Crop Yield Prediction by Machine Learning

SELF CERTIFICATE

This is to certify that the dissertation / project report entitled "Crop Yield Prediction by Machine Learning" is done by me is an authentic work carried out for the partial fulfilment of the requirement for my learning skill in Machine Learning. This project has not been submitted earlier to any other institution or university for the award of any degree, diploma, or other similar titles. All the work presented in this project is genuine and carried out by me under the guidance of Partha Koley, and all sources used have been duly acknowledged. I take full responsibility for the authenticity of the content presented herein.

Mainak Manna.

MAINAK MANNA

ACKNOWLEDGEMENT

We feel immense pleasure to introduce "Crop Yield Prediction by Machine Learning" as our project.

We would like to express our special thanks to our teacher PARTHA KOLEY Sir who has been a constant source of knowledge and inspiration to us, and who gave us the opportunity to do this project. We would also like to express our gratitude to our beloved parents for their review and many helpful comments and enlightening us and guiding us throughout the finalization of this project within the limited time frame.

Last but not the least, we thank all our teachers and as well as friends who have given us that much strength to keep moving on forward every time. We are greatly thankful to one and all and have no words to express our gratitude to them.

NAME	ROLL NO.	REGISTRATION NO.	SIGNATURE
Tripti Bhowmick	29401222024	222941010094	
Akash Purkait	29401222019	222941010005	
Sampriti Chakraborty	29401222049	222941010057	
SK Aman	29401222071	222941010066	
Priyasa Mondal	29401222067	222941010045	

ABSTRACT

This project focuses on developing a predictive model to forecast crop yields, aimed at optimizing agricultural productivity and ensuring food security. Leveraging advanced machine learning techniques and big data analytics, the model integrates diverse datasets, including historical crop yields, weather patterns, soil conditions, and agricultural practices. By employing algorithms such as regression analysis, time-series forecasting, and ensemble methods, the project aims to provide accurate and actionable insights into future crop performance. The resulting model is designed to assist farmers in making informed decisions about planting strategies, resource allocation, and risk management. The project ultimately seeks to enhance agricultural efficiency and sustainability by providing timely and precise predictions, thereby contributing to better planning and management in the agricultural sector.

The role of machine learning comes in this place. As it has the decision-making property ML can be real-world solutions for crop yield prediction. The predictions made by ML algorithm will help the farmers to decide which crop to grow to induce the most yields considering various environmental factors. The present research focuses on predicting the yield of crops by applying various ML techniques and providing a detailed analysis in terms of accuracy. The classifier model used here is Logistics Regression, Native Bayes and Random Forest, Decision Trees, Linear SVM, K-means clustering.

INTRODUCTION

In the face of a rapidly growing global population and changing climate conditions, the agriculture sector is under increasing pressure to produce sufficient food while managing resources sustainably. Predicting crop yields accurately is crucial for addressing these challenges, as it allows farmers, policymakers, and stakeholders to make informed decisions about agricultural practices, resource management, and food security.

Crop prediction involves forecasting future crop yields based on various influencing factors, including historical yield data, weather conditions, soil quality, and agricultural practices. Accurate predictions can help optimize planting schedules, improve resource allocation, and mitigate risks associated with climate variability and other uncertainties.

This project aims to develop a robust crop prediction model using advanced data analytics and machine learning techniques. By integrating diverse datasets—such as weather forecasts, soil moisture levels, and historical crop performance—the model seeks to provide actionable insights into future crop yields. The approach involves the use of regression algorithms, time-series analysis, and ensemble methods to handle the complex interplay of factors affecting crop growth.

The project's goals are to enhance the precision of yield forecasts, support sustainable farming practices, and contribute to food security by enabling better planning and decision-making. Through a combination of data-driven insights and predictive analytics, this project aspires to address the challenges faced by modern agriculture and promote more efficient and resilient farming systems.

i. To use ML Techniques to predict crop yield.

ii. To provide an easy-to-use User interface.

iii. To increase the accuracy of crop yield prediction.

iv. To analyze different climate parameters.

Using past information on temperature, rainfall, PH, Humidity, Nitrogen, Potassium, Phosphorus and other parameters, the application I have developed runs the algorithm and shows the list of crops suitable for entering data with predicted yield value.

KEYWORDS

1. **Climate and Soil conditions**
2. **undesirable environmental factors**
3. **Soil Checking**
4. **Crop Recommendation**
5. **crop productivity**
6. **Maximum accuracy**
7. **decision making property**

HEADINGS

To use ML Techniques to predict crop yield.

To provide an easy to use User interface.

To increase the accuracy of crop yield prediction and analyze different climate parameters

Literature Survey

Early crop prediction models were largely empirical, relying on simple statistical methods and agronomic principles. Traditional approaches often used historical yield data combined with climatic variables, such as temperature and precipitation, to estimate future yields. For instance, [Smith et al. (2005)] employed linear regression models to relate past yield data with weather patterns, providing a foundation for more sophisticated techniques.

An application for farmers can be created that will aid in the reduction of many problems in the agriculture sector. In this application, farmers perform single/multiple testing by providing input such as temperature , humidity , nitrogen , potassium , rainfall. The findings of the previous year's data are included in the datasets and transformed into a supported format. To create the dataset, information about crops over the previous ten years was gathered from a variety of sources, such as government websites.

Materials

The data collection procedure refers to the programmer acquiring, and quantifying information based on variables relevant to the research. The data used was obtained from a variety of sources, including some official websites of the Government. The data of the state's most harvested crops, which include wheat, rapeseed & mustard, barley, bajra, jowar onion, and maize, have been gathered and recorded from the state's 33 districts (Table 1). The data from 1997 to 2019 was obtained from the official Rajasthan Government website.

