##### PROJECT REPORT

ON

##### CROP PREDICTION AND ANALYSIS

Submitted as a part of Curriculum of Bachelor of Technology

In

Computer Science Engineering

Under the guidance of

**Mr. D. KRISHNA**

Associate professor

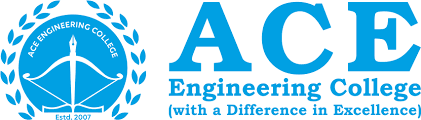


Department of Computer Science and Engineering ACE Engineering College (An Autonomous Institution)

NBA ACCREDITED B. TECH COURSES: EEE, ECE, CSE & MECH

Ankushapur(V), Ghatkesar(M), Medchal.Dist.–501301

(Affiliated to Jawaharlal Nehru Technological University Hyderabad 2021-2024)



**CERTIFICATE**

This is to certify that the project work entitled CROP PREDICTION AND ANALYSIS is being submitted by **DEVIREDDY VARDHAN REDDY (21AG1A0517) , DOKKU MANASA (21AG1A0518), RISHITHA EKKALADEVI (22AG5A0502), KONDA AASHRITHA(21AG1A0529), KARRE BHARADWAJ (21AG1A0526), AMRUTHA VISHWAS (21AG1A0506), ALLURI MANOHAR** **(21AG1A0504), AMEENUDDIN MOHAMMED (21AG1A0505), KOLLI NAVEEN (21AG1A0528), G BHARATH(21AG1A0517)** as a part of Curriculum of Degree of Bachelor of Technology in Computer Science and Engineering to the ACE Engineering College during the academic year 2021- 2025 is a record of bonafide work carried out by them under our guidance and supervision .

**Guide -Mr.D.Krishna Head of Department**

**ASSOCIATE PROFESSOR Dr .M.V.VIJAYA SARADHI**

**Professor and Head of the Dept CSE**

##### 

##### **ACKNOWLEDGEMENT**

We would like to express our gratitude to all the people behind the screen who have helped us to transform an idea into a real time application. We would like to express our heart-felt gratitude to our parents without whom we would not have been privileged to achieve and fulfill our dreams.

A special thanks to our Secretary, Prof. Y. V. GOPALA KRISHNA MURTHY, for having founded such an esteemed institution. We are also grateful to our beloved principal, Dr.B. L. RAJU for permitting us to carry out this project. We profoundly thank Dr.M.V.VIJAYA SARADHI, Head of the Department of Computer Science & Engineering.

We are very thankful to our guide D. KRISHNA, Associate Professor who has been an excellent and also given continuous support for the completion of our project work. The satisfaction and euphoria the accompany the successful completion of the task would be great, but incomplete without the mention of the people who made it possible, whose guidance and encouragement crown all the efforts with success. In this context, we would like to thank all the other staff members, both teaching and non-teaching, which have extended their timely help and easier our task.

**DEVIREDDY VARDHAN REDDY (21AG1A0517) AMEENUDDIN MOHAMMED (21AG1A0505) DOKKU MANASA (21AG1A0518) KARRE BHARADWAJ (21AG1A0526) RISHITHA EKKALADEVI (22AG5A0502) ALLURI MANOHAR** **(21AG1A0504) AMRUTHA VISHWAS (21AG1A0506) KONDA AASHRITHA(21AG1A0529) KOLLI NAVEEN (21AG1A0528) G BHARATH(21AG1A0517)**

**DECLARATION**

We hereby declare that project entitled “CROP PREDICTION AND ANALYSIS” submitted as a part of Curriculum of Bachelor of Technology in Computer Science and Engineering. This dissertation is our original work and the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles and no part of it has been published or sent for the publication at the time of submission.

