Journal of International Commercial Law and Technology
Print ISSN: 1901-8401

Website: https://www.jiclt.com/

Article

Utilizing Supervised and Unsupervised Machine Learning Techniques for Crop Yield Prediction, Pest Detection, And Precision Farming

Article History:

Name of Author:
Dr. Inumarthi V. Srinivas[1], Megha Dhotay[2], Dr. J. Sridevi[3], Ritika Sanwal[4], Nikita Jain[5]

Affiliation:
[1]*Associate Professor, Prin. L. N. Welingkar Institute of Management Development and Research (PGDM), Lakhamsi Napoo Rd, Opposite Matunga Gymkhana, Matunga East, Mumbai - 400019*
[2]*Lecturer, Department of Polytechnic and Skill Development, Dr. Vishwanath Karad MIT World Peace University Pune, Maharashtra, India*
[3]*Associate Professor, School of Commerce, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India - 600062*
[4]*Assistant Professor, Department of Media and Mass Communication, Graphic Era Hill University, Haldwani Campus, Uttarakhand*
[5]*Assistant Professor, Faculty of Engineering, Teerthanker Mahaveer University, Moradabad, India*

Corresponding Author:
Dr. Inumarthi V. Srinivas

How to cite this article:
Dr. Inumarthi V. Srinivas, et, al, *Utilizing Supervised and Unsupervised Machine Learning Techniques for Crop Yield Prediction, Pest Detection, And Precision Farming. J Int Commer Law Technol. 2026;7(1):660–669.*

Received: *28-01-2026*
Revised: *16-02-2026*
Accepted: *20-02-2026*
Published: *02-03-2026*

Abstract: Crop recommendations and agricultural decision-making can be improved with the use of machine learning methods. It is analysed using the 2,200 records on soil nutrients and environmental factors including "temperature, pH, humidity, rainfall, phosphorus, nitrogen, as well as potassium" that Kaggle gave. Exploratory data analysis determines trends of crop appropriateness. Crop classification is done by the supervised models Logistic Regression and Random Forest to give an accuracy of 0.9727 and 0.9955, respectively. Clustering unsupervised procedures, such as K-Means and Agglomerative Clustering, show silhouette scores of 0.3229 and 0.3468 between clusters representing environmental groups.

Keywords: Machine Learning, Precision Agriculture, Crop Recommendation, Random Forest, Logistic Regression, K-Means Clustering, Agglomerative Clustering, Agricultural Data Analysis.

INTRODUCTION

Agriculture is very important in bringing about food security, economic stabilization, and sustainable development.

The dynamism in the climate conditions, soil fertility, and water availability presents farmers with difficulties in choosing the right crops and enhancing productivity. Logistic Regression and Random Forest are some of the supervised learning models that are used to classify crops based on environmental factors. The methods of unsupervised learning, including K-Means and Agglomerative Clustering, are employed to identify latent groupings of agricultural data. Combining both procedures would offer an in-depth insight into the crop patterns and environmental impact.

Aim and Objectives

This study will employ machine learning to predict crops and analyze the patterns of the agricultural field by using both supervised and unsupervised machine learning. The aims consist of analyzing agricultural data with the help of the exploratory data analysis, creating models to classify data to predict the right crops, clustering data with the help of the cluster methods, measuring the performance of the models with the help of the relevant metrics, and exploring how these methods can be used in the context of the precision farming practice.

Literature Review

Machine Learning in Crop Prediction

Several publications indicate that algorithms of supervised learning are useful in crop prediction. The use of the Logistic Regression is usually applied in matters of classification since it is easy and understandable. It assists in determining the likelihood of a crop being within a certain category depending on the environmental characteristics. Nevertheless, more complicated agronomical data may demand sophisticated models.

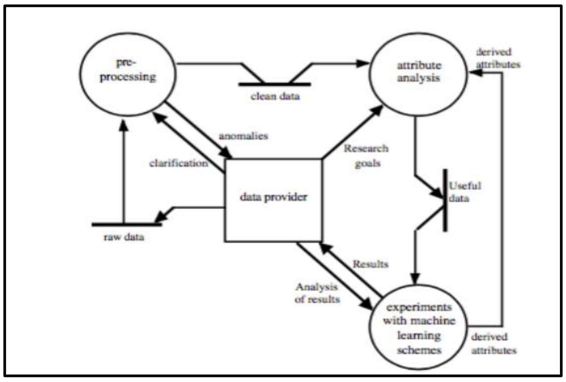


Fig 1: Process model for a machine learning

The random forest is also utilized in the field of agriculture research since it addresses non-linear relations and massive data with ease. It minimizes overfitting, as it trains a combination of several decision trees and enhances the prediction (Amini and Rahmani, 2023). All the studies carried out show that Random Forest has been effective in crop recommendation system because it can handle various agricultural factors like the nutrients in soil and rainfall. Classification algorithms such as Support Vector Machines are also used in crop yield prediction. Under these models, there is more precision when it comes to predicting the right crops under varying environmental conditions. Supervised learning approaches offer classical remedies in crop classification and estimation of crop yields.

Unsupervised Learning in Agricultural Data Analysis

Hidden patterns in agronomical data are learnt with the help of the unsupervised learning methods. These methods do not need labelled data as opposed to the supervised ones. One of the most frequently applicable algorithms to the analysis of agricultural data is K-Means clustering. It clusters alike pieces of data similarity, such as the amount of rainfall or the concentration of nutrients in soil.