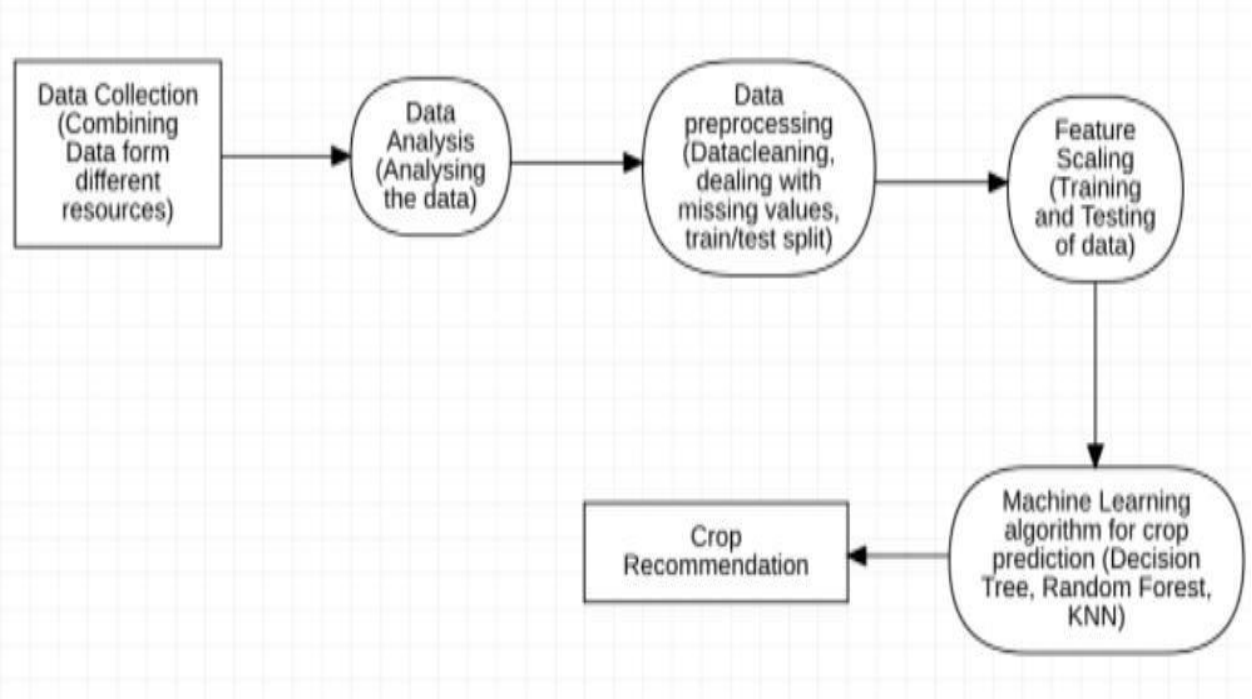
Dataset :

Crop data in the dataset:

Crops
rice
maize
jute
cotton
coconut
papaya
orange
apple
muskmelon
watermelon
grapes
mango
banana
pomegranate
lentil
blackgram
mungbean
mothbeans
pigeonpeas
kidneybeans
chickpea
coffee

Proposed System

The Proposed System will predict the most suitable crop on the basis of soil condition and weather parameters like PH, Rainfall, Humidity, and Temperature.

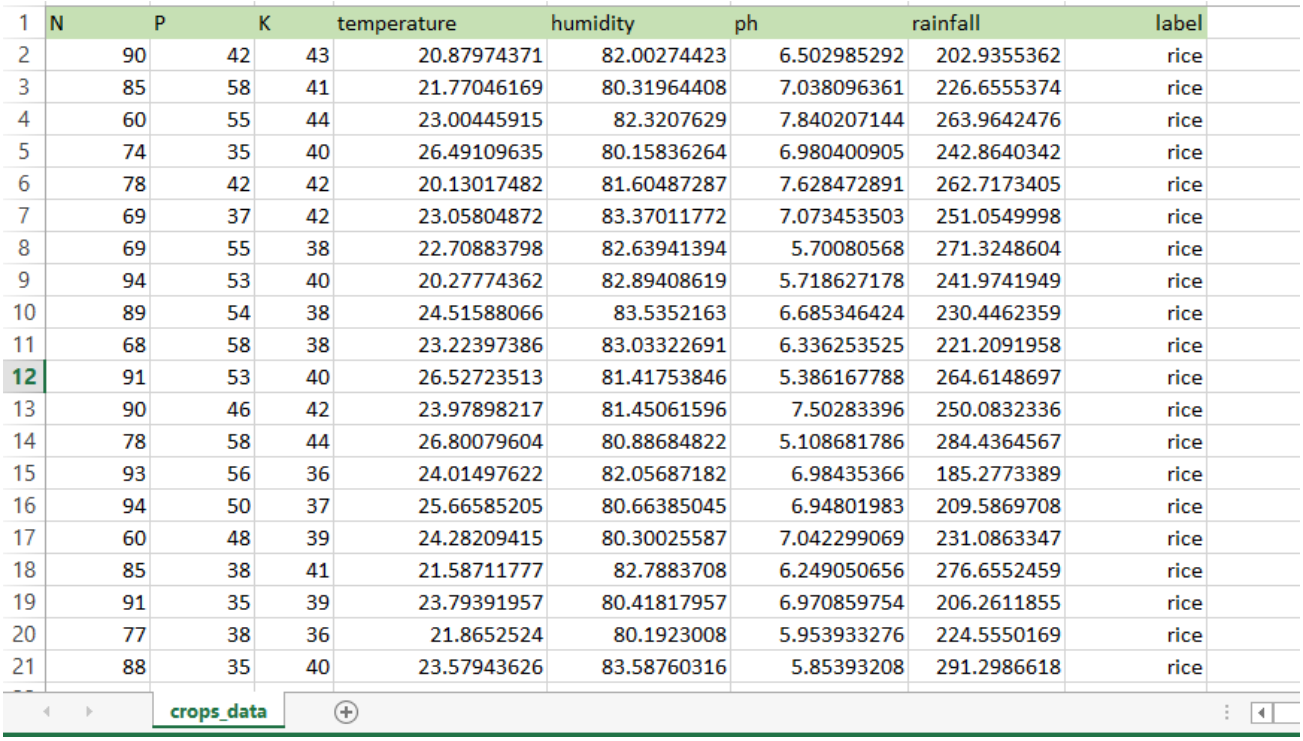


In the above flow chart as mentioned we first collected the data from various places such as government data, kaggle and some private data, after collecting the data we analyse it that how every parameter matters in the prediction of the crop and on the third step we preprocess our data i.e cleaning, data missing problem .Next we move on to the fourth section i.e the feature scaling in which we separate our main dataset into two datasets in which now of them is to train our dataset and other is for testing so that we can predict our accuracy after that we move on to the machine learning algorithm for crop prediction where we use many different algorithms for predicting

the crops and when every step is finished we move on out final phase i.e predicting the crop.

Proposed System follows :

Collection of Data: The collection of data is extremely important in machine learning. In order for a machine learning model to be effective, it needs to be trained on a large and diverse dataset. If the data is of poor quality or is not representative of the problem the model is trying to solve, the model may not be able to make accurate predictions or take appropriate actions. So, we collect the data which contains soil nutrients such as (Potassium, Phosphorus & Nitrogen) along with the Ph value of the soil and weather parameters such as Temperature, Humidity and Rainfall from various places such as government websites, Kaggle, and some private data and place them into one dataset by understanding every dataset and put them together when every data set is ready to have the same exact column name with no duplicated data.