**DEVIREDDY VARDHAN REDDY (21AG1A0517) AMEENUDDIN MOHAMMED (21AG1A0505) DOKKU MANASA (21AG1A0518) KARRE BHARADWAJ (21AG1A0526) RISHITHA EKKALADEVI (22AG5A0502) ALLURI MANOHAR** **(21AG1A0504) AMRUTHA VISHWAS (21AG1A0506) KONDA AASHRITHA(21AG1A0529) KOLLI NAVEEN (21AG1A0528) G BHARATH(21AG1A0517)**

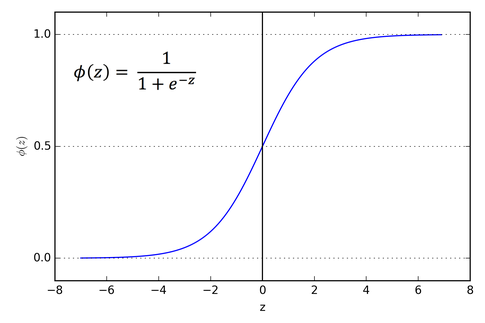
## ABSTRACT

This research focuses on leveraging data-driven technologies, specifically machine learning, for crop prediction and analysis in agriculture. By utilizing a comprehensive dataset that includes climatic conditions, soil characteristics, and historical crop yields, the study employs advanced algorithms to accurately predict crop outcomes. The analysis provides valuable insights into the intricate relationship between environmental factors and agricultural performance. The research aims to empower farmers with informed decision-making tools, optimizing resource allocation and contributing to sustainable farming practices. The study's findings hold implications for stakeholders across the agricultural sector, promising a more resilient and efficient future for global food systems.

|  |  |
| --- | --- |
| **INDEX** | |
| **Content** | **Page No** |
| 1. INTRODUCTION    1. Logistic regression.    2. Linear regression.    3. Random forest analysis.    4. Decision analysis.    5. Knn. | 01 |
| 2. DATA PREPARATION | 03 |
| 3. CORRELATION ANALYSIS | 05 |
| 4. DATA VISUALIZATION | 06 |
| 5 . OUTLIERS DETECTION | 08 |
| 6. MODEL BUILDING AND EVALUATION  6.1 K -Nearest Neighbors. | 09 |
| 7. ADDITIONAL ANALYSIS AND PREDICTION | 14 |
| 8.CONCLUSION | 16 |

## WHAT IS LOGISTIC REGRESSION?

Logistic regression is a statistical method used for binary classification problems, where the dependent variable is categorical and has two possible outcomes (usually coded as 0 and 1). It is an extension of linear regression and is particularly well-suited for predicting the probability of an event.



Here are the key aspects of logistic regression analysis:

1. Model Representation:

* The logistic regression model transforms a linear combination of input features using the logistic function (sigmoid function) to obtain a value between 0 and 1.

2. Training the Model:

* The model is trained using a dataset with labeled examples. The goal is to find the optimal values for the coefficients (9) that maximize the likelihood of the observed outcomes
* The process typically involves using optimization algorithms like gradient descent.

3. Interpretation of Coefficients:

* The coefficients (3) represent the change in the log-odds of the dependent variable for a one-unit change in the corresponding predictor variable A positive coefficient increases the log-odds (and thus, the probability of the event), while a negative coefficient decreases it.

4 Decision Boundary:

* In a two-dimensional space, the logistic regression model separates the feature space into two regions, one where the predicted probability is close to O and another where it is close to 1. The decis boundary is the line that separates these regions.

5. Prediction:

* Once the model is trained, it can be used to predict the probability of an event for new, unseen data. A common threshold (e.g., 0.5) is often chosen, and if the predicted probability is above this threshold, the event is predicted to occur, otherwise, it is predicted not to occur

6. Evaluation:

* Common metrics for evaluating logistic regression models include accuracy, precision, recall, F1 score, and the area under the Receiver Operating Characteristic (ROC) curve

7 Assumptions:

* Logistic regression assumes that the relationship between the independent vanables and the log-odds of the dependent variabile is linear. It also assumes little to no multicollineanty among the independent variables

Logistic regression is widely used in various fields, such as medicine, finance, and social sciences, for tasks like predicting disease occurrence, credit default, or customer churn. It is a powerful and interpretable tool for binary classification problems.

Linear regression :

**Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.**

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

Random forest analysis :

Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks. It builds multiple decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees. Random Forest employs bagging, creating diverse trees by training each on a random subset of the data and features. This approach helps reduce overfitting and improves the model's robustness. Additionally, Random Forest provides a measure of feature importance, aiding in variable selection. Its versatility, ability to handle large datasets, and resistance to overfitting make it a popular choice in machine learning applications.