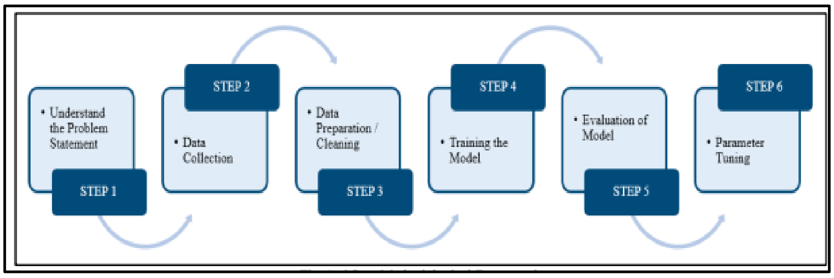
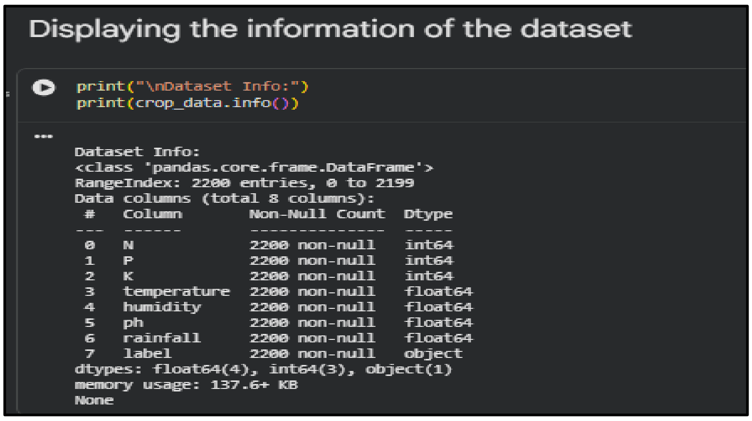


Fig 2: Machine Learning Model Development Workflow for Crop Prediction

Agglomerative Clustering, as well as Hierarchical clustering, is implemented to discover structured clusters of the farming data (Panigrahi *et al.* 2023). The approach uses clusters in a sequence and assists in learning about what environmental variables are related to one another. The purpose of clustering is to be able to segment agricultural fields, determine the types of soils, and determine crop patterns.

Methodology

Dataset Description



Displaying the information of the dataset

```
print("\nDataset Info:")
print(crop_data.info())
```

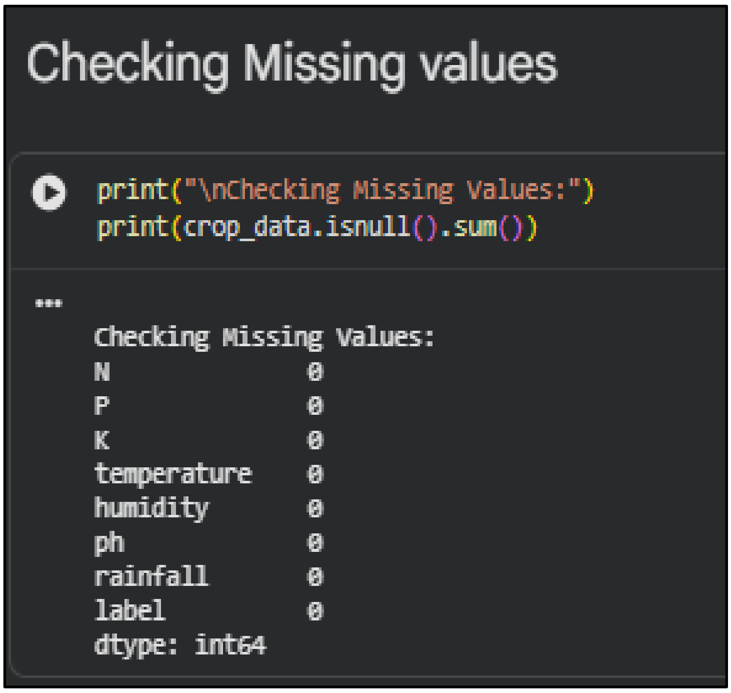
```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2200 entries, 0 to 2199
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   N            2200 non-null   int64  
 1   P            2200 non-null   int64  
 2   K            2200 non-null   int64  
 3   temperature  2200 non-null   float64
 4   humidity     2200 non-null   float64
 5   ph           2200 non-null   float64
 6   rainfall     2200 non-null   float64
 7   label        2200 non-null   object 
dtypes: float64(4), int64(3), object(1)
memory usage: 137.6+ KB
None
```

Fig 3: Displaying the information

The data employed in this research was collected on Kaggle and comprises agricultural data for crop suggestion. It has 2,200 records and 8 columns, which are “nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, rainfall, and crop label”. The data has numerical and categorical variables. Checking data reveals that 2,200 non -elements exist in all columns, meaning that there are no missing numbers. This guarantees high-quality analysis and model construction.

Dataset Link: “https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset/data”

Data Preprocessing



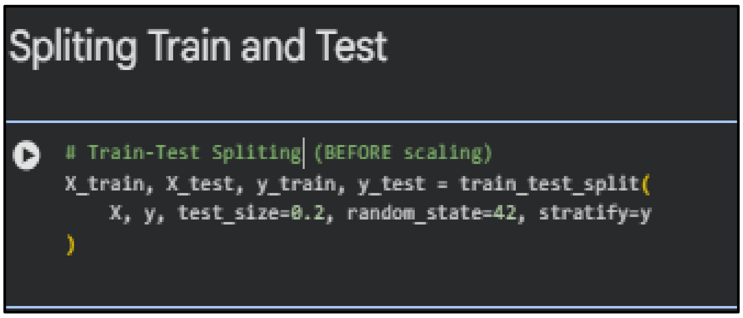
Checking Missing values

```
print("\nChecking Missing Values:")
print(crop_data.isnull().sum())
```

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

Fig 4: Handling Missing Values

As depicted in the figure, zero missing values are found in all the variables, such as N, P, K, temperature, humidity, pH, rainfall, and label. This confirms the completeness of databanks. The figure and the data in this paper guarantee the reliability of the data and aid proper training of the model without the necessity of imputation and elimination of the data.