	N	P	K	temperature	humidity	ph	rainfall	label
2	90	42	43	20.87974371	82.00274423	6.502985292	202.9355362	rice
3	85	58	41	21.77046169	80.31964408	7.038096361	226.6555374	rice
4	60	55	44	23.00445915	82.3207629	7.840207144	263.9642476	rice
5	74	35	40	26.49109635	80.15836264	6.980400905	242.8640342	rice
6	78	42	42	20.13017482	81.60487287	7.628472891	262.7173405	rice
7	69	37	42	23.05804872	83.37011772	7.073453503	251.0549998	rice
8	69	55	38	22.70883798	82.63941394	5.70080568	271.3248604	rice
9	94	53	40	20.27774362	82.89408619	5.718627178	241.9741949	rice
10	89	54	38	24.51588066	83.5352163	6.685346424	230.4462359	rice
11	68	58	38	23.22397386	83.03322691	6.336253525	221.2091958	rice
12	91	53	40	26.52723513	81.41753846	5.386167788	264.6148697	rice
13	90	46	42	23.97898217	81.45061596	7.50283396	250.0832336	rice
14	78	58	44	26.80079604	80.88684822	5.108681786	284.4364567	rice
15	93	56	36	24.01497622	82.05687182	6.98435366	185.2773389	rice
16	94	50	37	25.66585205	80.66385045	6.94801983	209.5869708	rice
17	60	48	39	24.28209415	80.30025587	7.042299069	231.0863347	rice
18	85	38	41	21.58711777	82.7883708	6.249050656	276.6552459	rice
19	91	35	39	23.79391957	80.41817957	6.970859754	206.2611855	rice
20	77	38	36	21.8652524	80.1923008	5.953933276	224.5550169	rice
21	88	35	40	23.57943626	83.58760316	5.85393208	291.2986618	rice

crops_data

Analysing of Data : In order to find patterns and trends that may be utilised to train the model, data analysis is a crucial phase in the crop prediction project. We can choose the most pertinent and practical elements to incorporate into the model by carefully examining the data, and we can also see any potential issues or biases that can affect the model's performance. This makes it possible for us to create a model that is more precise and efficient and that can anticipate crop growth with greater reliability. We analyse the data by using many python libraries like Matplotlib, Seaborn, Numpy and pandas by plotting different graphs with many parameters to analyse the correlation between them and to know more about the importance of every parameter.

Data Preprocessing : In order to find patterns and trends that may be utilised to train the model, data analysis is a crucial phase in the crop prediction project. We can choose the most pertinent and practical elements to incorporate into the model by carefully examining the data, and we can also see any potential issues or biases that can affect the model's performance. This makes it possible for us to create a model that is more precise and efficient and that can anticipate crop growth with greater reliability. It plays a key role in developing a model that is accurate, reliable, and efficient. By carefully preprocessing the data, we can improve the performance of the model and ensure that it is able to make accurate predictions about crop growth. We preprocess our data by dividing them into two different tables in which one contains all the parameters except the crop name and the other table consists of only crop name with its unique id.

Feature Scaling: We can change the data so that each feature is on a comparable scale by using feature scaling. Several techniques, including normalization, standardization, and min-max scaling, can be used to accomplish this. We can increase the data's suitability for machine learning techniques and the model's

accuracy and dependability by scaling the features. After Data Preprocessing, we have to scale the dataset into two parts: training and testing data which help to find out the accuracy of the model.

Model Selection: We can choose the best algorithms to utilize for the crop prediction project by conducting a thorough examination of the data and the issue we are attempting to solve. This may entail contrasting various algorithms according to how well they work on the data, how sophisticated they are, and how well they generalize to new data. We can make sure that the model is able to estimate crop growth accurately by choosing the most suitable algorithm. In Model selection we select KNN, Decision Tree and Random Forest, Logistic Regression, Naive Bayes, Linear SVM. After selecting 3 different models we implement every single one and at the end we take the Decision Tree as it predicts more accurately than the other two algorithms.

Crop Prediction: The system will suggest the best crop for cultivation based on the amount of anticipated rainfall, the composition of the soil, and weather parameters. This approach also displays the amount of seed needed to cultivate a recommended crop in parts per million. weather parameters. This approach also displays the amount of seed needed to cultivate a recommended crop in parts per million.

Confusion Matrix and Classification Report

Confusion Matrix and Classification Report are the methods imported from the metrics module in the scikit learn library that are calculated using the actual labels of test datasets and predicted values. Confusion Matrix gives the matrix of frequency of true negatives, false negatives, true positives and false positives. Classification Report is a metric used for evaluating the performance of a classification algorithm's predictions.

It gives three things: Precision, Recall and f1-score of the model.

Precision refers to a classifier's ability to identify the number of positive predictions which are relatively correct. It is calculated as the ratio of true positives to the sum of true and false positives for each class.

$$Precision = \frac{TP}{TP + FP}$$

Where Precision-Positive Prediction Accuracy; TP-True Positive; FP-False Positive

Recall is the capability of a classifier to discover all positive cases from the confusion matrix. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class.

$$Recall = \frac{TP}{TP + FN}$$

Where Recall- The percentage of positives that were correctly identified; FN-False Negative.

F1 score is a weighted harmonic mean of precision and recall, with 0.0 being the worst and 1.0 being the best.Since precision and recall are used in the computation, F1 scores are often lower than accuracy measurements.

$$\text{F1 score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

Where P-Precision; R-Recall

Accuracy: The number of correct predictions divided by the total number of predictions accuracy. The accuracy of the model is calculated using the accuracy_score method of scikit learn module.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Where TP-True Positive; FP-False Positive; TN : True Negative; FN : False Negative

Proposed Work

K-Nearest Neighbour Classifier :

The K-Nearest Neighbour algorithm is based on the supervised learning technique and is a simple machine learning algorithm. The K-NN technique saves all possible and classifies the incoming data depending on how similar they are to the actual data. This means that the K-NN technique can swiftly classify new instances into a precisely defined category. The KNN technique can be used in both regression and classification problems but it is most likely to be used in classification.
KNN technique has two properties. First, the model is based on the dataset or, it is not required to identify parameters for the distribution. Hence it is referred to as non-parametric. Second, there is no learning taking place; instead, it just stores the training data. The classification of the dataset happens during the testing phase, due to which the testing phase becomes time-consuming and takes a lot of memory. This property is known as lazy-learner.

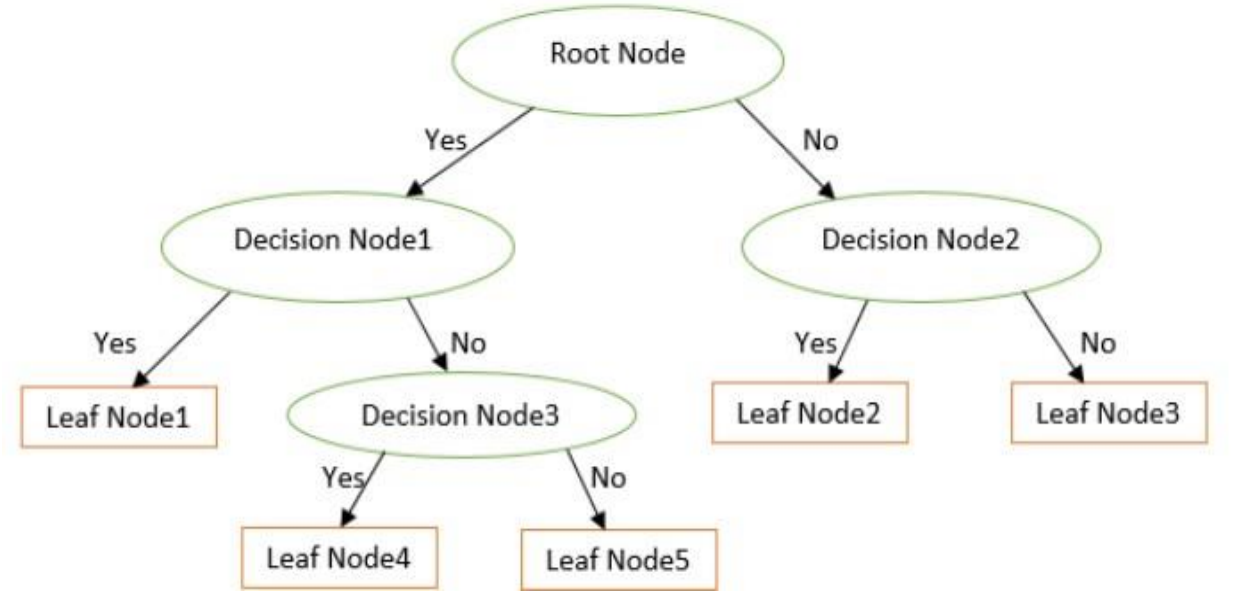
Step 1: K, i.e., the number of neighbours is selected. The primary deciding factor is the number of neighbours.
Step 2: Using distance measures, determine the distance between two points like Euclidean distance