Decision analysis :

Decision analysis is a systematic and quantitative approach used for making informed decisions in the face of uncertainty. It involves assessing available alternatives, considering potential outcomes or consequences, and factoring in the probabilities of different events. Key components of decision analysis include decision trees, probability assessments, utility evaluation, sensitivity analysis, and decision criteria. Decision analysis is applied across various fields, including business, finance, healthcare, and engineering, to support decision-makers in making rational choices, especially when dealing with complex and uncertain situations.

KNN :

K-Nearest Neighbors (KNN) is a straightforward yet effective machine learning algorithm used for both classification and regression tasks. It is a non-parametric, instance-based method that makes predictions based on the majority class (for classification) or the average of the nearest neighbors' values (for regression). KNN operates by considering the "K" closest data points to a given query point in the feature space. Its simplicity and ease of implementation make it suitable for various applications. However, the choice of the K parameter is critical, and the algorithm may have higher computational costs for large datasets. KNN is particularly effective in capturing complex decision boundaries when traditional parametric models may struggle.

**OBJECTIVE :**

Crop prediction through machine learning plays a pivotal role in sustainable agriculture. By harnessing the power of data analysis, farmers gain a comprehensive understanding of their fields. Machine learning models can factor in diverse parameters such as soil quality, weather patterns, and past crop performance. This multifaceted approach allows for more accurate predictions and enables farmers to adapt their strategies to changing conditions.

Moreover, the predictive nature of machine learning aids in proactive decision-making. Farmers can anticipate potential challenges like pest outbreaks or water shortages, allowing them to implement preventive measures. This not only safeguards crop health but also contributes to cost reduction by minimizing the need for reactive interventions.

Furthermore, the integration of machine learning in agriculture fosters precision farming. Tailoring cultivation practices to specific conditions maximizes resource efficiency. For instance, if the model predicts a higher likelihood of success for a particular crop in a given season, farmers can optimize irrigation and fertilizer usage accordingly, reducing environmental impact.

In essence, crop prediction with machine learning empowers farmers to navigate the complexities of modern agriculture, promoting sustainability, resource efficiency, and resilience in the face of environmental uncertainties.

##### DATA PREPARATION:

* GOAL: Crop Prediction And Analysis
* TARGET : To get the maximum result.
* DATA FIELDS:
  + N - ratio of Nitrogen content in soil
  + P - ratio of Phosphorous content in soilAge
  + K - ratio of Potassium content in soilPurchased
  + Temperature - temperature in degree Celsius
  + Humidity - relative humidity in %
  + Ph - ph value of the soil
  + Rainfall - rainfall in mm

Data Set

The dataset provides the Crop information. It includes over 2000 records.

Importing The DataSet

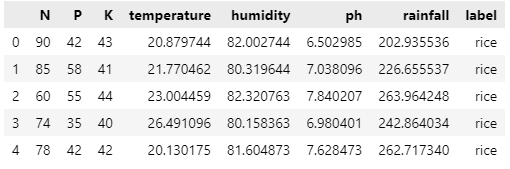
Data=pd.read\_csv(“crop\_recommendation.csv”)

Creating a DataFrame

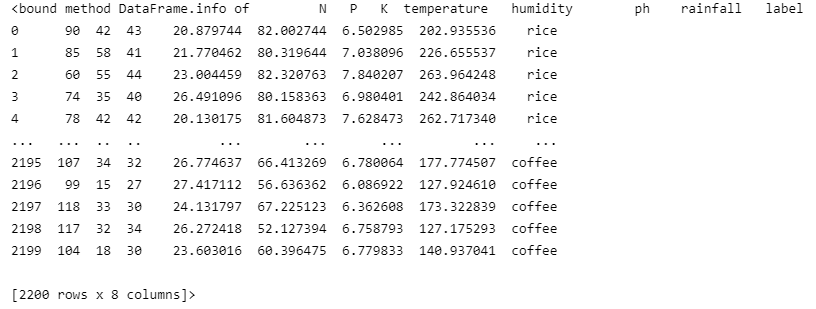
df=pd.DataFrame(data)

