CROP ANALYSIS AND PREDICITION

Sruti Lakshmi Narasimhan

INTRODUCTION

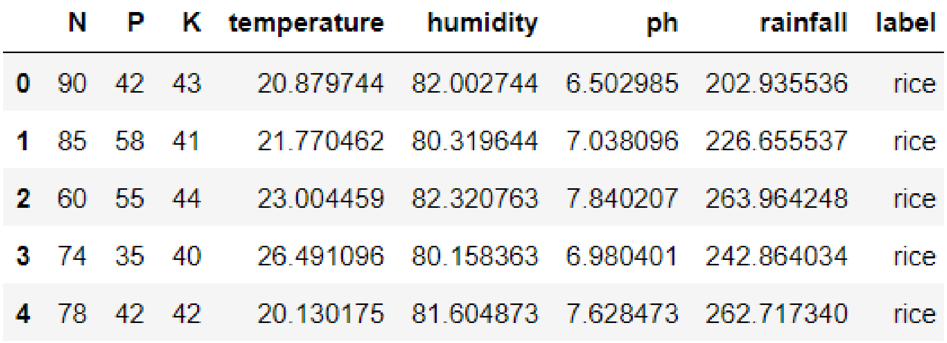
India is characterized by small farms. Over 75% of total land capitals within the country are less than 5 acres. Most crops are rain nourished, with just about 45% of the land irrigated. As per some estimations, about 55% of total population of India depends on farming. In the US, because of heavy mechanization of agriculture, it is about 5%.

Agriculture in India plays a major role in economy and employment. The common difficulty present among the Indian farmers are they don't opt for the proper crop based on their soil necessities. Because of this productivity is affected. This problem of the farmers has been solved through precision agriculture. This method is characterized by a soil database collected from the farm, crop provided by agricultural experts, achievement of parameters such as soil through soil testing lab dataset. The data from soil testing lab given to recommendation system it will use the collect data and do ensemble model with majority voting technique using support vector machine (SVM) and ANN as learners to recommend a crop for site specific parameter with high accuracy and efficiency.

ABOUT THE DATASET

Precision agriculture is in trend nowadays. It helps the farmers to get informed decision about the farming strategy. Here, I present you a dataset which would allow the users to build a predictive model to recommend the most suitable crops to grow in a particular farm based on various parameters.

This dataset was build by augmenting datasets of rainfall, climate and fertilizer data available for India.



	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

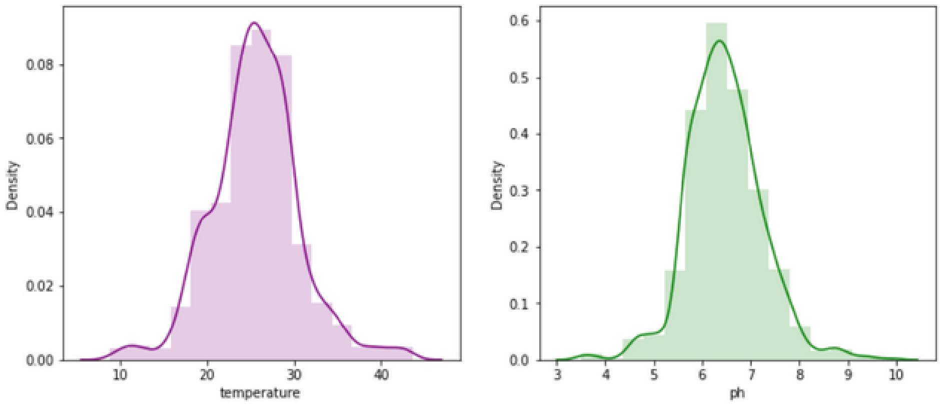
The Data fields in this dataset are: N, P, K, Temprature, Humidity, ph, rainfall and label.
N - ratio of Nitrogen content in soil
P - ratio of Phosphorous content in soil
K - ratio of Potassium content in soil
temperature - temperature in degree Celsius
humidity - relative humidity in %
ph - ph value of the soil
rainfall - rainfall in mm
label- suitable crop

DATA VISUALISATION

There are 22 different crops and they are rice, maize, chickpea, kidneybeans, pigeonpeas, mothbeans, mungbean, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, coffee. The count of each crop is 100