$$\text{Euclidean distance} = d(b,a) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$

Step 3: K nearest neighbours are taken into account according to the calculated Euclidean distance.
Step 4: Figure the number of data points in each class surrounded by these
Step 5: The class with the highest number of neighbours is assigned to the new data points.
Step 6: The label is voted on, and the model is ready.

Decision Tree Classifier :

Decision Tree is a supervised learning technique used where each path is a set of decision instances from the training set. The non attributes. The decision is made by comparing the instance with the split and jumping to the next node. This splitting continues, generating sub nodes which determine class labels for that instance. It divides recursive partitioning. With high accuracy, decision trees are capable of handling high data. It's a flowchart diagram-style representation that closely parallels human result, decision trees are simple to explain and apprehend.



Step 1: Starting with the root node of the tree, which consists of the

Step 2: The most appropriate attribute is obtained from the dataset by applying the Attribute Selection Measure (ASM).

Step 3: The S is divided into subdivisions that attributes.

Step 4: The node is formed in the decision tree with the most appropriate attri Step 5: The tree formation is setup by iteratively repeating this method for each child until one of the following requirements is met:

• The tuples are entirely correlated with the same attribute value.

• There are no further attributes accessible.

• There aren't any more instances.

Steps to Split. The dataset used in the project has with the numerical values in the following ways:

Step 1: Sorting all the values.

Step 2: It will consider a threshold value from

Step 3: Feature value will split into two parts such that threshold value, and the right node contains feature values greater than

Step 4: The next feature value will be considered as a threshold value and again create the same split .

Step 5: Entropy/Gini and Information Gain are split with better information gain is considered.

$$I(Attribute) = \frac{\sum p_i + n_i}{p + n}$$

Where pi denotes the number of yes values; n attribute; p and n are the numbers of yeses and noes of the entire sample, respectively.

$$Information\ Gain = Entropy(S) - I(Attribute)$$

Where Entropy(S) denotes entropy of sample S; I(Attribute) denotes Average Information of the particular attribute.

Step 6: Repeat from Step 2 to Step 5. In this way it will get branches for the decision tree .

Entropy and Gini Index : The criteria for measuring Information Gain are the Gini index and Entropy. Information gain is a measurement of how much the reduction in entropy. Entropy and Gini Index are the metrics that measure the impurity of the nodes. A node is considered as impure if it has multiple classes, else Entropy is a metric that gives the degree of impurity in a specified attribute. The following formula can be used to compute entropy:

Entropy : is a metric that gives the degree of impurity in a specified attribute. The following formula can be used to compute entropy:

$$Entropy(S) = -P(yes)\ log_2\ P(yes) - P(no)\ log2\ P(no)$$

$$Entropy(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

Where S denotes a complete sample, P(yes) denotes the probability of yes , P(no) denotes the probability of no .

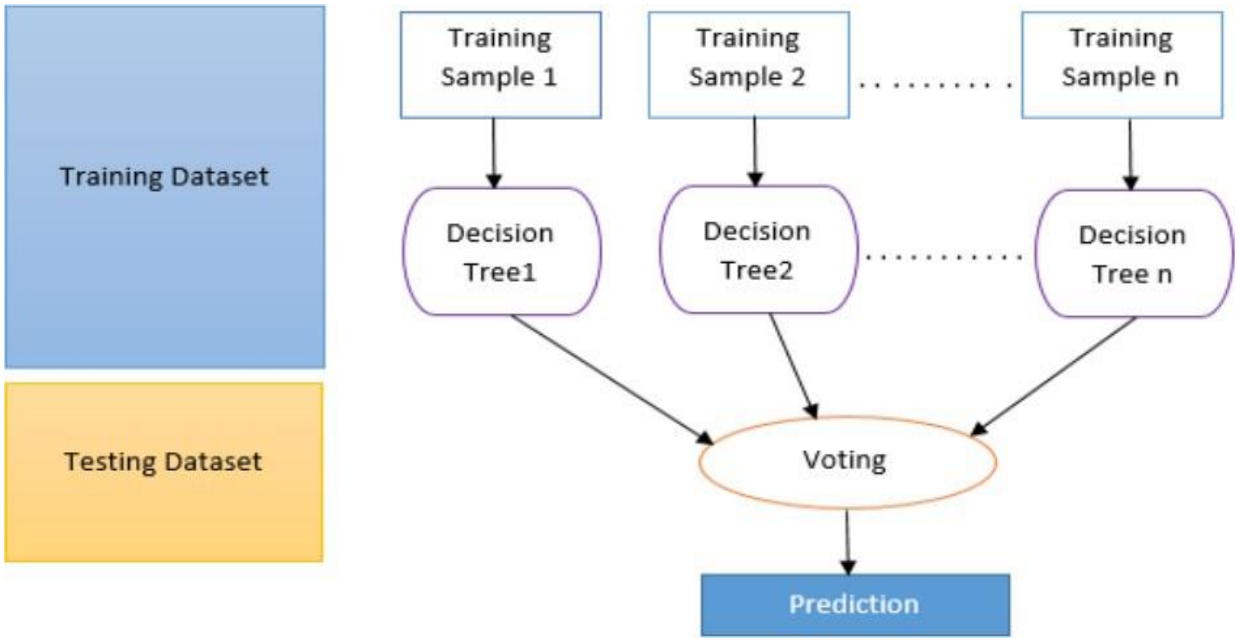
Gini Index: Gini is estimated by deducting the sum of squared probabilities of each class from one. The lower Gini Index value is preferred rather than a higher value. Scikit is the default value and supports "Gini" criteria for Gini Index.

$$Gini\ Index = 1 - \sum_{i=1} (P_i^2)$$

Where Pi denotes the probability of a tuple , say R belonging to class Ci .

