic binding, making it suitable for rapid application development, scripting, and connecting components together. Python's syntax prioritizes readability, reducing the cost of program maintenance. It supports modules and packages, promoting modularity and code reuse. Python is freely available as an interpreter and comes with an extensive standard library for all major platforms.

Fig. 3. Python

The project utilizes several Python libraries, including:

A. Pandas: Pandas is a Python library used for data manipulation and analysis. It provides data structures and operations for working with numerical tables and time series.

B. NumPy: NumPy (Numerical Python) is a Python library used for working with arrays. It offers functions for linear algebra, Fourier transform, and matrix operations.

C. Matplotlib: Matplotlib is a Python library for creating 2D graphs and plots. Its pyplot module simplifies plotting by providing features to control line styles, font properties, and axes formatting. It supports various types of graphs and plots, such as histograms, bar charts, power spectra, and error charts. It is often used in conjunction with NumPy.

D. Seaborn: Seaborn is an open-source Python library built on top of Matplotlib. It facilitates data visualization and exploratory data analysis. Seaborn integrates well with Pandas and allows for easy customization of graphs.

E. ipywidgets: Ipywidgets (also known as Jupyter-widgets) are interactive HTML widgets for Jupyter notebooks and the IPython kernel. They enhance the interactivity of notebooks and enable users to control and visualize changes in data.

F. Scikit-learn (Sklearn): Scikit-learn is a robust and widely-used Python library for machine learning. It provides efficient tools for tasks such as classification, regression, clustering, and dimensionality reduction. Scikit-learn is built on top of NumPy, SciPy, and Matplotlib.

Jupyter Notebook: Jupyter Notebook is an open-source web application used for creating and sharing documents containing live code, equations, visualizations, and text. It serves various purposes, including data cleaning, numerical simulation, statistical modeling, data visualization, and machine learning.

Fig. 4. Jupyter

FLOW IN THE PROPOSED SYSTEM:

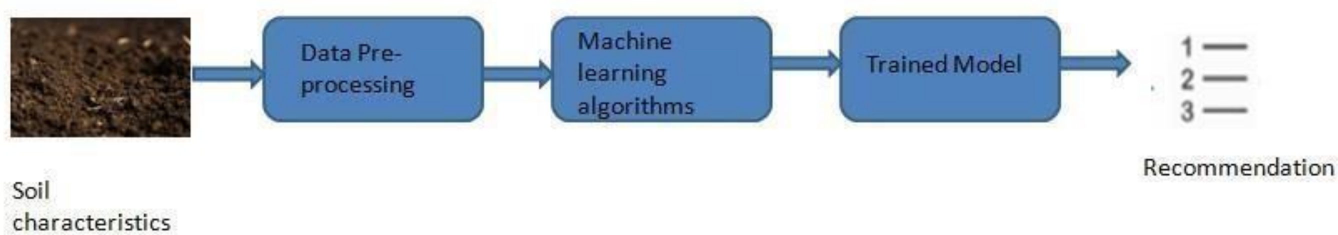


Fig. 5. Flow in the proposed system

A. Acquisition of training dataset:

The accuracy of a machine learning algorithm can be influenced by the number of parameters used and the quality of the dataset. In our project, the dataset consists of attributes such as N, P, K, and pH values representing different types of soils, along with the corresponding crops that can be grown in those soils.

By employing an appropriate machine learning algorithm, we can train the dataset to predict the most suitable crop based on the given input parameters. The dataset used in our project was sourced from Kaggle and is titled "Crop_recommendation." It is provided in the form of a CSV (Comma Separated Value) file. A CSV file is a plain text file where each line represents a data record, and the values within each record are separated by commas. This file format is commonly used for storing tabular data consisting of numbers and text.

To utilize this dataset in Python, we need to import the .csv file. Once the .csv file is imported, we can read its contents using the following Python command:

```
In [11]: # Lets check the head of the dataset
         data.head()
```

Out[11]:

	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

B. Data preprocessing:

Data preprocessing is an essential step that consists of two main components. The first step is Data Cleaning, where the original dataset is examined for missing values. Missing values can significantly impact the quality of the data and subsequently affect the performance of the model. In order to address this issue, we identify and remove any instances of missing values in the dataset. Missing values are typically represented by a dot or a specific placeholder in the dataset. To mitigate their impact, we replace these missing values with large negative values, which are considered outliers by the model. The second step in data preprocessing involves generating the class labels. Since we are utilizing a supervised learning approach, each entry in the dataset should be associated with a corresponding class label. This step is crucial in preparing the data for the subsequent modeling and prediction stages. During the preprocessing step, we create the necessary class labels to facilitate the supervised learning process.

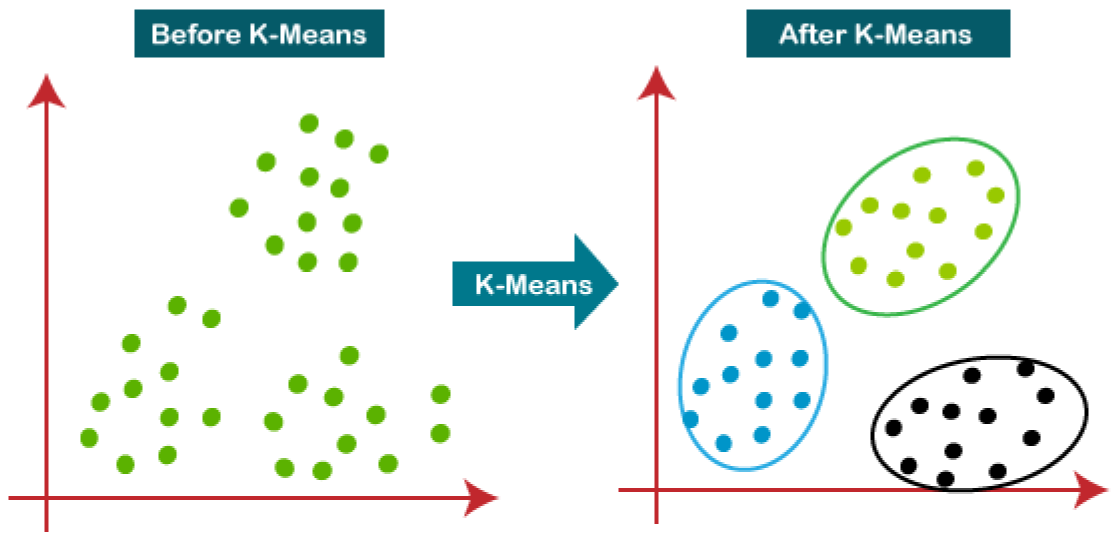
```
In [12]: # Lets check if there is any missing value present in the dataset
         data.isnull().sum()

Out[12]: N              0
         P              0
         K              0
         temperature    0
         humidity       0
         ph             0
         rainfall       0
         label          0
         dtype: int64
```

To handle missing values in the dataset, we can utilize either the Fill-NA function or the drop-NA function available in the pandas library of Python. The Fill-NA function replaces missing values with statistical measures like mean, median, or mode, while the drop-NA function eliminates the rows or columns containing missing values from the dataset.