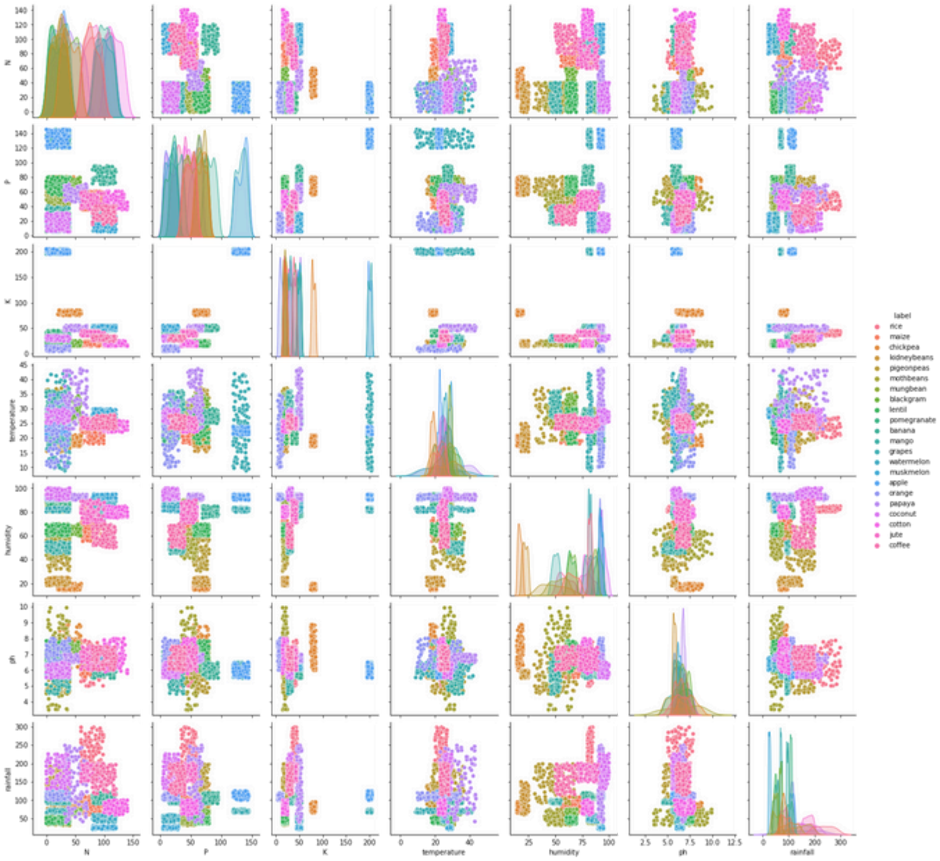
01. Distribution of temprature and ph

It is symmetrical and bell shaped, showing that trials will usually give a result near the average, but will occasionally deviate by large amounts. It's also fascinating how these two really resemble each other!



02. Pair plot analysis

A very important plot to visualize the diagonal distribution between two features for all the combinations! It is great to visualize how classes differ from each other in a particular space.