Random Forest Classifier

The Random Forest method consists of multiple decision tree classifiers to enhance the model's performance. Here ensemble learning is a supervised learning algorithm. Decision trees are created at random using the instances from the training set. Each of the decision trees the model is decided by majority voting. One of the reasons for its popularity as a machine learning approach is that it can handle the issue of overfitting trees.



Step 1: K instances are chosen at random from the given training

Step 2: Decision trees are created for the chosen instances.

Step 3: The N is selected for the number of estimators to be created.

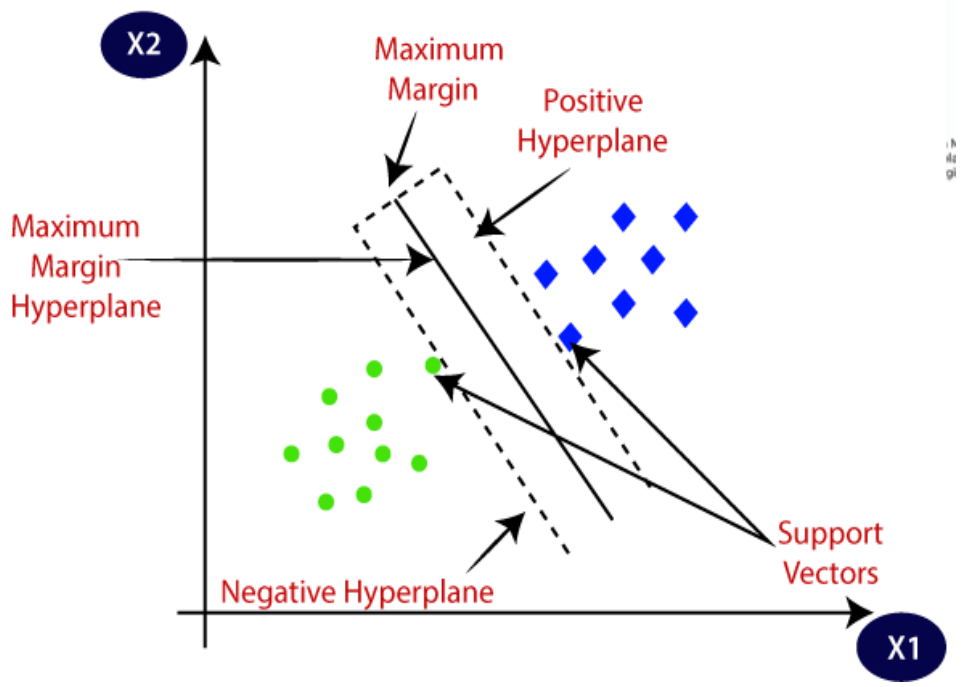
Step 4: Here Step 1 & Step 2 is repeated.

Step 5: For the new instance, the predictions of each estimator is determined, and the category with the Hughes votes is assigned .

SVM :

SVM is used for Classification as well as Regression problems .The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine



Here is this diagram of two different categories that are classified using a decision boundary or hyperplane .

Type of SVM :

1. **Linear**
2. **Non -Linear**

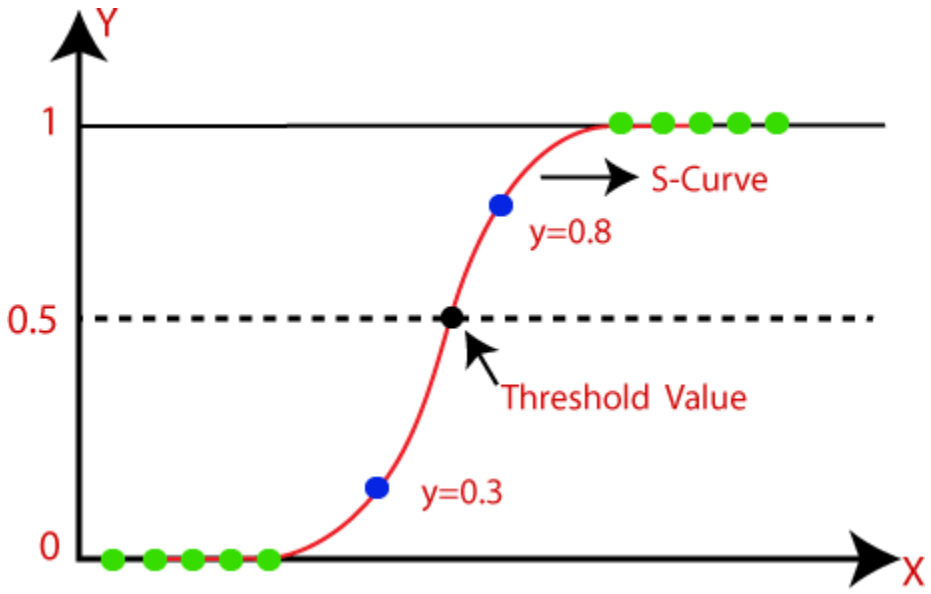
Here we will use a linear SVM for prediction .

Logistic Regression :

Logistic Regression is used for predicting the categorical dependent variable using a given set of independent variables.The outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Logistic Rcgícssio⇥ Equatio⇥ :

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:



We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$\frac{y}{1-y}$ **; 0 for y= 0, and infinity for y=1**

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Stcps i⇥ Logistic Rcgícssio⇥:

- **Data Preprocessing step**
- **Fitting Logistic Regression to the Training set**
- **Predicting the test result**
- **Test accuracy of the result(Creation of Confusion matrix)**
- **Visualizing the test set result.**

Naive Bayes :

- **The Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.**
- **It is mainly used in text classification that includes a high-dimensional training dataset.**
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.**
- **Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.**

Bayes' Theorem:

- **Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.**
- **The formula for Bayes' theorem is given as:**

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Working of Naïve Bayes' Classifier :

Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

1. **Convert the given dataset into frequency tables.**
2. **Generate Likelihood table by finding the probabilities of given features.**
3. **Now, use Bayes theorem to calculate the posterior probability.**

Problem: If the weather is sunny, then the Player should play or not?

Solution: To solve this, first consider the below dataset:

	Outlook	Play
0	Rainy	Yes

1	**Sunny**	**Yes**
2	**Overcast**	**Yes**
3	**Overcast**	**Yes**
4	**Sunny**	**No**
5	**Rainy**	**Yes**
6	**Sunny**	**Yes**
7	**Overcast**	**Yes**
8	**Rainy**	**No**
9	**Sunny**	**No**
10	**Sunny**	**Yes**
11	**Rainy**	**No**
12	**Overcast**	**Yes**
13	**Overcast**	**Yes**

Ïícquc→ıcQ tablc foí tkc Wcatkc Co→ıditio→ıs:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Likclikood tablc wcatkcí co→ditio→:

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

ApplQi→g BaQcs lkcoícm :