C. Machine Learning Algorithm:

The project employs various machine learning algorithms, including:

K-Means Clustering: K-Means Clustering is an unsupervised learning algorithm that groups unlabeled datasets into distinct clusters. The algorithm determines the number of pre-defined clusters (K) to create during the process. For example, if K=2, there will be two clusters; if K=3, there will be three clusters, and so on.



- It is an iterative algorithm that divides the unlabelled dataset into different clusters in such a way that each dataset belongs only one group that has similar properties. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

- The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best possible cluster. The value of k should be predetermined in this algorithm.

- The k-means clustering algorithm mainly performs two tasks:
  1) Determines the best value for K center points or centroids by an iterative process.
  2) Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

  Hence, each cluster has data points with some commonalities, and it is away from other clusters.

- If k is given, the K-means algorithm can be executed in the following steps:

i. Partition of objects into k non-empty subsets
ii. Identifying the cluster centroids (mean point) of the current partition.
iii. Assigning each point to a specific cluster
iv. Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
v. After re-allotting the points, find the centroid of the new cluster formed.

- The below diagram explains the working of the K-means Clustering Algorithm:

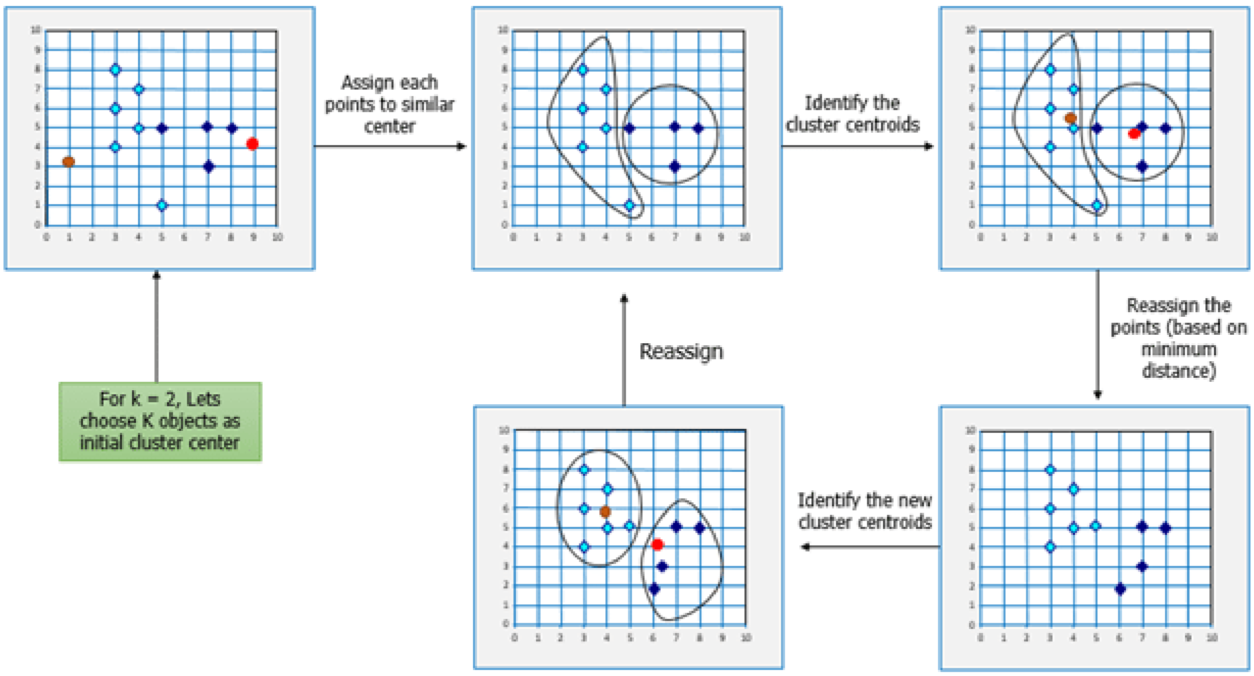


Fig. K-Means Clustering Algorithm Flow

So for a particular example, each iteration can be shown as:

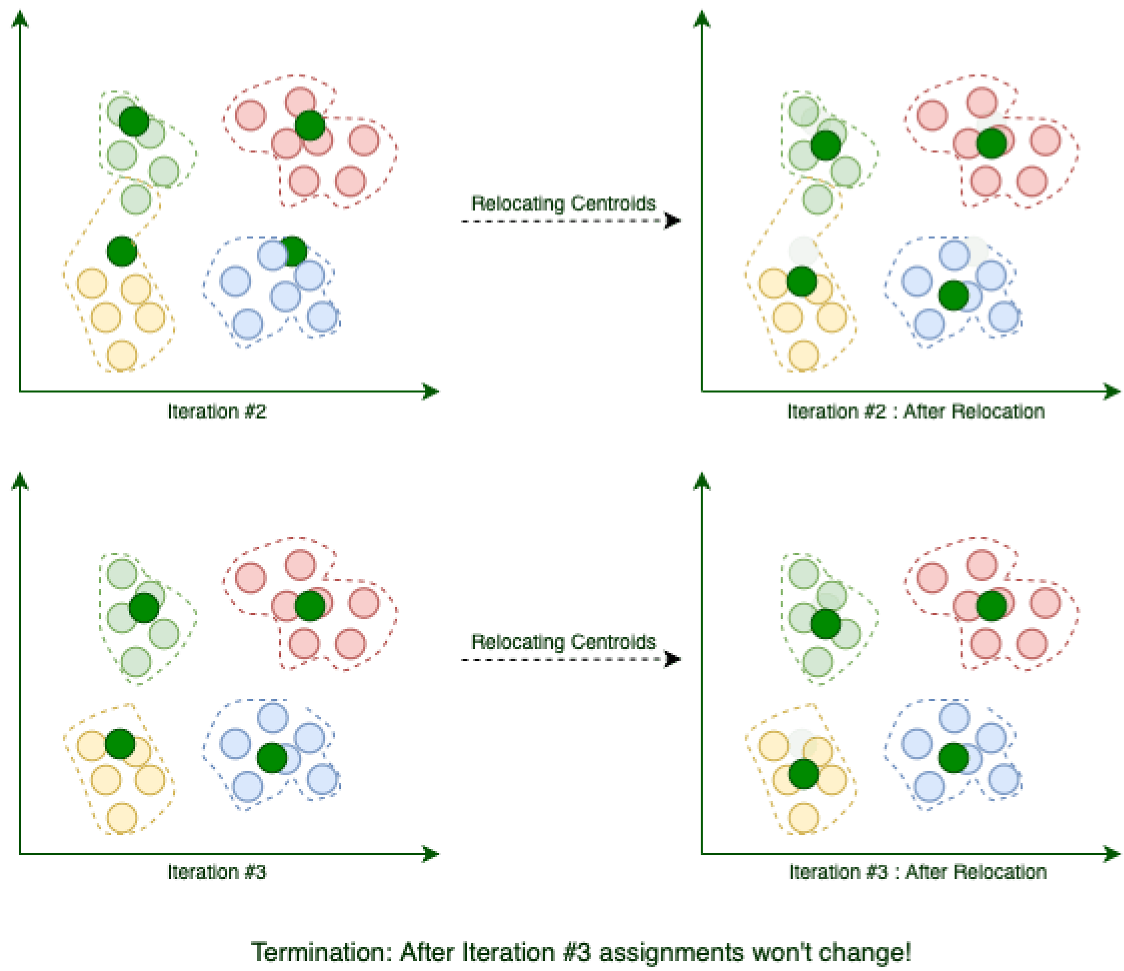


Fig: Iterations in K-Means Clustering

- The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters, but here we are discussing the most appropriate method to find the number of clusters or value of K. The method is given below:
-

ELBOW METHOD:

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS = \sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} distance(P_i\ C_3)^2$$

In the above formula of WCSS,

$\sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2$: : is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the 2 others.

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method.

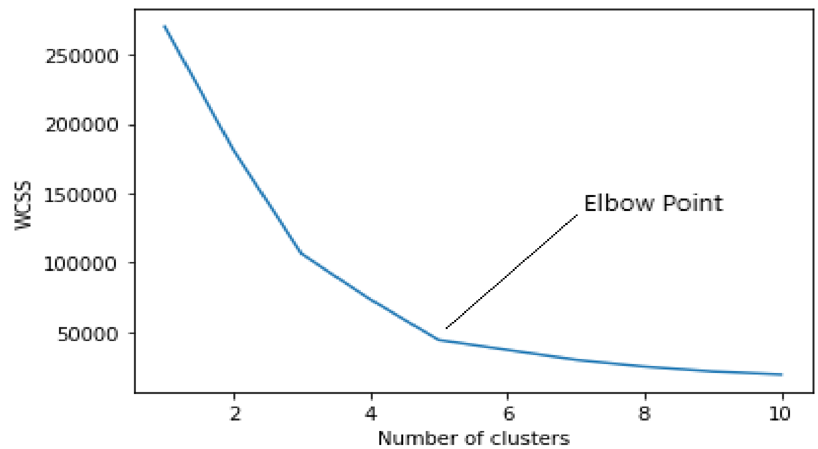
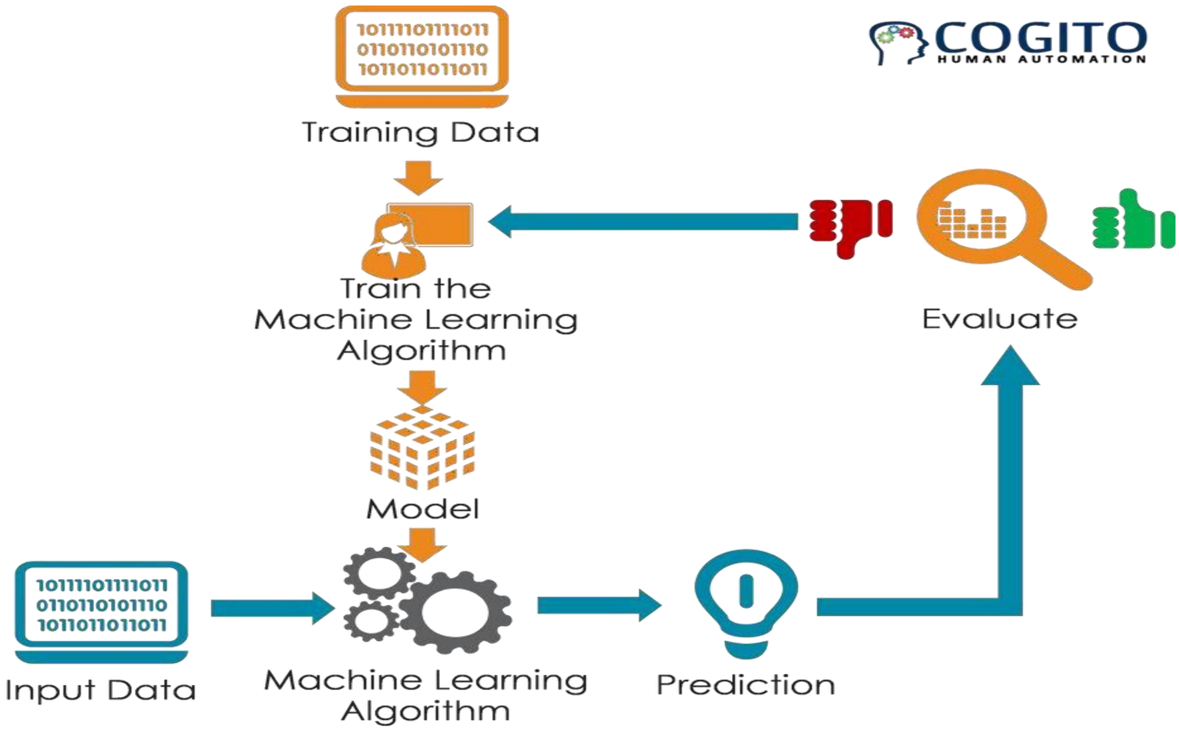