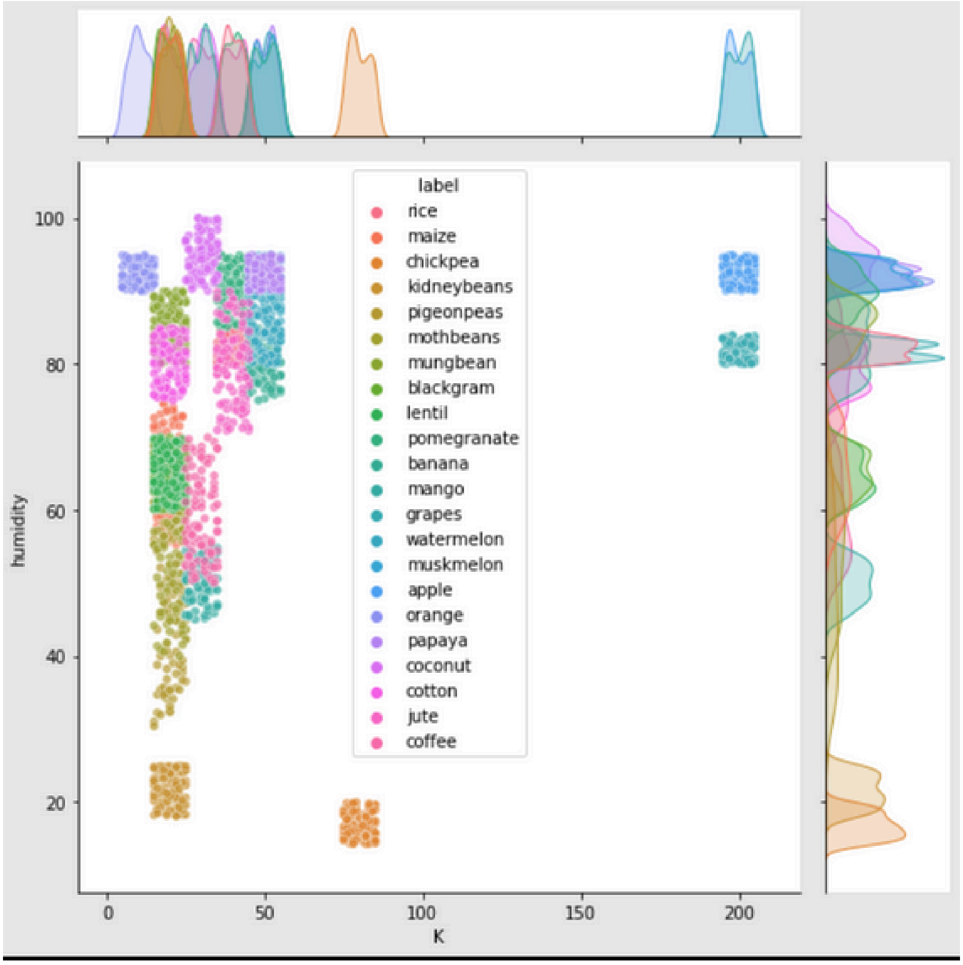


03. **Rain fall required**

During rainy season, average rainfall is high (average 120 mm) and temperature is mildly chill (less than 30'C).

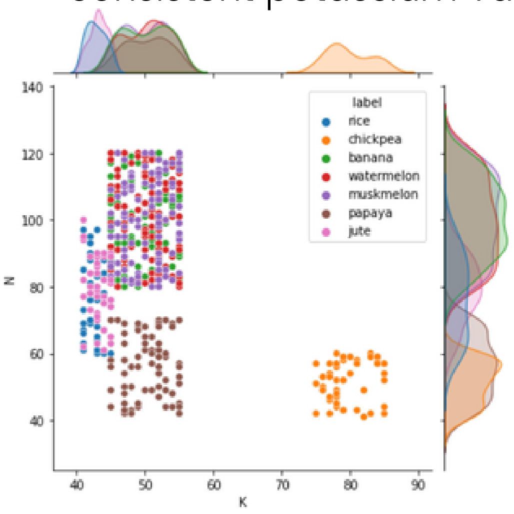
Rain affects soil moisture which affects ph of the soil. Here are the crops which are likely to be planted during this season.

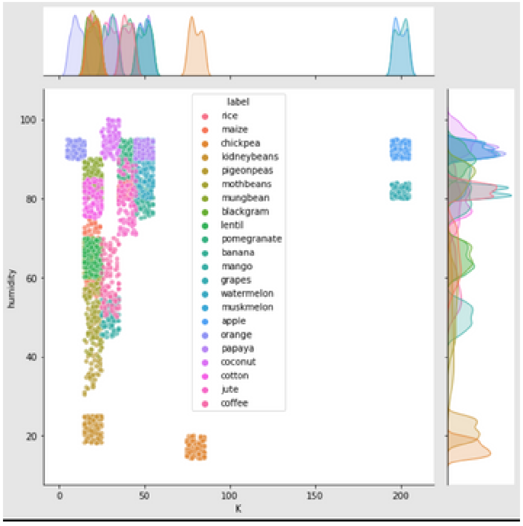
- Rice needs heavy rainfall (>200 mm) and a humidity above 80%. No wonder major rice production in India comes from East Coasts which has average of 220 mm rainfall every year!
- Coconut is a tropical crop and needs high humidity therefore explaining massive exports from coastal areas around the country.