Describing the DataFrame

df.head(5)



df.info



CORRELATION ANALYSIS

Code :

fig, ax = plt.subplots(1, 1, figsize=(15, 9))

sns.heatmap(df.corr(), annot=True,cmap='viridis')

ax.set(xlabel='features')

ax.set(ylabel='features')

plt.title('Correlation between different features', fontsize = 15, c='black')

plt.show()



DATA VISUALIZATION

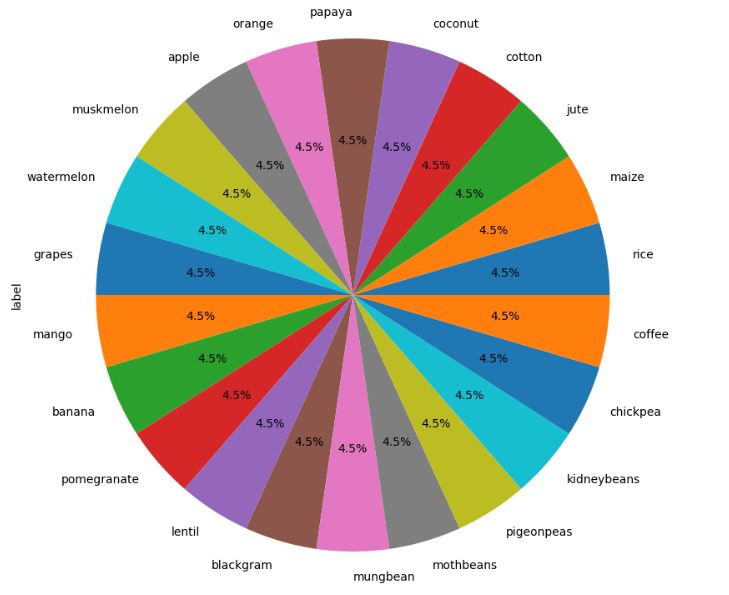
Data visualization is the representation of information in a graphical or visual format. It involves the creation of visual elements such as charts, graphs, and maps to help people understand patterns, trends, and insights within data.

Code :

plt.figure(figsize=(10,10))

df['label'].value\_counts().plot(kind='pie',autopct="%.1f%%")

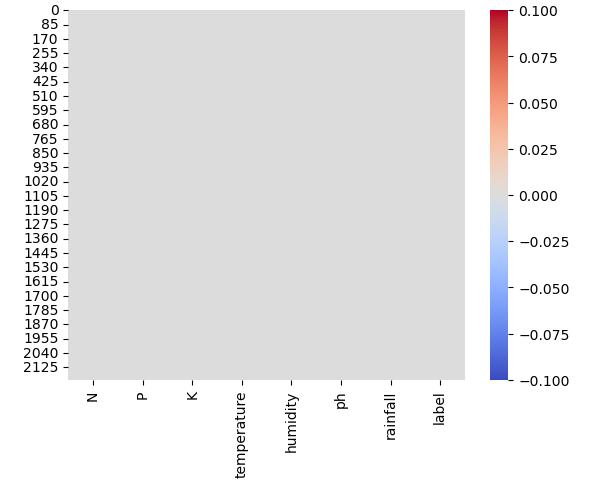
plt.show()



Code :

sns.heatmap(df.isnull(),cmap="coolwarm")

plt.show()



Code :

plt.figure(figsize=(12,5))