Fig. Elbow Method

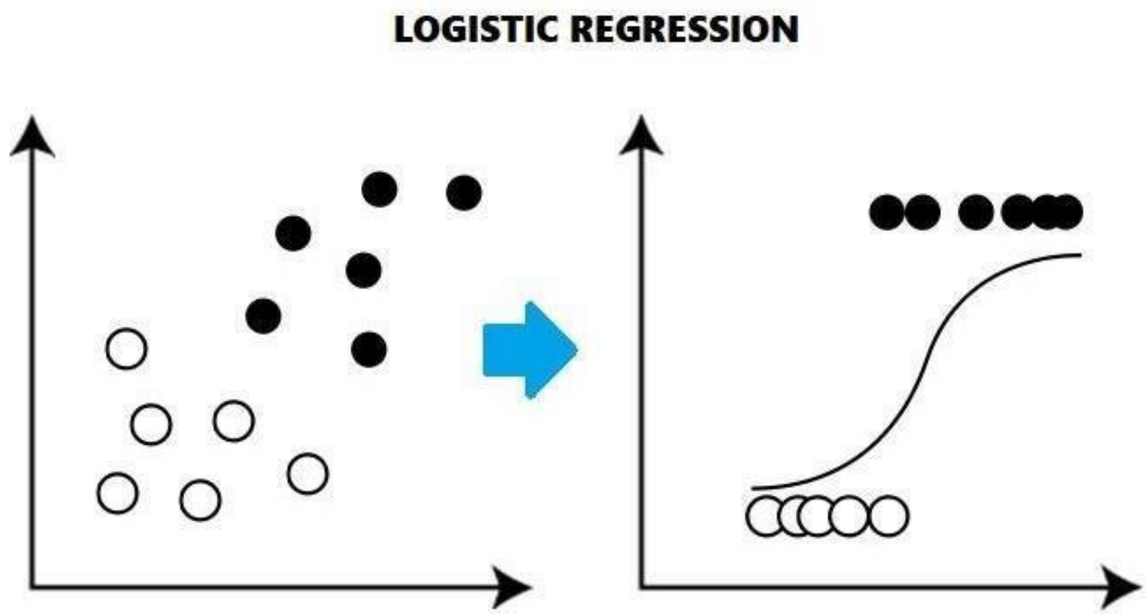
Trained model:

Trained models are obtained after applying the dataset to the machine learning algorithms. Our project suggests a crop prediction system which is based on the K-Means Clustering algorithm. Soil properties such Nitrogen, Phosphorus, Potassium, pH value, etc. are given as input to the model. The algorithm will look for a crop which will have the value closest to the inputted values. The output will be all the crops which are suitable for the inputted values. The result is calculated based on the most comparable or closest values. Because of its higher convergence speed and simplicity this algorithm is preferred over other algorithms. The input for the algorithm is the soil properties such as Nitrogen, Phosphorus, Potassium, pH value, etc.



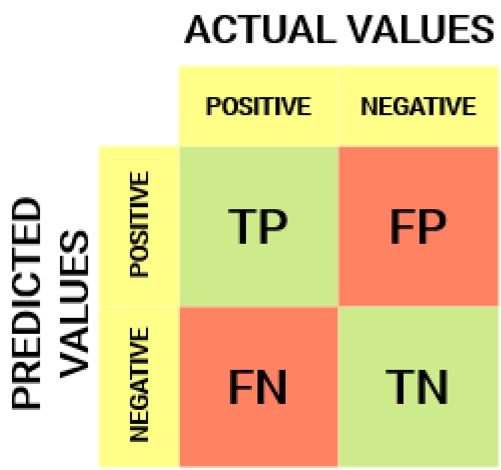
Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables



The process of building machine learning models is based on a feedback loop, where the model is developed, feedback is collected through metrics, and improvements are made until the desired level of accuracy is achieved. Evaluation metrics play a significant role in determining the model's performance, as they can differentiate between the various results produced by the model.

However, building a predictive model is not sufficient; we also need to check the accuracy of the model before making any predictions. In our project, we have used the Confusion Matrix as the evaluation metric. This is an N x N matrix that is used to evaluate the performance of a classification model, where N is the number of target classes. The Confusion Matrix compares the actual target values with the predicted values of the model, providing a comprehensive understanding of the model's performance and the types of errors it is making.



- The target variable has two values: Positive or Negative.
- The columns represent the actual values of the target variable.
- The rows represent the predicted values of the target variable.
- **True Positive (TP):**
  i. The predicted value matches the actual value
  ii. The actual value was positive and the model predicted a positive value
- **True Negative (TN):**
  i. The predicted value matches the actual value
  ii. The actual value was negative and the model predicted a negative value
- **False Positive (FP) – Type 1 error:**
  i. The predicted value was falsely predicted
  ii. The actual value was negative but the model predicted a positive value
- **False Negative (FN) – Type 2 error:**
  i. The predicted value was falsely predicted
  ii. The actual value was positive but the model predicted a negative value