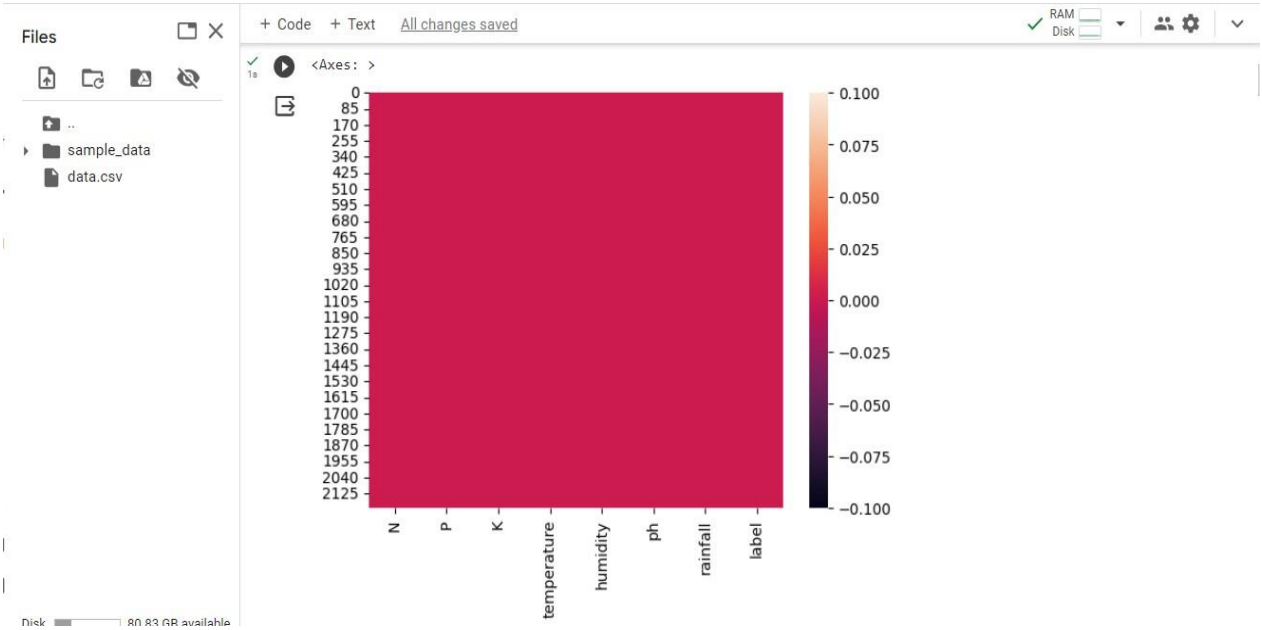
P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)

P(Sunny|Yes)= 3/10= 0.3

P(Sunny)= 0.35

P(Yes)=0.71

Results and Discussion :



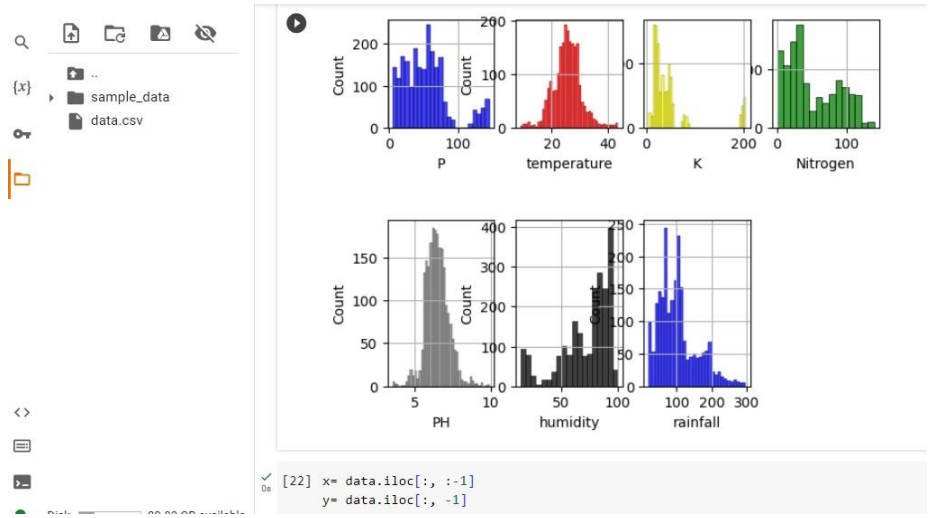
+ Code + Text All changes saved

[50]

crops rice

```
min N required  60
Avg N required  79.89
Max N required  99
min P required  35
Avg P required  47.58
Max P required  60
min K required  35
Avg K required  39.87
Max K required  45
min temperature required  20.0454142
Avg temperature required  23.6893322105
Max temperature required  26.92995077
min humidity required  80.12267476
Avg humidity required  82.27282153889999
Max humidity required  84.96907151
min ph required  5.005306977
Avg ph required  6.425470922139999
Max ph required  7.868474653
min rainfall required  182.5616319
Avg rainfall required  236.18111359399998
Max rainfall required  298.5601175
```

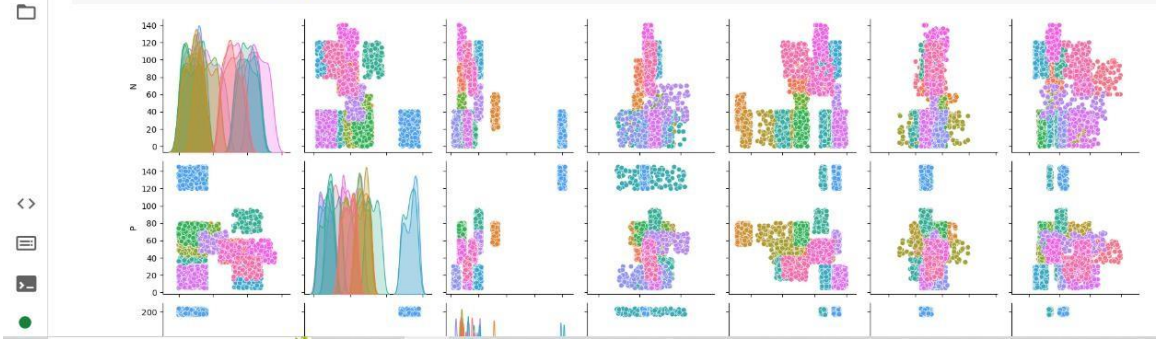
```
plt.subplot(3,4,4)
sns.histplot(data['N'],color="green")
plt.xlabel("Nitrogen")
plt.grid()

plt.subplot(3,4,1)
sns.histplot(data['P'],color="blue")
```




```
[22] x= data.iloc[:, :-1]
     y= data.iloc[:, -1]
```