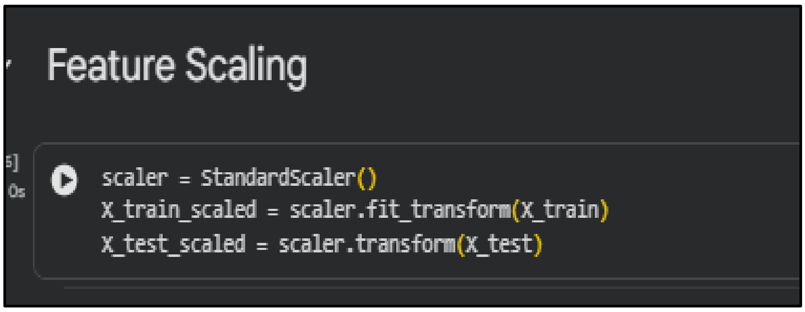


Spliting Train and Test

```
# Train-Test Spliting (BEFORE scaling)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Fig 5: Splitting data into train and test

The figure shows how the dataset is split into a training and a testing set and on an 80:20 proportion. This guarantees objective model assessment. In this paper, data splitting eliminates overfitting and enables appropriate validation of crop prediction performance.



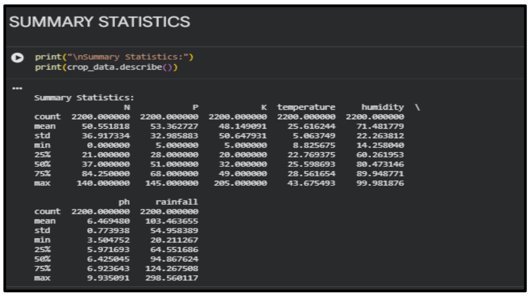
Feature Scaling

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Fig 6: Performing Feature scaling

The figure demonstrates the use of standardization on training and testing data, Standard Scaler. Scaling brings down agricultural variables, rainfall and temperature. Feature scaling gives stability to the model and increases the performance of both supervised and unsupervised learning methods.

Performing EDA



SUMMARY STATISTICS

```
print("\nSummary Statistics:")
print(crop_data.describe())
```

```
Summary Statistics:
                 N            P            K  temperature     humidity  \
count  2200.000000  2200.000000  2200.000000  2200.000000  2200.000000
mean     50.551818    53.362727    48.149091    25.616244    71.481779
std      36.917334    32.985883    50.647931     5.063749    22.263812
min       0.000000     5.000000     5.000000     8.825675    14.258040
25%      21.000000    28.000000    20.000000    22.769375    60.261953
50%      37.000000    51.000000    32.000000    25.598693    80.473146
75%      84.250000    68.000000    49.000000    28.561654    89.948771
max     140.000000   145.000000   205.000000    43.675493    99.981876