Chapter 3

Functionality/Working of Project

IMPLEMENTATION

```
#for manipulations
import numpy as np
import pandas as pd

#for data visualization
import matplotlib.pyplot as plt
import seaborn as sns

#for interactivity
from ipywidgets import interact

#Let's read the data set
data = pd.read_csv("Crop_recommendation.csv")
#Let's Check the shape of the dataset
print("Shape of the dataset: ", data.shape)
#Let's check the head of the dataset
data.head()
#Let's check if there is any missing value present in the dataset i.e data cleaning
#isnull() returns boolean value and sum() returns the count
data.isnull().sum()
#Let's check the unique crops present in this Dataset and their occurence
data['label'].value_counts()
```

```
#Let's check the unique crops present in this Dataset and their occurence
data['label'].value_counts()
#Let's check the summary statistics for each Factor

@interact
def summary(crops = list(data['label'].value_counts().index)):
    x = data[data['label'] == crops]
    print("---------------------------------------------")
    print("Statistics for Nitrogen")
    print("Minimum Nitrogen required : ", x['N'].min())
    print("Average Nitrogen required : ", x['N'].mean())
    print("Maximum Nitrogen required : ", x['N'].max())
    print("---------------------------------------------")
    print("Statistics for Phosphorus")
    print("Minimum Phosphorus required : ", x['P'].min())
    print("Average Phosphorus required : ", x['P'].mean())
    print("Maximum Phosphorus required : ", x['P'].max())
    print("---------------------------------------------")
    print("Statistics for Potassium")
    print("Minimum Potassium required : ", x['K'].min())
    print("Average Potassium required : ", x['K'].mean())
    print("Maximum Potassium required : ", x['K'].max())
    print("---------------------------------------------")
    print("Statistics for Temperature")
    print("Minimum Temperature required : {0:.2f}".format(x['temperature'].min()))
    print("Average Temperature required : {0:.2f}".format(x['temperature'].mean()))
    print("Maximum Temperature required : {0:.2f}".format(x['temperature'].max()))
    print("---------------------------------------------")
    print("Statistics for Humidity")
    print("Minimum Humidity required : {0:.2f}".format(x['humidity'].min()))
    print("Average Humidity required : {0:.2f}".format(x['humidity'].mean()))
```

```
    print("Maximum Humidity required : {0:.2f}".format(x['humidity'].max()))
    print("---------------------------------------------")
    print("Statistics for PH")
    print("Minimum PH required : {0:.2f}".format(x['ph'].min()))
    print("Average PH required : {0:.2f}".format(x['ph'].mean()))
    print("Maximum PH required : {0:.2f}".format(x['ph'].max()))
    print("---------------------------------------------")
    print("Statistics for Rainfall")
    print("Minimum Rainfall required : {0:.2f}".format(x['rainfall'].min()))
    print("Average Rainfall required : {0:.2f}".format(x['rainfall'].mean()))
    print("Maximum Rainfall required : {0:.2f}".format(x['rainfall'].max()))
#Let's check the summary statistics for each of the Crops

@interact
def compare(conditions = ['N', 'P', 'K', 'temperature', 'humidity', 'ph', 'rainfall']):
    print("Average value for", conditions, "is {0:.2f}".format(data[conditions].mean()))
    print("---------------------------------------------")
    print("Rice : {0:.2f}".format(data[(data['label'] == 'rice')][conditions].mean()))
    print("Black Grams : {0:.2f}".format(data[(data['label'] == 'blackgram')][conditions].mean()))
    print("Banana : {0:.2f}".format(data[(data['label'] == 'banana')][conditions].mean()))
    print("Jute : {0:.2f}".format(data[(data['label'] == 'jute')][conditions].mean()))
    print("Coconut : {0:.2f}".format(data[(data['label'] == 'coconut')][conditions].mean()))
    print("Apple : {0:.2f}".format(data[(data['label'] == 'apple')][conditions].mean()))
    print("Papaya : {0:.2f}".format(data[(data['label'] == 'papaya')][conditions].mean()))
    print("Muskmelon : {0:.2f}".format(data[(data['label'] == 'muskmelon')][conditions].mean()))
    print("Grapes : {0:.2f}".format(data[(data['label'] == 'grapes')][conditions].mean()))
    print("Watermelon : {0:.2f}".format(data[(data['label'] == 'watermelon')][conditions].mean()))
    print("Kidney Beans : {0:.2f}".format(data[(data['label'] ==
'kidneybeans')][conditions].mean()))
    print("Mung Beans : {0:.2f}".format(data[(data['label'] == 'mungbean')][conditions].mean()))
    print("Oranges : {0:.2f}".format(data[(data['label'] == 'orange')][conditions].mean()))
```

```
    print("Chick Peas : {0:.2f}".format(data[(data['label'] == 'chickpea')][conditions].mean()))
    print("Lentils : {0:.2f}".format(data[(data['label'] == 'lentil')][conditions].mean()))
    print("Cotton : {0:.2f}".format(data[(data['label'] == 'cotton')][conditions].mean()))
    print("Maize : {0:.2f}".format(data[(data['label'] == 'maize')][conditions].mean()))
    print("Moth Beans : {0:.2f}".format(data[(data['label'] == 'mothbeans')][conditions].mean()))
    print("Pigeon Peas : {0:.2f}".format(data[(data['label'] == 'pigeonpeas')][conditions].mean()))
    print("Mango : {0:.2f}".format(data[(data['label'] == 'mango')][conditions].mean()))
    print("Pomegranate : {0:.2f}".format(data[(data['label'] == 
'pomegranate')][conditions].mean()))
    print("Coffee : {0:.2f}".format(data[(data['label'] == 'coffee')][conditions].mean()))
@interact
def compare(conditions = ['N', 'P', 'K', 'temperature', 'humidity', 'ph', 'rainfall']):
    print("Crops which require greater than average", conditions, '\n')
    print(data[data[conditions] > data[conditions].mean()]['label'].unique())
    print("------------------------------------------------------------")
    print("Crops which require equal to or smaller than average", conditions, '\n')
    print(data[data[conditions] <= data[conditions].mean()]['label'].unique())
plt.figure(figsize=(15,8))

plt.subplot(2, 4,  1)
sns.distplot(data['N'], color = 'lightgrey')
plt.xlabel("Ratio of Nitrogen", fontsize = 12)
plt.grid()

plt.subplot(2, 4,  2)
sns.distplot(data['P'], color = 'lightblue')
plt.xlabel("Ratio of Phosphorus", fontsize = 12)
plt.grid()

plt.subplot(2, 4, 3)
sns.distplot(data['K'], color = 'darkblue')
```