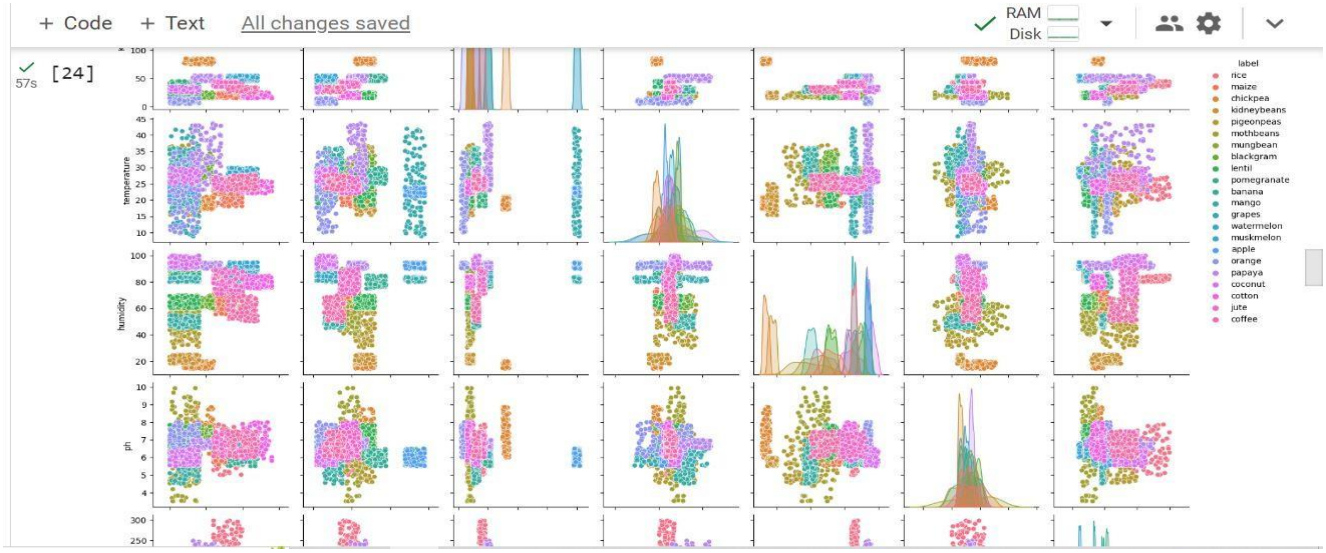
```
#shuffle the data to avoid bias
[23] x_train, x_test, y_train, y_test= train_test_split(x, y,test_size= 0.2,shuffle= True,random_state= 0)
x_train= np.asarray(x_train)
y_train= np.asarray(y_train)
x_test= np.asarray(x_test)
y_test= np.asarray(y_test)
```

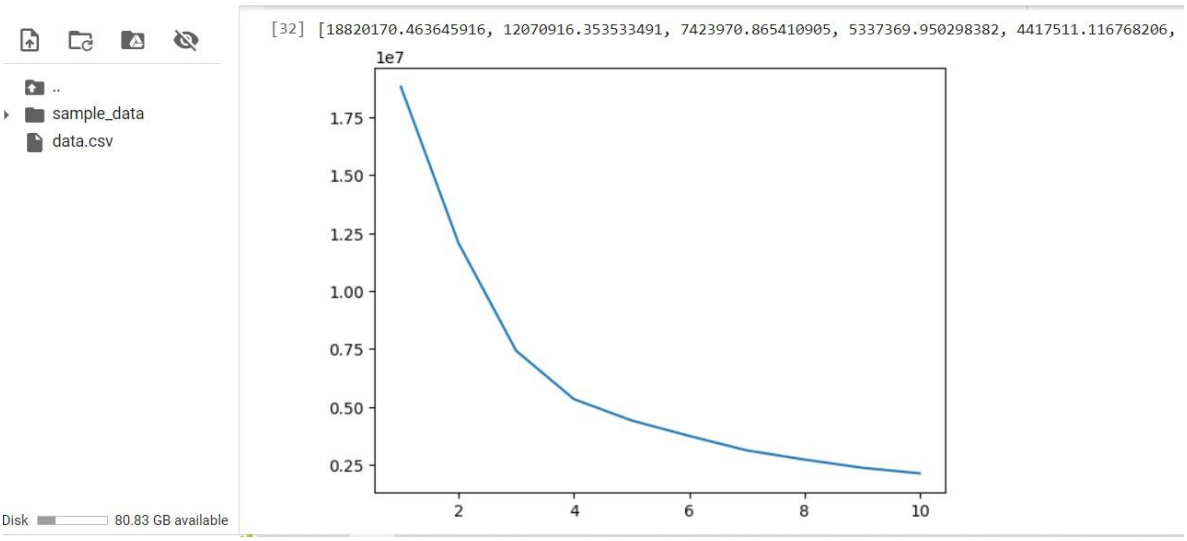
```
[24] before= sns.pairplot(data, hue= 'label')
```

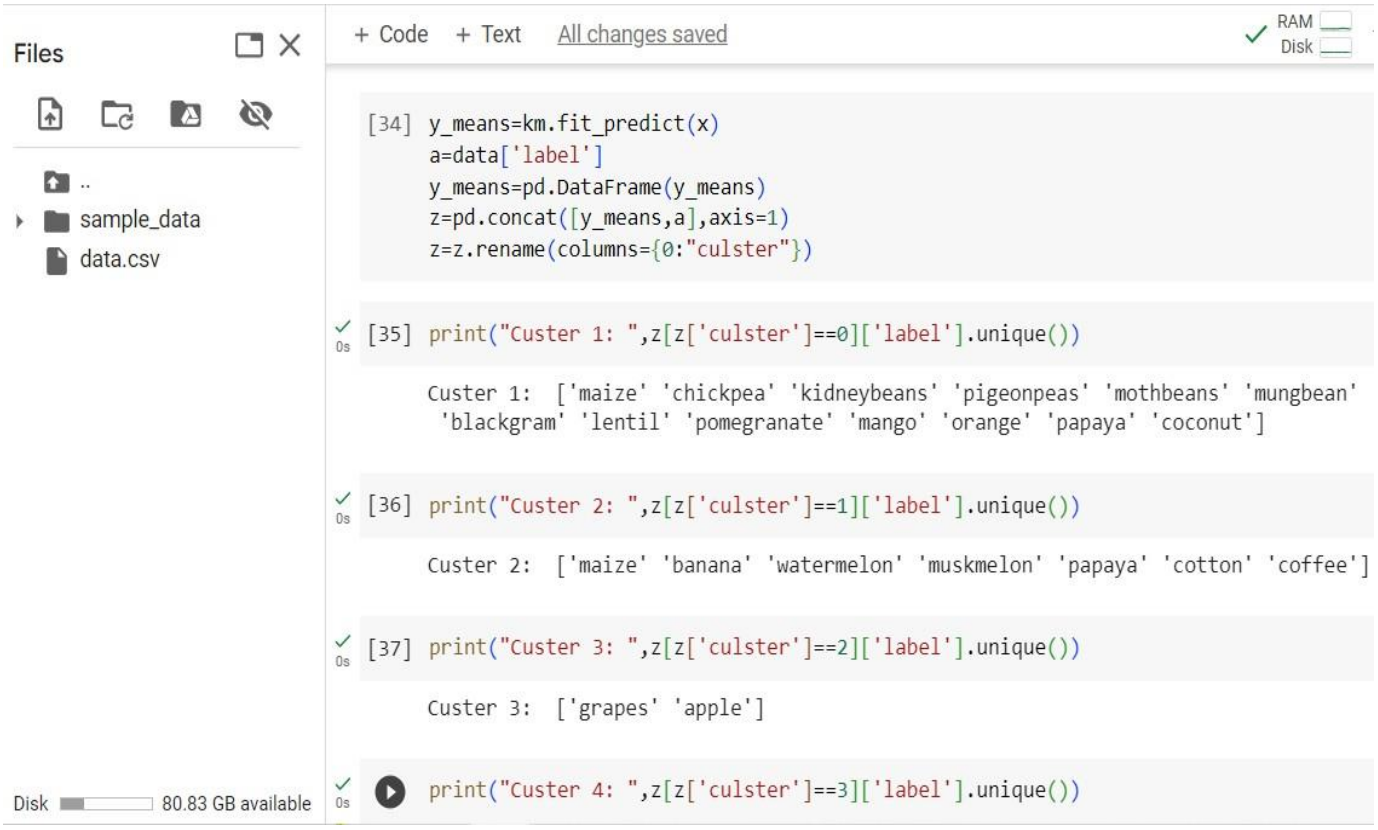


+ Code + Text All changes saved RAM Disk

[24]