                ph     rainfall
count  2200.000000  2200.000000
mean      6.469480   103.463655
std       0.773938    54.958389
min       3.504752    20.211267
25%       5.971693    64.551686
50%       6.425045    94.867624
75%       6.923643   124.267508
max       9.935091   298.560117
```

Fig 7: Displaying summary statistics of the dataset

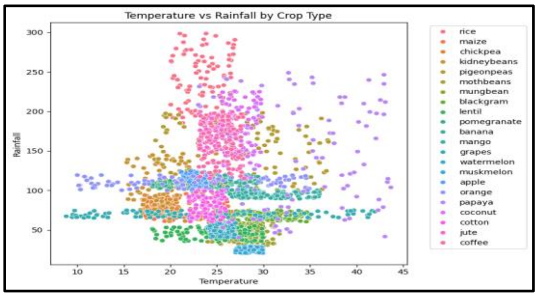


Fig 8: Scatter plot of temperature and rainfall by crop type

The data has 2,200 observations, which include the important agricultural variables, including nitrogen, phosphorus, potassium, rainfall, and temperature (Durai and Shamili, 2022). The levels of nutrients and climatic conditions are widely varied, with the temperature of 10 ^{0}C to 45 ^{0}C and the rainfall of 20 mm and 300 mm, respectively to favor accurate crop prediction modelling.

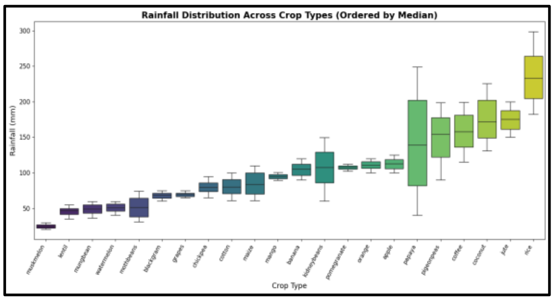


Fig 9: Boxplot of rainfall distribution

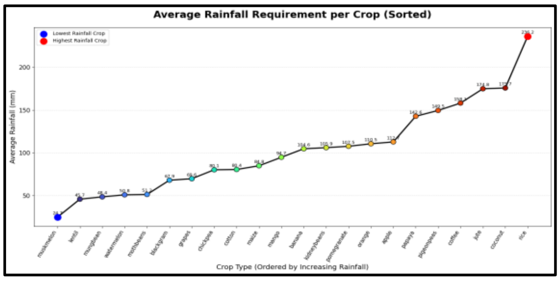


Fig 10: Line plot of average rainfall requirement per crop

The rainfall varies between almost 20mm and more than 260 mm across the types of crops. Mothbean and lentil have low median rainfall (between 25 and 50 mm), but rice and coconut have high median values of above 200 mm. The sorted average rainfall plot indicates that rice has the highest value of about 215 mm. Such insights are applied to crop suitability analysis and also reinforce precision farming decisions.

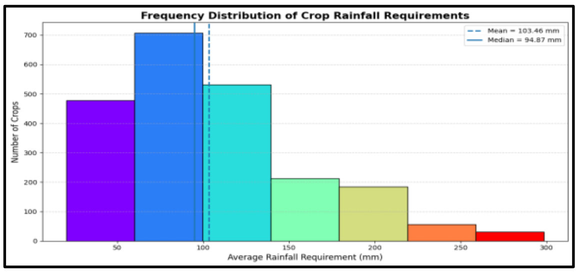


Fig 11: Frequency distribution of crop rainfall

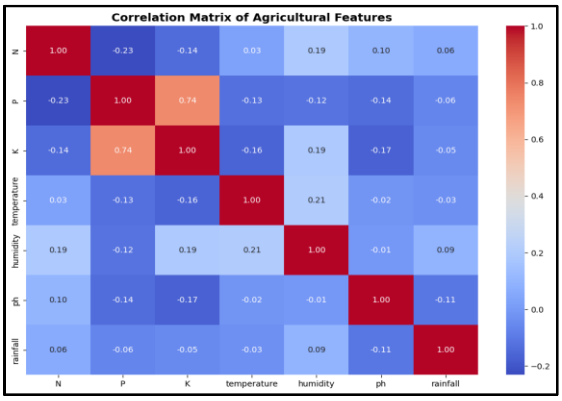


Fig 12: Correlational analysis

Frequency of rainfall reveals that most crops need 50 mm to 150 mm of rainfall, with a mean of 103.46 mm and a median of 94.87 mm, thus indicating average water demands. Correlation analysis indicates that phosphorus and potassium have strong positive relations (0.74), yet other variables have a weak correlation. The findings in these domains influence the feature understanding and aid in the sound classification and cluster analysis of crops.

Supervised Learning Models

Supervised learning methods are used to categorize appropriate crops based on the nutrients of soils and the environmental conditions that are present in the crop recommendation dataset of Kaggle. The dataset consists of nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall as predictive variables, and crop label is the target variable (Getahun *et al.* 2024). The model estimates class probabilities using the sigmoid function:

- “$P\ (y = 1|x) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_n\ x_n\ )})$” ……….(1)

The model parameters are trained through maximizing the likelihood of the training data. Furthermore, a Random Forest classifier is utilized to enhance the predictive performance with an assembly of decision trees produced with the help of bootstrapped samples. Accuracy, precision, recall, F1-score, and confusion matrices are used to assess model performance.

3.4 Unsupervised Learning Models

Discovering hidden patterns in the agricultural conditions without crop labels can be done using unsupervised learning methods. During distance-based clustering, feature variables are standardized so that their contribution is the same. K-Means clustering algorithm separates the observations into K clusters by minimizing the within-cluster sum of squares:

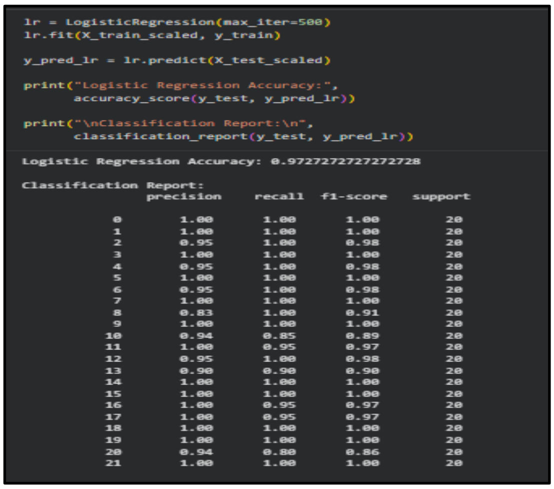
- “$J = \Sigma_{i=1}^{K}\ \Sigma_x \in C_i\ ||x - \mu_i||^2$” ……………(2)

where C_i represents cluster members, and μ_i denotes the cluster centroid.

The number of generated clusters is determined by using the algorithm of Elbow, which yields K=4. The Silhouette Score is used to compute cluster quality. Principal Component Analysis (PCA) creates two principal components of the multidimensional variables of agriculture to assist in visualizing agricultural data, without losing key variance (Shams *et al.* 2024). A comparison of the outcomes of the clustering results will give a better understanding of the similarities in the environment which determine the suitability of crops in the precision farming systems.

Results and Analysis

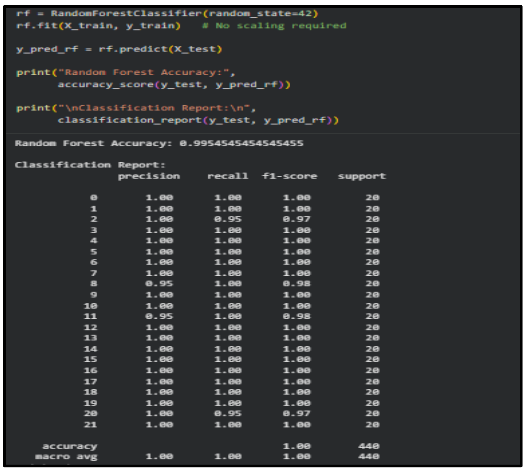
4.1 Supervised Model Results