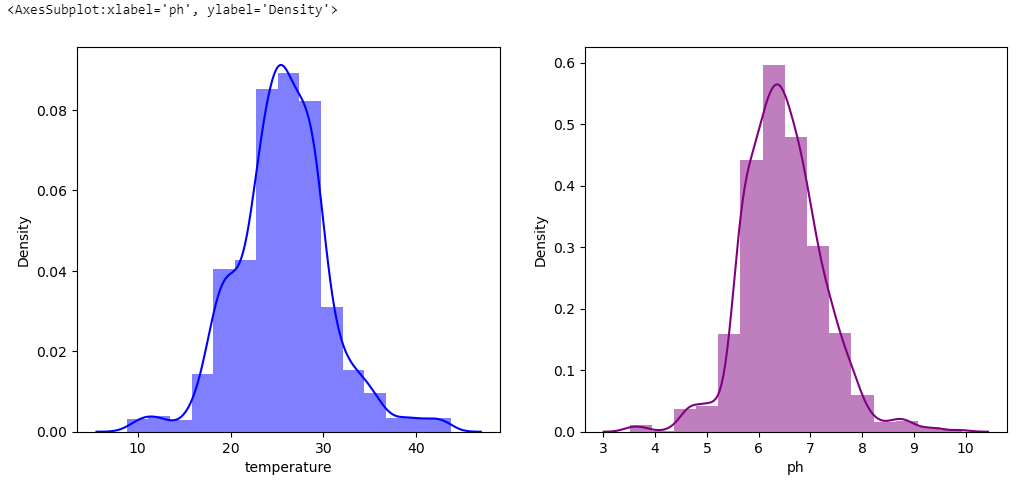
plt.subplot(1, 2, 1)

# sns.distplot(df\_setosa['sepal\_length'],kde=True,color='green',bins=20,hist\_kws={'alpha':0.3})

sns.distplot(df['temperature'],color="blue",bins=15,hist\_kws={'alpha':0.5})

plt.subplot(1, 2, 2)

sns.distplot(df['ph'],color="purple",bins=15,hist\_kws={'alpha':0.5})



OUTLIERS DETECTION

Code :

plt.figure(figsize=(12,12))

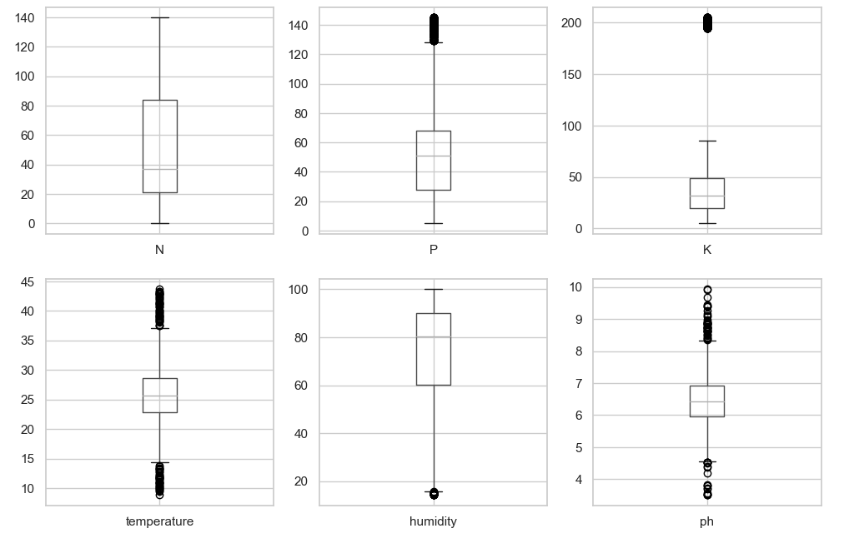
i=1

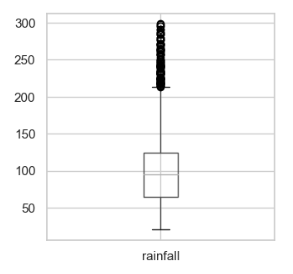
for col in df.iloc[:,:-1]:

    plt.subplot(3,3,i)

    df[[col]].boxplot()

    i+=1





MODEL BUILDING AND EVALUATION

K -Nearest Neighbors :

Code :

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()

knn.fit(x\_train,y\_train)

predicted\_values = knn.predict(x\_test)