04. Potassium and Nitrogen analysis

This graph correlates with average potassium (K) and average nitrogen (N) value (both>50).
These soil ingredients direcly affects nutrition value of the food. Fruits which have high nutrients typically has consistent potassium values



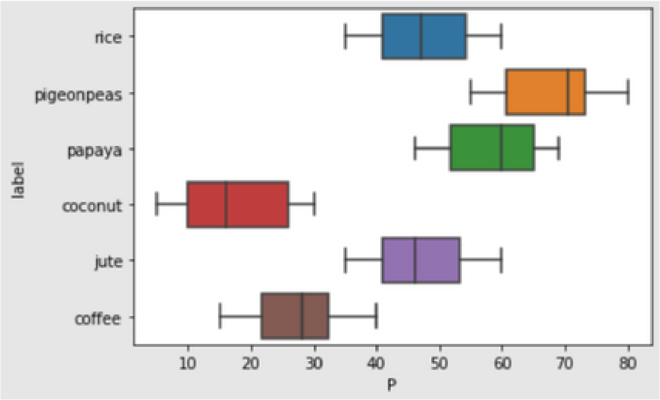


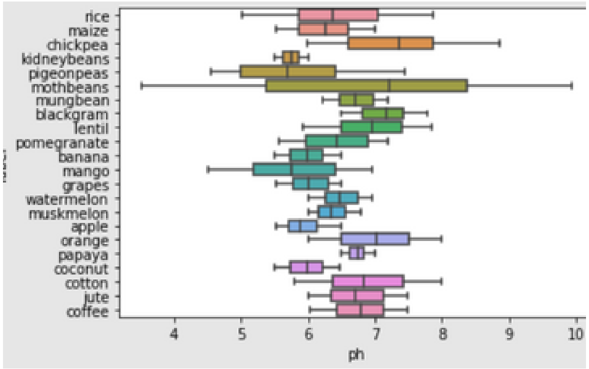
05. Humidity and Potassium analysis

Now plotting a specfic case of pairplot between `humidity` and `K` (potassium levels in the soil.)
sns.jointplot() can be used for bivariate analysis to plot between humidity and K levels based on Label type. It further generates frequency distribution of classes with respect to features

06. Ph analysis

We can see ph values are critical when it comes to soil. A stability between 6 and 7 is preffered



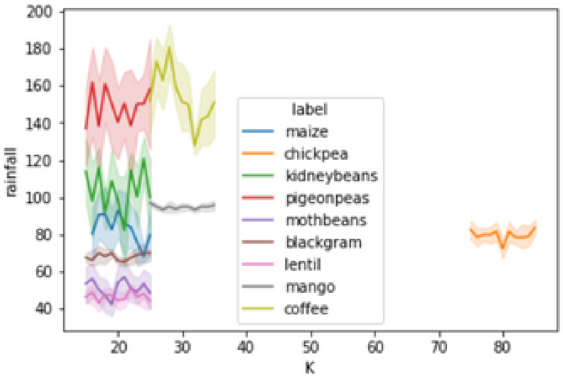


07. Phosphorus analysis

We can infer that Phosphorous levels are quite differentiable when it rains heavily (above 150 mm).