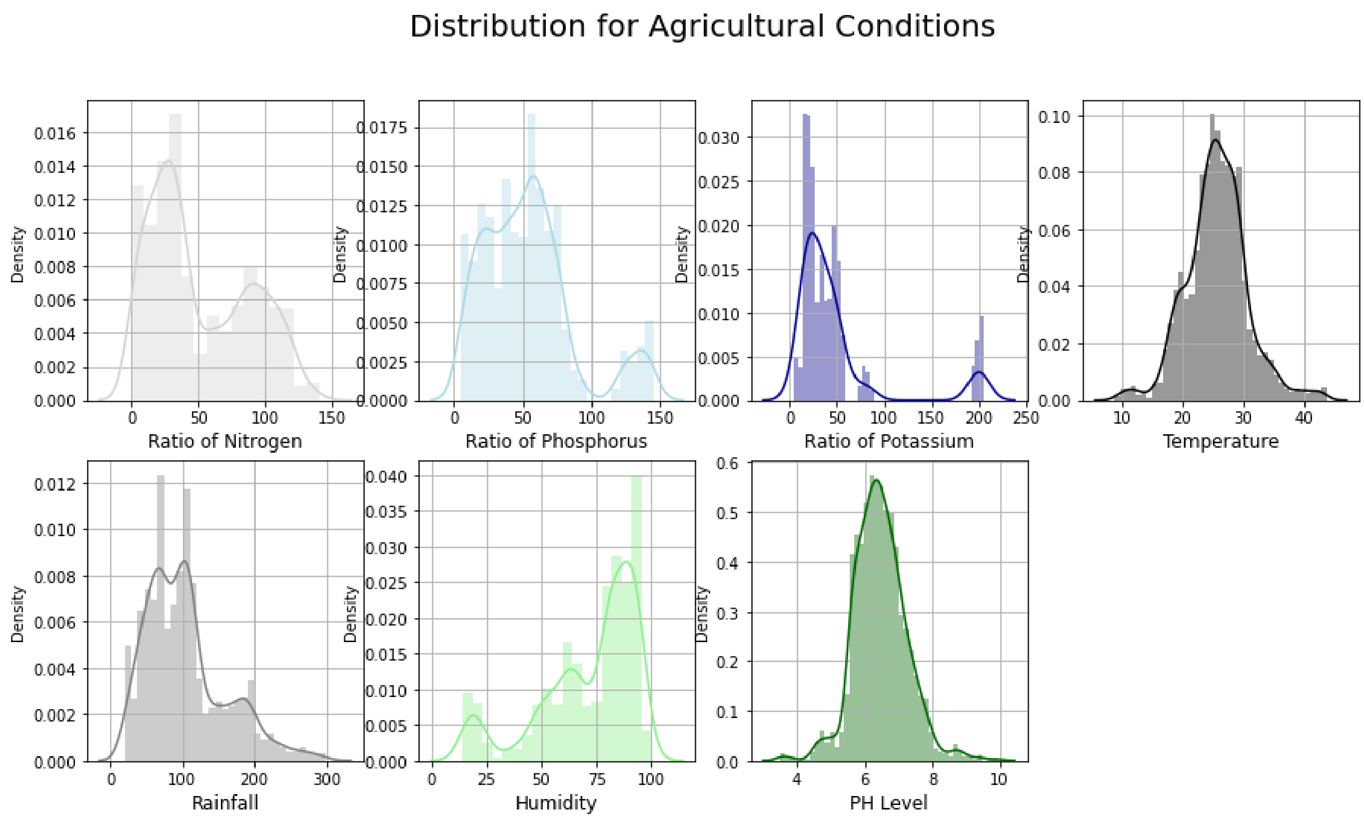
```
plt.xlabel("Ratio of Potassium", fontsize = 12)
plt.grid()

plt.subplot(2, 4, 4)
sns.distplot(data['temperature'], color = 'black')
plt.xlabel("Temperature", fontsize = 12)
plt.grid()

plt.subplot(2, 4, 5)
sns.distplot(data['rainfall'], color = 'grey')
plt.xlabel("Rainfall", fontsize = 12)
plt.grid()

plt.subplot(2, 4, 6)
sns.distplot(data['humidity'], color = 'lightgreen')
plt.xlabel("Humidity", fontsize = 12)
plt.grid()

plt.subplot(2, 4, 7)
sns.distplot(data['ph'], color = 'darkgreen')
plt.xlabel("PH Level", fontsize = 12)
plt.grid()

plt.suptitle("Distribution for Agricultural Conditions", fontsize = 20)
plt.show()
# Some Interseting Patterns

print("Crops which requires very High rainfall:",data[data['rainfall'] > 200]['label'].unique())
print("Crops which requires very Low rainfall:",data[data['rainfall'] < 40]['label'].unique())
print("Crops which requires very High ratio of Nitrogen Content in soil :",data[data['N'] >
120]['label'].unique())
```

Chapter 4

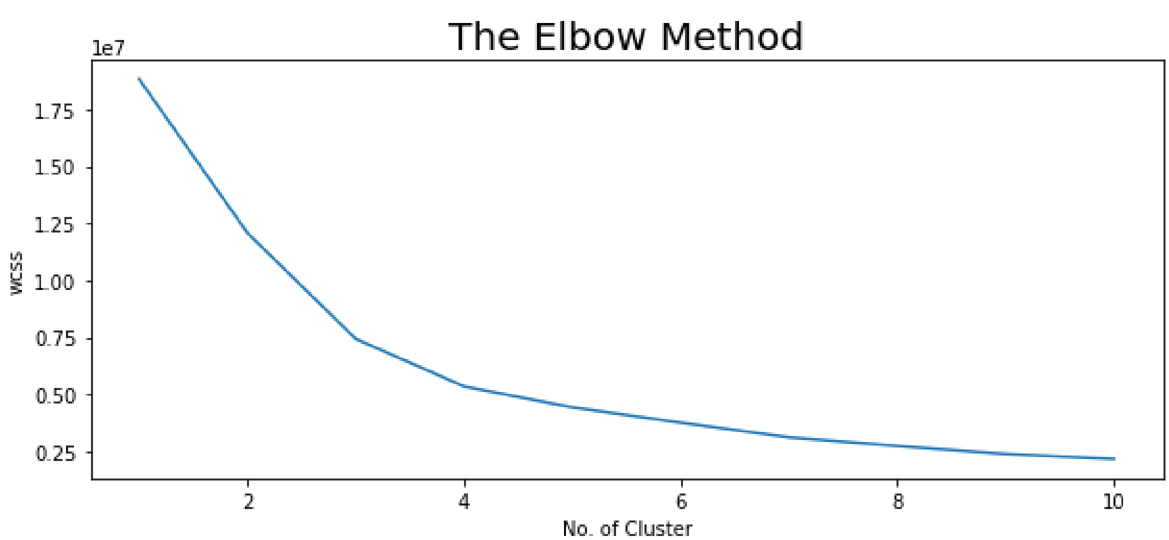
Results and Discussion

The proposed system aims to recommend the most suitable crop for a specific land based on parameters such as annual rainfall, temperature, humidity, and soil pH. The user is required to enter additional parameters. The system provides output that includes the recommended crop, required seeds per acre, market price, approximate yield of the recommended crop, and the required NPK values based on the inputted NPK values. Our project offers several advancements compared to previous papers.One significant improvement is the utilization of a large dataset, enabling us to provide details for a greater number of crops and predict suitable crops for various soil conditions. We have implemented different machine learning models in our project and compared their accuracies to select the best model for accurate predictions. This approach allows us to obtain results quickly.