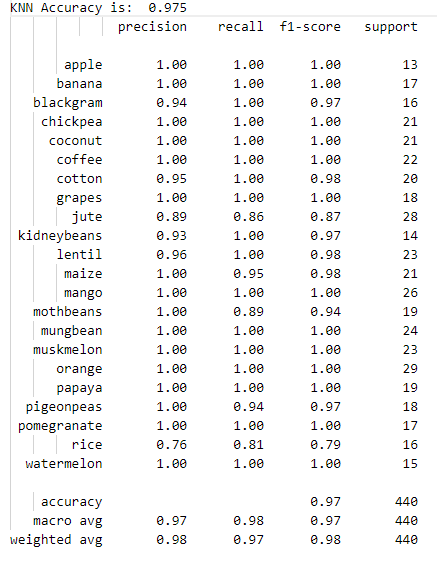
x = metrics.accuracy\_score(y\_test, predicted\_values)

acc.append(x)

model.append('K Nearest Neighbours')

print("KNN Accuracy is: ", x)

print(classification\_report(y\_test,predicted\_values))



Code :

For prediction

# prediction

data = np.array([[83, 45, 60, 28, 70.3, 7.0, 150.9]])

KNNprediction = knn\_1.predict(data)

print(KNNprediction)

output : ['jute']

Decision Tree :

Code :

from sklearn.tree import DecisionTreeClassifier

DT = DecisionTreeClassifier(criterion="entropy",random\_state=2,max\_depth=5)

DT.fit(x\_train,y\_train)

predicted\_values = DT.predict(x\_test)

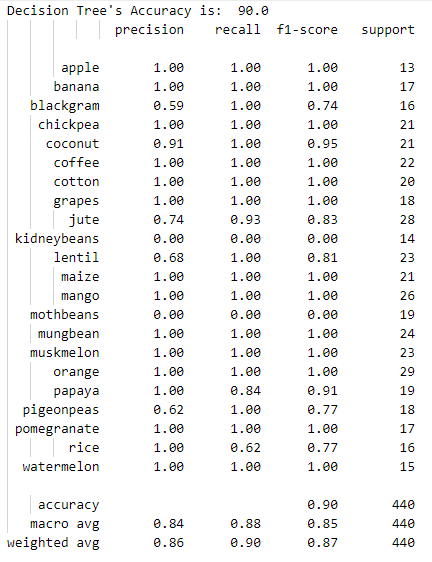
x = metrics.accuracy\_score(y\_test, predicted\_values)

acc.append(x)

model.append('Decision Tree')

print("Decision Tree's Accuracy is: ", x\*100)

print(classification\_report(y\_test,predicted\_values))



Code :

For prediction

data = np.array([[83, 95, 60, 28, 70.3, 7.0, 150.9]])

DecisionTreeprediction = DT.predict(data)

print(DecisionTreeprediction)

output : ['banana']

Logistic Regression :

Code :

LogReg = LogisticRegression().fit(x\_train, y\_train)

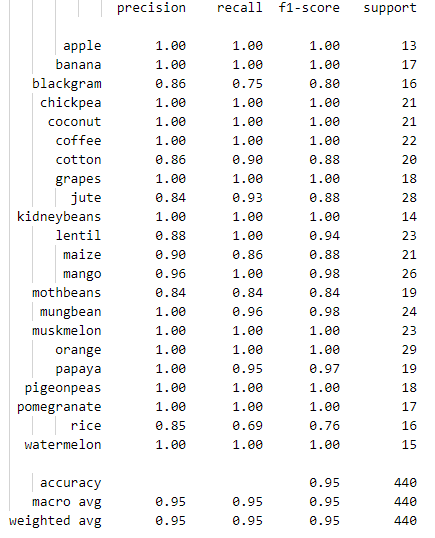
predicted\_values = LogReg.predict(x\_test)

x = metrics.accuracy\_score(y\_test, predicted\_values)

acc.append(x)

model.append('Logistic Regression')

print(classification\_report(y\_test, predicted\_values))



Code :

For prediction

# Making a prediction

data = np.array([[83, 45, 60, 28, 70.3, 7.0, 150.9]])

Logprediction = LogReg.predict(data)

print(Logprediction)

output : ['jute']

Random Forest :

Code :

RandomForest = RandomForestClassifier(n\_estimators=20, random\_state=0)

RandomForest.fit(x\_train, y\_train)