08. Potassium analysis

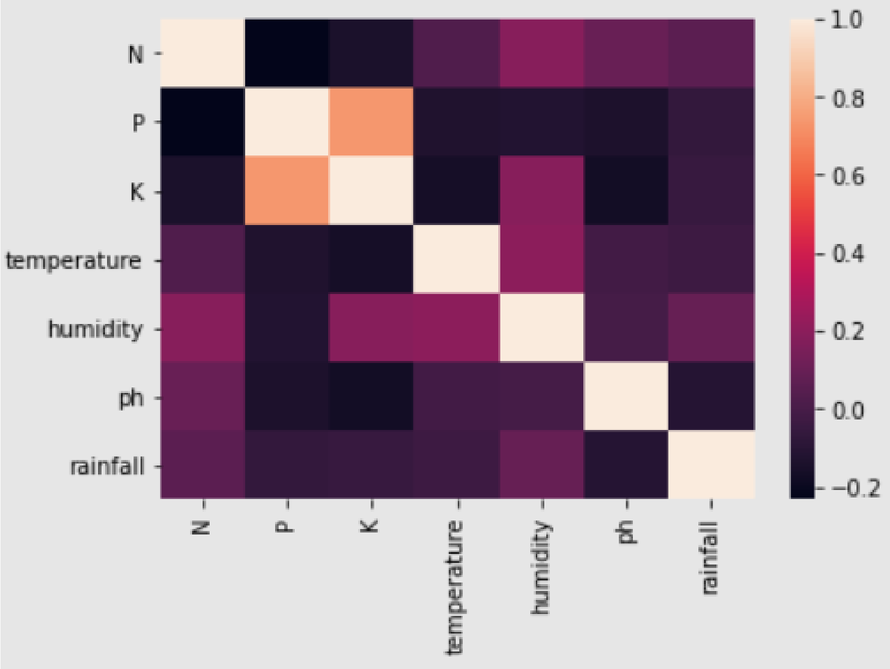
When humidity is less than 65, almost the same potassium levels(approx. 14 to 25) are required for few crops which could be grown just based on the amount of rain expected over the next few weeks.



DATA
PRE PROCESSING
AND FETURE SCALING

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process.

Making the data ready for the machine learning model. Now we are performing correlation visualization between features. We can see how Phosphorous levels and Potassium levels are highly correlated.



Feature scaling is required before creating training data and feeding it to the model. As we saw earlier, two of our features (temperature and ph) are gaussian distributed, therefore scaling them between 0 and 1 with MinMaxScaler.

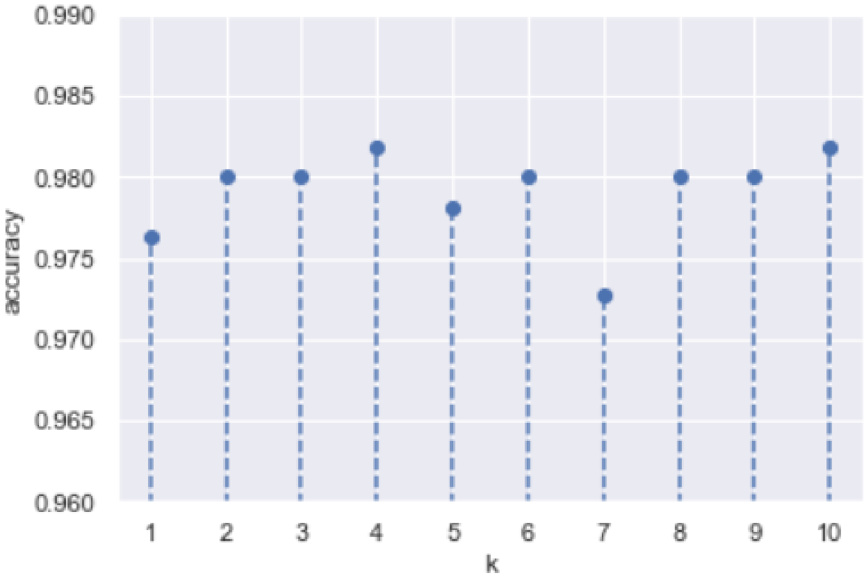
MODEL SELECTION

The ML models used in this project are KNN, SVC Decision tree and gradient boosting

01. KNN

The accuracy is 97%.

The different values of n_neighbors to fine tune and get better results



02. Gradient boosting

The accuracy is 99%.

03. Support Vector Classifier (svc)

The accuracy of

Linear Kernel Accuracy: 0.9745454545454545

Rbf Kernel Accuracy: 0.9872727272727273

Poly Kernel Accuracy: 0.9890909090909091

Then used Grid search CV to find the best parameters.