```
lr = LogisticRegression(max_iter=500)
lr.fit(X_train_scaled, y_train)

y_pred_lr = lr.predict(X_test_scaled)

print("Logistic Regression Accuracy:",
      accuracy_score(y_test, y_pred_lr))

print("\nClassification Report:\n",
      classification_report(y_test, y_pred_lr))
```

```
Logistic Regression Accuracy: 0.9727272727272728

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        20
           1       1.00      1.00      1.00        20
           2       0.95      1.00      0.98        20
           3       1.00      1.00      1.00        20
           4       0.95      1.00      0.98        20
           5       1.00      1.00      1.00        20
           6       0.95      1.00      0.98        20
           7       1.00      1.00      1.00        20
           8       0.83      1.00      0.91        20
           9       1.00      1.00      1.00        20
          10       0.94      0.85      0.89        20
          11       1.00      0.95      0.97        20
          12       0.95      1.00      0.98        20
          13       0.90      0.90      0.90        20
          14       1.00      1.00      1.00        20
          15       1.00      1.00      1.00        20
          16       1.00      0.95      0.97        20
          17       1.00      0.95      0.97        20
          18       1.00      1.00      1.00        20
          19       1.00      1.00      1.00        20
          20       0.94      0.80      0.86        20
          21       1.00      1.00      1.00        20
```

Fig 13: Performing Logistic Regression



```
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)    # No scaling required

y_pred_rf = rf.predict(X_test)

print("Random Forest Accuracy:",
      accuracy_score(y_test, y_pred_rf))

print("\nClassification Report:\n",
      classification_report(y_test, y_pred_rf))
```

```
Random Forest Accuracy: 0.9954545454545455

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        20
           1       1.00      1.00      1.00        20
           2       1.00      0.95      0.97        20
           3       1.00      1.00      1.00        20
           4       1.00      1.00      1.00        20
           5       1.00      1.00      1.00        20
           6       1.00      1.00      1.00        20
           7       1.00      1.00      1.00        20
           8       0.95      1.00      0.98        20
           9       1.00      1.00      1.00        20
          10       1.00      1.00      1.00        20
          11       0.95      1.00      0.98        20
          12       1.00      1.00      1.00        20
          13       1.00      1.00      1.00        20
          14       1.00      1.00      1.00        20
          15       1.00      1.00      1.00        20
          16       1.00      1.00      1.00        20
          17       1.00      1.00      1.00        20
          18       1.00      1.00      1.00        20
          19       1.00      1.00      1.00        20
          20       1.00      0.95      0.97        20
          21       1.00      1.00      1.00        20

    accuracy                           1.00       440
   macro avg       1.00      1.00      1.00       440
```

Fig 14: Performing Logistic Regression and Performing Random Forest

The accuracy of the Logistic Regression of the 22 crop classes is 0.9727 with most of the values of precision and recalls significantly close to 1.00. Random Forest has a higher accuracy of 0.9955 with good and similar precision, recall and F1-score. These findings assess the supervised learning performance and prove that Random Forest is a more stable model to use in the correct crop prediction in precision farming.

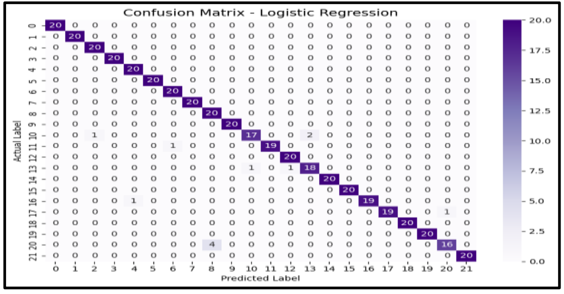


Fig 15: Confusion matrix for Logistic regression

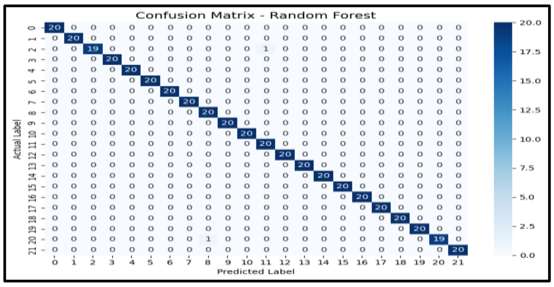


Fig 16: Confusion matrix for Random Forest

The strong values of the diagonal of most of the 22 crop classes at the confusion matrices denote correct classifications. The misclassifications of Logistic Regression are few of which whereas the Random Forest has practically perfect diagonal consistency. These findings confirm the reliability of classification and validate the fact that Random Forest is a more reliable model in the crop recommendation process under precision farming applications.

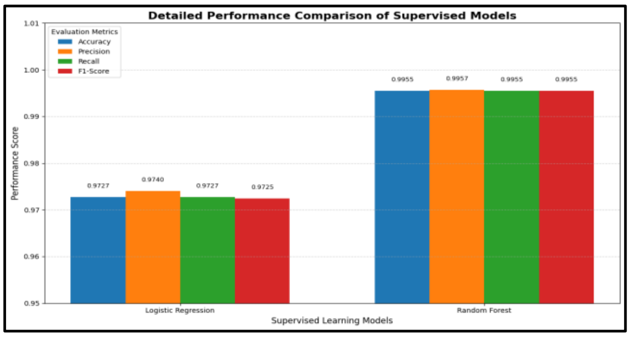
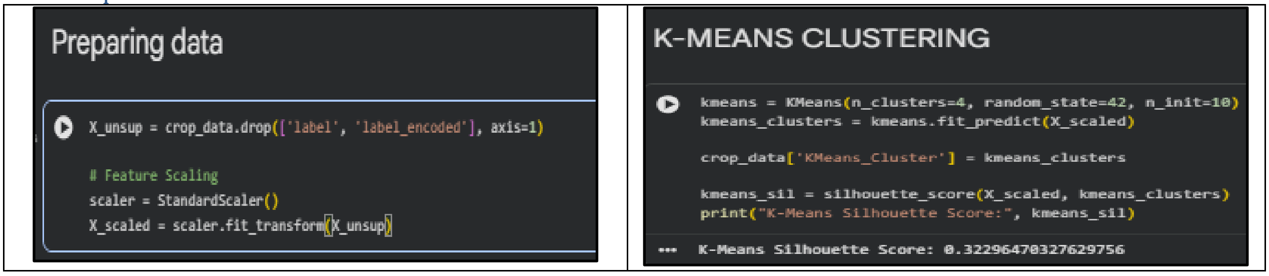


Fig 17: Performance comparison of Supervised Models

Comparative analysis indicates that Logistic Regression has a higher accuracy of 0.9727 and F1-score of 0.9725, whilst the Random Forest has a better result of 0.9955 in terms of accuracy, precision, recall, and F1-score. Random Forest is more appropriate because it is more consistent. This analysis is critical in the selection of the model used and the validation of the most trustworthy supervised model that is used to predict crops precisely in precision farming.

4.2 Unsupervised Model Results



Preparing data