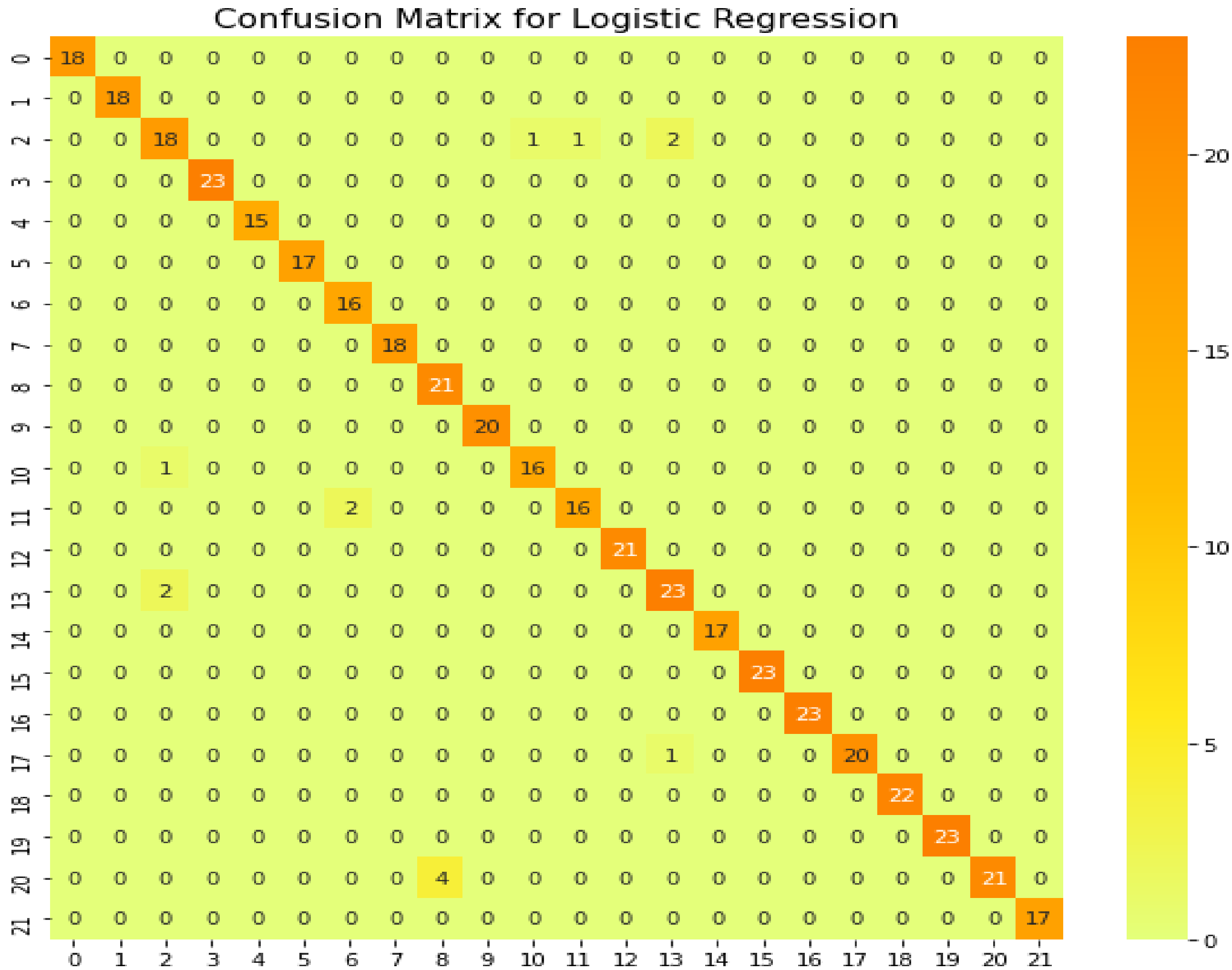
Our project is designed to be easily accessible to all farmers, and we can incorporate additional features as technology advances. By utilizing a machine learning model with high predictive accuracy, our project yields the best results. Furthermore, our recommendation system can adapt to changes in soil composition caused by natural disasters like floods and soil erosion, providing a reliable method for predicting suitable crops under altered soil conditions.



Distribution for Agricultural Conditions
Ratio of Nitrogen
Ratio of Phosphorus
Ratio of Potassium
Temperature
Rainfall
Humidity
PH Level
Density



The Elbow Method
1e7
wcss
No. of Cluster



Here we have taken a sample input providing all the required parameters and the suggested crop that we get is: **‘RICE’** as can be seen in the given output:

jupyter Untitled Last Checkpoint: Last Tuesday at 00:46 (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

```
      lentil       0.94      0.94      0.94        17
       maize       0.94      0.89      0.91        18
       mango       1.00      1.00      1.00        21
   mothbeans       0.88      0.92      0.90        25
    mungbean       1.00      1.00      1.00        17
   muskmelon       1.00      1.00      1.00        23
      orange       1.00      1.00      1.00        23
      papaya       1.00      0.95      0.98        21
  pigeonpeas       1.00      1.00      1.00        22
 pomegranate       1.00      1.00      1.00        23
        rice       1.00      0.84      0.91        25
  watermelon       1.00      1.00      1.00        17

    accuracy                           0.97       440
   macro avg       0.97      0.97      0.97       440
weighted avg       0.97      0.97      0.97       440
```

In [19]:

```
# Our model is ready,now we use model to predict Crop name
data.head()
```

Out[19]:

	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

In [20]:

```
prediction = model.predict((np.array([[90,
                                       40,
                                       40,
                                       20,
                                       80,
                                       7,
                                       200]])))
print("The suggested Crop for Given Climatic condition is :", prediction)
```

```
The suggested Crop for Given Climatic condition is : ['rice']
```

Chapter 5

CONCLUSION AND FUTURE SCOPE

CONCLUSION

Currently, many farmers lack effective utilization of technology and data analysis, resulting in potential mistakes in crop selection that can negatively impact their income. To address this issue and minimize losses, we have developed a user-friendly system specifically designed for farmers. This system predicts the most suitable crop for a given plot of land and provides essential information such as recommended nutrient additions, required seed quantities, expected yield, and market prices.By leveraging this system, farmers can make informed decisions when selecting crops for cultivation, leading to improved agricultural practices and fostering innovation in the agricultural sector. This initiative aims to enhance farmers' decision-making processes and contribute to the overall development of the agricultural industry.

FUTURE SCOPE

To gather the necessary data, we utilize GPS coordinates to determine the location of the land. Access to the government's Rain Forecasting System enables us to predict suitable crops based on the provided GPS location. Additionally, we can develop models to mitigate food shortages and prevent overproduction.