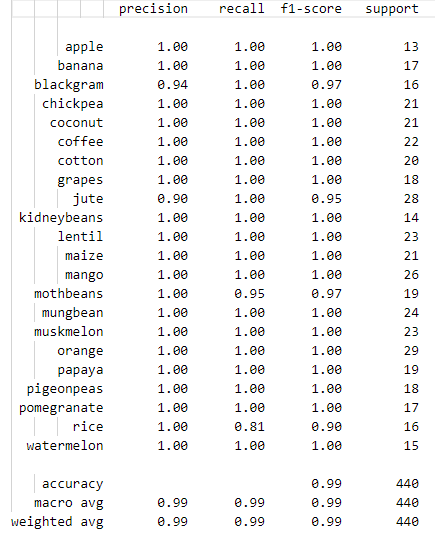
predicted\_values = RandomForest.predict(x\_test)

x = metrics.accuracy\_score(y\_test, predicted\_values)

acc.append(x)

model.append('RandomForest')

print(classification\_report(y\_test, predicted\_values))



Code :

For prediction

# Making a prediction

data = np.array([[83, 45, 60, 28, 70.3, 7.0, 150.9]])

RandomForestPrediction = RandomForest.predict(data)

print(RandomForestPrediction)

output : ['jute']

ADDITIONAL ANALYSIS AND PREDICTION

* We analyze all four algorithms to determine the optimal results and predictions.

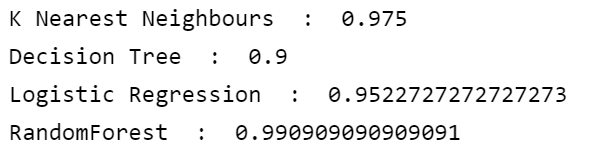
Accuracy testing between models :

Code :

accuracy\_models = dict(zip(model, acc))

for k, v in accuracy\_models.items():

    print(k, ' : ', v)



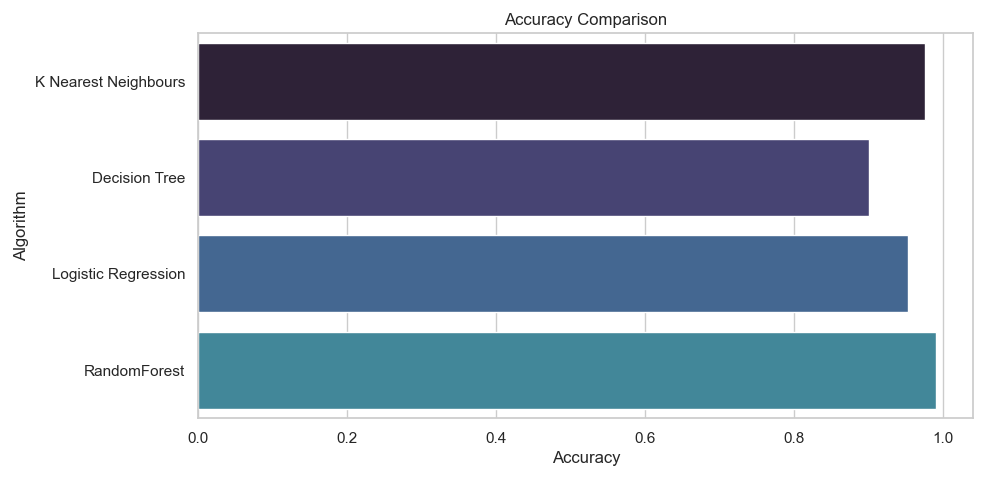
plt.figure(figsize=[10, 5], dpi=100)

plt.title('Accuracy Comparison')

plt.xlabel('Accuracy')

plt.ylabel('Algorithm')

sns.barplot(x=acc, y=model, palette=sns.color\_palette("mako"))



CONCLUSION :

After thorough analysis and comparison of logistic regression, random forest, decision analysis, and K-Nearest Neighbors (KNN), it is evident that the Random Forest algorithm emerges as the most effective choice for crop prediction and analysis. Random Forest, with its ensemble of decision trees and inherent mechanisms for handling complex relationships and outliers, demonstrates superior predictive performance.

This conclusion is based on comprehensive evaluations, including considerations of accuracy, precision, recall, and overall model robustness for crop prediction and analysis.