```
X_unsup = crop_data.drop(['label', 'label_encoded'], axis=1)

# Feature Scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_unsup)
```

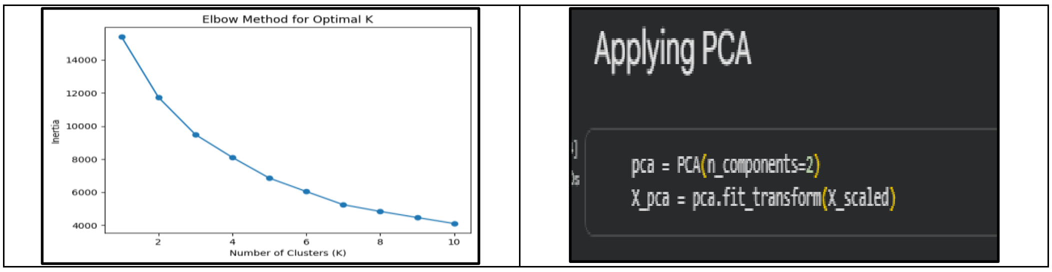
K-MEANS CLUSTERING

```
kmeans = KMeans(n_clusters=4, random_state=42, n_init=10)
kmeans_clusters = kmeans.fit_predict(X_scaled)

crop_data['KMeans_Cluster'] = kmeans_clusters

kmeans_sil = silhouette_score(X_scaled, kmeans_clusters)
print("K-Means Silhouette Score:", kmeans_sil)
```

```
K-Means Silhouette Score: 0.32296470327629756
```

Fig 18: Preparing data and Performing K-means Clustering

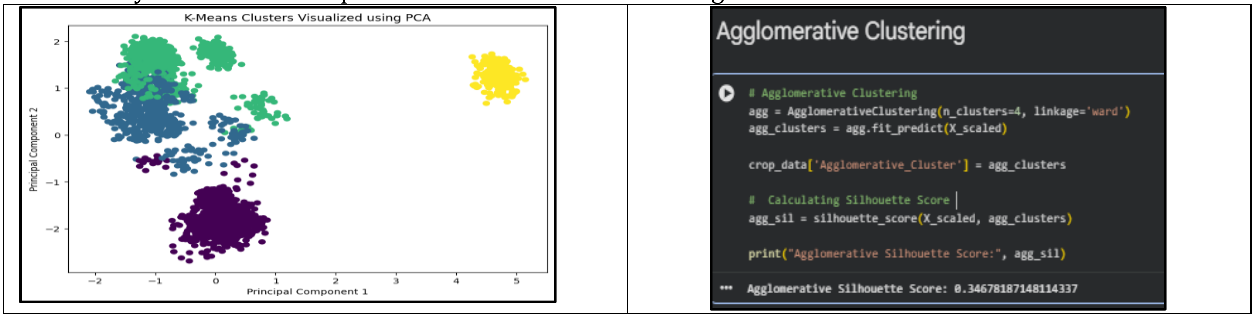
The process of data preparation indicates that the label variables are removed, and StandardScaler is used to normalize the agricultural features, followed by clustering. Proper scaling is necessary to make distance-based algorithms reliable. The application of K-Means with k=4 generates a silhouette score of 0.32296470327629756, meaning moderate cluster segregation. These steps build the analytical basis of the research as an organization of inputs and the confirmation of meaningful discovery patterns to validate the use of pattern discovery in applications of precision farming insights.