In the future, this system can be enhanced by incorporating Internet of Things (IoT) technology to obtain real-time soil data. By installing sensors on the farm, information regarding current soil conditions can be collected, thereby improving the accuracy and reliability of the system's predictions. This integration of IoT facilitates smart farming practices, enabling farmers to make more informed decisions.

References

[1] Ashwani kumar Kushwaha, Swetabhattachrya “crop yield prediction using agro algorithm in hatoop”

[2] Girish L, Gangadhar S, Bharath T R, Balaji K S, Abhishek K T “Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning”.

[3] Rahul Katarya, Ashutosh Raturi, Abhinav Mehndiratta, Abhinav Thap-per “Impact of Machine Learning Techniques in Precision Agriculture”.

[4]] Pijush Samui, Venkata Ravibabu Mandla, Arun Krishna and Tarun Teja “Prediction of Rainfall Using Support Vector Machine and Relevance Vector Machine”.

[5] Himani Sharma, Sunil Kumar “A Survey on Decision

[6] Pavan Patil, Virendra Panpatil, Prof. Shrikant Kokate “Crop Prediction System using Machine Learning Algorithms”.

[7] Prof. D.S. Zingade ,Omkar Buchade ,Nilesh Mehta ,Shubham Ghodekar ,Chandan Mehta “Crop Prediction System using Machine Learning”

[8] Kevin Tom Thomas , Varsha S , Merin Mary Saji , Lisha Varghese , Er. Jinu Thomas “Crop Prediction using Machine Learning”.

[9] Nischitha K, Dhanush Vishwakarma, Mahendra N, Manjuraju M.R, Ashwini “Crop Prediction using Machine Learning Approaches”

[10] “CROP YIELD PREDICTION USING K-MEANS CLUSTERING” Capstone Design Spring 2020 Amine Bouighoulouden Dr. Ilham Kissani.

[11] “Crop prediction based on soil and environmental characteristics using feature selection techniques” by A. Suruliandi,G. Mariammal S.P. Raj

[12] “Crop Yield Prediction Using Supervised Machine Learning Algorithm” Hardik Joshi, Monika Gawade, Manasvi Ganu, Prof. Priya Porwal

[13] P. Johri, J. N. Singh, A. Sharma and D. Rastogi, quot;Sustainability of Coexistence of Humans and Machines: An Evolution of Industry 5.0 from Industry 4.0,quot; 2021 10th International Conference on System Modeling amp; Advancement in Research Trends (SMART), MORADABAD, India, 2021, pp.410-414, doi: 10.1109/SMART52563.2021.9676275.

[14] Johri, P., Singh, J.N., Khatri, S.K., Bagchi, A., Rajesh, E. (2022). Role of Satellites in Agriculture. In: Moh, M., Sharma, K.P., Agrawal, R., Garcia Diaz, V. (eds) Smart IoT for Research and Industry. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-71485-7 6

[15] Kevin Tom Thomas , Varsha S , Merin Mary Saji , Lisha Varghese , Er. Jinu Thomas “Crop Prediction using Machine Learning”.

[16] Nischitha K, Dhanush Vishwakarma, Mahendra N, Manjuraju M.R, Ashwini “Crop Prediction using Machine Learning Approaches”

[17] “CROP YIELD PREDICTION USING K-MEANS CLUSTERING” Capstone Design Spring 2020 Amine Bouighoulouden Dr. Ilham Kissani.

[18] “Crop prediction based on soil and environmental characteristics using feature selection techniques” by A. Suruliandi,G. Mariammal & S.P. Raja

[19] "Crop Yield Prediction Using Supervised Machine Learning Algorithm" Hardik Joshi, Monika Gawade, Manasvi Ganu, Prof. Priya Porwal.

[20] https://github.com/yrnigam/Crop_Sowing_Prediction

Gmail Nafish Alam <md.nafish178@gmail.com>

Fwd: Acceptance Notification 5th IEEE ICAC3N-23 & Registration: Paper ID 140

1 message

Dr.J.N. Singh <singhjn2000@gmail.com> Thu, May 4, 2023 at 2:57 PM
To: md.nafish178@gmail.com

Registration fee details are available @ https://icac3n.in/register.

pls deposit fee Rs.5500 after having a discount
and send me a fee receipt.
Bank details are given in the link

Thanks
Dr.J.N.Singh
Professor

---------- Forwarded message ---------
From: **Microsoft CMT** <email@msr-cmt.org>
Date: Thu, May 4, 2023 at 10:47 AM
Subject: Acceptance Notification 5th IEEE ICAC3N-23 & Registration: Paper ID 140
To: jitendra singh <singhjn2000@gmail.com>

Dear jitendra singh,
Galgotias University

Greetings from ICAC3N-23 ...!!!

Congratulations....!!!!!

On behalf of the 5th ICAC3N-23 Program Committee, we are delighted to inform you that the submission of "Paper ID- 140 " titled " Crop Prediction Using Machine Learning " has been accepted for presentation and further publication with IEEE at the ICAC3N- 23 subject to incorporate the reviewers and editors comments in your final paper. All accepted papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

For early registration benefit please complete your registration by clicking on the following Link:

https://forms.gle/8e6RzNbho7CphnYN7 on or before 10 May 2023.

Registration fee details are available @ https://icac3n.in/register.

You must incorporate following comments in your final paper submitted at the time of registration for consideration of publication with IEEE:

The topic chosen "Crop Prediction Using Machine Learning" is interesting and relevent.
The formatting of paper is not proper. Formatting must be strictly as per template. Otherwise it will not be published.
Author list formatting is not proper. All authors information must be complete and should be in proper format and as per the sequence desired.
Citation of references within the content is not proper. Make only relevant citation. All references must be cited in content properly.
References are not in proper format. Format and assign number to the references properly.
An overview of paper is desired to eradicate typo and grammatical error.
Formatting and Quality of figures is poor. Use High quality images.
All figure and tables must be properly captioned and numbered as per IEEE conference template.
Add a comparison table in literature review section with the work already done in this filed.
Conclusion and result section needs to be improved and require better explanation.

Editor Comments/Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Transfer of E-copyright to IEEE and Presenting paper in conference is compulsory for publication of

paper in IEEE.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences. Plagiarism must be less then 20% (checked through Turnitin).
5. Change in paper title, name of authors or affiliation of authors will not be allowed after registration of papers.
6. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
7. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp) or write us at icac3n23@gmail.com.

Regards:
Organizing committee
ICAC3N - 2023