[32] [18820170.463645916, 12070916.353533491, 7423970.865410905, 5337369.950298382, 4417511.116768206,
1e7
1.75
1.50
1.25
1.00
0.75
0.50
0.25
2
4
6
8
10
..
sample_data
data.csv
Disk
80.83 GB available


Files
+ Code
+ Text
All changes saved
RAM
Disk
..
sample_data
data.csv
[34] y_means=km.fit_predict(x)
a=data['label']
y_means=pd.DataFrame(y_means)
z=pd.concat([y_means,a],axis=1)
z=z.rename(columns={0:"culster"})
[35] print("Custer 1: ",z[z['culster']==0]['label'].unique())
Custer 1: ['maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'mothbeans' 'mungbean'
'blackgram' 'lentil' 'pomegranate' 'mango' 'orange' 'papaya' 'coconut']
[36] print("Custer 2: ",z[z['culster']==1]['label'].unique())
Custer 2: ['maize' 'banana' 'watermelon' 'muskmelon' 'papaya' 'cotton' 'coffee']
[37] print("Custer 3: ",z[z['culster']==2]['label'].unique())
Custer 3: ['grapes' 'apple']
print("Custer 4: ",z[z['culster']==3]['label'].unique())
Disk
80.83 GB available

Discussion:

General discussion : **Such research is susceptible to threats to validity, and potential threats to validity can be external, construct validity, and reliability. The external validity and construct validity are addressed for this SLR study since the initial search string was broad, and the query returned a substantial number of studies: 567 publications in total.**

Search-related discussion : **There is a possibility that valuable publications might have been missed. More synonyms could have been used, and a broader search could have returned new studies. However, the search string resulted in a high number of publications indicating a broad enough search.**

Analysis-related discussion : **Another issue that could be a threat to the validity of the way the analysis is conducted. For example, not all publications stated what kind of evaluation parameters were used, and sometimes just a few examples of features were explained.**

RQ1-Related (algorithms) discussion : **Linear Regression is the second most used algorithm. Linear Regression is used as a benchmarking algorithm in most cases to check whether the proposed algorithm is better than Linear Regression or not. Therefore, although it is shown in many articles, it does not mean that it is the best-performing algorithm.**

RQ2-related (features) discussion : **Groups are created for features and algorithms to visualize the main features and algorithms. Due to this decision, detailed information is lost, but clarity has been maintained. The most used features are soil type, rainfall, and temperature. Apart from those features that are used in several studies, some features were used in specific studies.**

RQ3-related (evaluation parameters and approaches) discussion : **There are not many evaluation parameters reported in the selected papers. Almost every study used RMSE as the measurement of the quality of the model. Other evaluation parameters are MSE, R2, and MAE. Some parameters were used in specific studies, but most of these parameters look like some of the previously mentioned parameters, with a small difference. These are MAPE, LCCC, MFE, SAE, RCV, RSAE, and MCC. Most of the models had outcomes with high accuracy values for their evaluation parameters, which means that the model made correct predictions. As the evaluation approach, the 10-fold cross-validation approach was preferred by researchers.**

Q4-related (challenges) discussion : **Challenges were reported based on the explicit statements in the articles. However, there might be additional challenges that were not**

stated in the identified papers. The challenges are mainly in the field of improvement of a working model. When more data is gathered to train and test, much more can be said about the precision of the model. Another challenge is the implementation of the models into the farm management systems. When applications are made that the farmer can use, then only can the models be useful to make decisions, and also during the growing season. When specific parameters for that specific place are measured and added, predictions will have higher precision.

Table of crop type :

***Crop type* - The name of the crop being studied.**

***Dataset* - The source of the data used for the study.**

***Feature selection* - The features or variables used to train the machine learning algorithm.**

***Machine learning algorithm* - The name of the algorithm used for crop yield prediction.**

***Model accuracy* - The accuracy of the model in predicting crop yield (usually in percentage).**

***Evaluation metric* - The metric used to evaluate the performance of the model (e.g., RMSE, MAE, R^2).**

***Experiment setup* - The parameters used for the experiment, such as training and testing split, hyperparameters tuning, and cross-validation.**

Accuracy :

The accuracy of crop yield prediction using machine learning algorithms depends on various factors such as the quality and quantity of data used, the choice of machine learning algorithm, and the specific crop and region being studied. Generally, the accuracy of the prediction is evaluated using evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

It is important to note that the accuracy of the machine learning model is not the only factor to consider when predicting crop yield. Other factors such as the cost and feasibility of data collection and the impact of external factors such as weather and pests on crop yield also need to be considered. Nevertheless, crop yield prediction using machine learning algorithms has the potential to significantly improve crop yields and reduce waste, leading to increased profitability for farmers.

Conclusion :

Crop yield prediction using machine learning algorithms has the potential to revolutionize the field of agriculture. By analyzing large amounts of data on factors such as weather, soil conditions, and historical crop yields, machine learning algorithms can provide accurate predictions of crop yields, allowing farmers to make more informed decisions and increase their productivity.

This project aimed to develop a sophisticated crop prediction model by leveraging advanced data analytics and machine learning techniques to address the growing challenges in modern agriculture. By integrating diverse datasets, including historical yield data, weather patterns, soil conditions, and remote sensing information, the project has achieved significant advancements in forecasting crop yields with enhanced accuracy.

Overall, crop yield prediction using machine learning algorithms has the potential to significantly improve crop yields and reduce waste, leading to increased profitability for farmers and a more sustainable agricultural industry.This study showed that the selected publications use a variety of features, depending on the scope of the research and the

The application of machine learning algorithms, such as regression models, ensemble methods, and neural networks, has substantially improved the accuracy of yield forecasts compared to traditional statistical methods. This enhancement allows for more reliable predictions, which can help in optimizing planting schedules and resource allocation.

References :

- https://www.javatpoint.com/crop-yield-prediction-using-machine-learning
- https://link.springer.com/chapter/10.1007/978-3-031-43145-6_5
- https://github.com/topics/crop-prediction