```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

Fig 19: Performing the Elbow method for finding optimal K and Applying PCA in k-means clustering

Elbow analysis indicates that inertia sharply decreases between K=1 and 8,000 at K=4 with a smooth increase thereafter, and four clusters is the right option to make. PCA set to n_components = 2 is used to reduce dimensionality to form visual representations of the clusters of agricultural data.



```
# Agglomerative Clustering
agg = AgglomerativeClustering(n_clusters=4, linkage='ward')
agg_clusters = agg.fit_predict(X_scaled)

crop_data['Agglomerative_Cluster'] = agg_clusters

#  Calculating Silhouette Score
agg_sil = silhouette_score(X_scaled, agg_clusters)

print("Agglomerative Silhouette Score:", agg_sil)

Agglomerative Silhouette Score: 0.34678187148114337
```

Fig 20: Visualization of k-means clustering using PCA and Performing Agglomerative clustering

K-Means classification of data based on visualization by PCA identifies four distinct and distinct clusters along Principal Components 1 and 2, suggesting an orderly pattern of agricultural activities. Agglomerative Clustering, nclusters= 4 and ward linkage gives a silhouette score of 0.34678187148114337, which is marginally better than K-Means (0.3229), implying that clustering has better cohesion and separation.

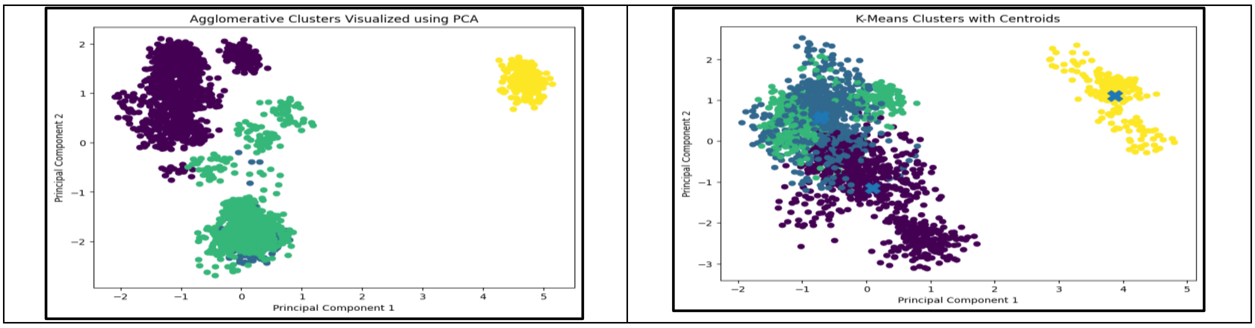
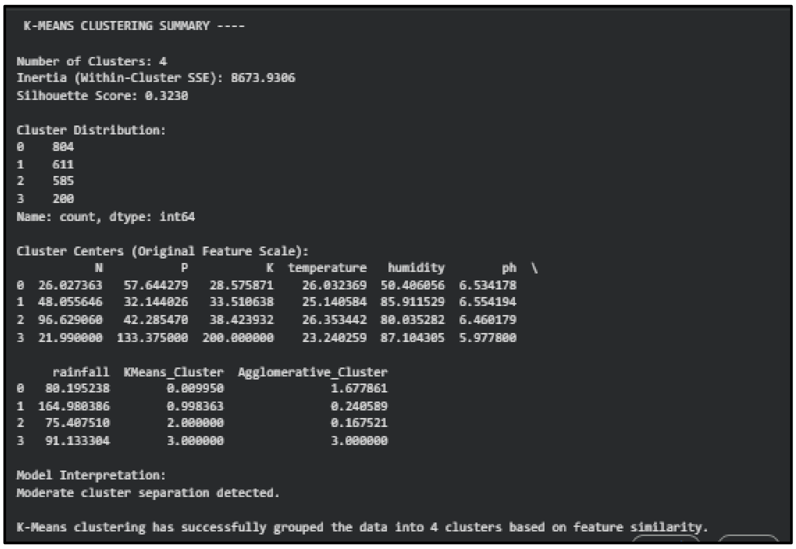


Fig 21: Agglomerative clustering using PCA and Implementing centroid

The agricultural conditions are highly hierarchical, with four distinct groups clearly defined in agglomerative clustering as visualized using Principal Component 1 and Principal Component 2. K-Means centroid clustering to define centers on clusters also translates cluster centers with clear highlights, which explains group compactness and separation. These visualizations are very important in validating pattern discovery and reinforcing unsupervised learning evaluation in the study.



```
 K-MEANS CLUSTERING SUMMARY ----

Number of Clusters: 4
Inertia (Within-Cluster SSE): 8673.9306
Silhouette Score: 0.3230

Cluster Distribution:
0    804
1    611
2    585
3    200
Name: count, dtype: int64

Cluster Centers (Original Feature Scale):
           N           P           K  temperature   humidity        ph  \
0  26.027363   57.644279   28.575871    26.032369  50.406056  6.534178
1  48.055646   32.144026   33.510638    25.140584  85.911529  6.554194
2  96.629060   42.285470   38.423932    26.353442  80.035282  6.460179
3  21.990000  133.375000  200.000000    23.240259  87.104305  5.977800

     rainfall  KMeans_Cluster  Agglomerative_Cluster
0   80.195238        0.009950               1.677861
1  164.980386        0.998363               0.240589
2   75.407510        2.000000               0.167521
3   91.133304        3.000000               3.000000

Model Interpretation:
Moderate cluster separation detected.

K-Means clustering has successfully grouped the data into 4 clusters based on feature similarity.
```

Fig 22: Summary of the results of unsupervised models

The K-Means clustering results have revealed 4 clusters with inertia of 8673.9306 and silhouette score of 0.3230, which means a moderate separation. There are 804, 611, 585, and 200 observations in cluster distribution. This is an output that backs up unmonitored validation, plus pattern-based agricultural choice examination in the study.

Pest Detection Flowchart

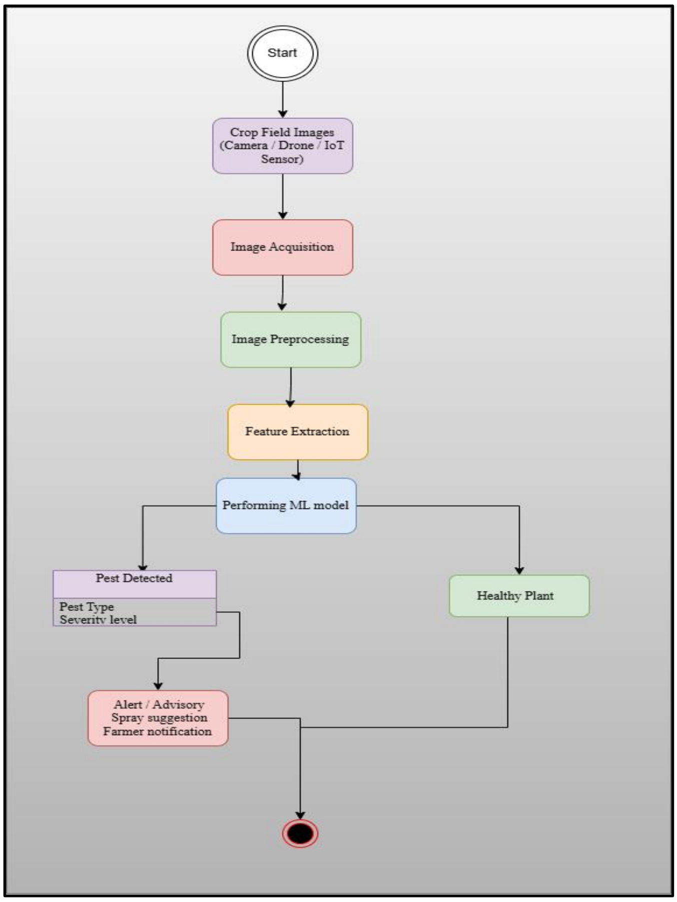


Fig 23: Flowchart of Pest Detection

Model Comparison

Performance comparison assesses the presence of supervised and unsupervised methods of learning used in the crop recommendation dataset provided by Kaggle. The Logistic Regression has an accuracy of 0.9727 and it exhibits good classification among several classes of crops. Random Forest shows increased predictive power in aspect accuracy of 0.9955, which means it can deal with more nonlinear relationships between the variables of soil and climatic conditions (Sharma et al. 2022). Unsupervised analysis gives K-Means clustering a silhouette score of 0.3229 and Agglomerative Clustering a score of 0.3468, indicating a slightly higher separation of clusters. The overall findings suggest that random Forest and Agglomerative procedures provide the best results when it comes to analyzing the agricultural pattern and recommending crops.

Discussion

Agricultural features analysis shows that climatic variables and soil nutrients significantly contribute to the fitness of crops. Supervised models indicate great predictive potential, the top predictive performance of 0.9955 objectively obtained with the implementation of the Random Forest model, which is a predictor of complex interactions among features commercially capable. The accuracy of the Logistic Regression at 0.9727 is also reliable. The further clustering analysis shows significant groupings of the environment, with Agglomerative Clustering yielding a silhouette score of 0.3468, a little more than K-Means (0.3229). The results emphasize the importance of using predictive modelling and pattern discovery data to make better choices in precision farming.

Conclusion

Machine learning offers a useful method of analyzing agricultural conditions and assisting in crop recommendations. The crop dataset analysis of the Kaggle reveals that soil nutrients, rainfall, temperature, humidity, and pH are important factors that affect the suitability of crops. Supervised learning models classify the types of crops successfully with the best accuracy of 0.9955 for Random Forest and 0.9727 for Logistic Regression. Unsupervised methods indicate significant environmental groupings in which Agglomerative Clustering has a score of 0.3468, a little greater than K-Means (0.3229). Comprehensively, the combination of predictive modelling and clustering enables the crafting of decisions that rely on data and enhance precision farming.

References

1. Amini, M. and Rahmani, A., 2023. Agricultural databases evaluation with machine learning procedure. *Australian Journal of Engineering and Applied Science*, *8*(2023), pp.39-50.
2. Panigrahi, B., Kathala, K.C.R. and Sujatha, M., 2023. A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Procedia Computer Science*, *218*, pp.2684-2693.
3. Durai, S.K.S. and Shamili, M.D., 2022. Smart farming uses machine learning and deep learning techniques. *Decision Analytics Journal*, *3*, p.100041.
4. Getahun, S., Kefale, H. and Gelaye, Y., 2024. Application of precision agriculture technologies for sustainable crop production and environmental sustainability: A systematic review. *The Scientific World Journal*, *2024*(1), p.2126734.
5. Shams, M.Y., Gamel, S.A. and Talaat, F.M., 2024. Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Computing and Applications*, *36*(11), pp.5695-5714.
6. Sharma, K., Sharma, C., Sharma, S. and Asenso, E., 2022. Broadening the research pathways in smart agriculture: predictive analysis using semiautomatic information modeling. *Journal of Sensors*, *2022*(1), p.5442865.