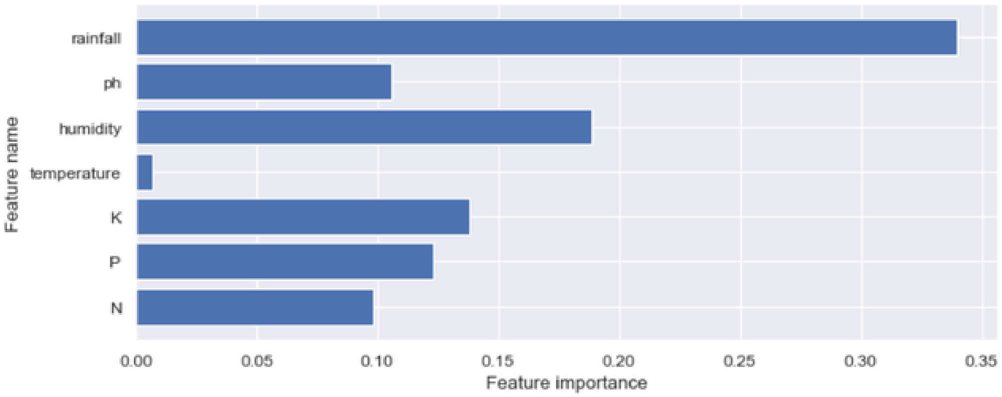
Points to be highligted

1. Interestingly liner kernel also gives satisfactory results but fine tuning increases the computation and might be inefficient in some cases
2. The accuracy can be increased in poly kernel by tweaking parameters but might lead to intensive overfitting.
3. RBF has better result than linear kernel.
4. Poly kernel so far wins by a small margin.

04. Decision Tree

The accuracy Is 98.7%

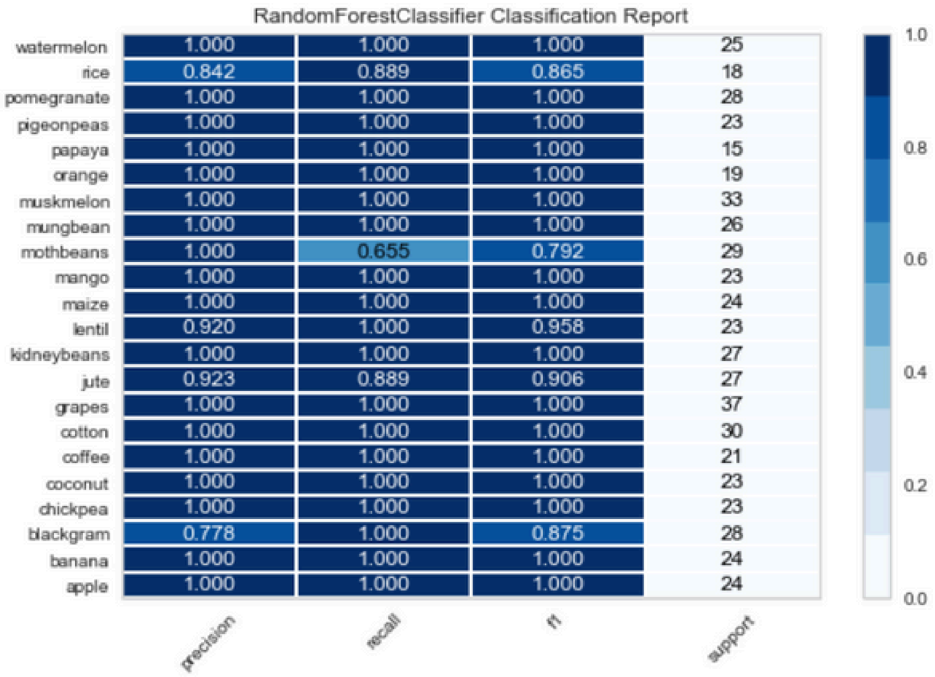
Visualizing the import features which are taken into consideration by decision trees.



05. Random forest

The accuracy Is 97% in both training and test set

We are creating classification report for the random forest using yellowbrick as they are great for visualizing in a tabular format



RandomForestClassifier Classification Report

	precision	recall	f1	support
watermelon	1.000	1.000	1.000	25
rice	0.842	0.889	0.865	18
pomegranate	1.000	1.000	1.000	28
pigeonpeas	1.000	1.000	1.000	23
papaya	1.000	1.000	1.000	15
orange	1.000	1.000	1.000	19
muskmelon	1.000	1.000	1.000	33
mungbean	1.000	1.000	1.000	26
mothbeans	1.000	0.655	0.792	29
mango	1.000	1.000	1.000	23
maize	1.000	1.000	1.000	24
lentil	0.920	1.000	0.958	23
kidneybeans	1.000	1.000	1.000	27
jute	0.923	0.889	0.906	27
grapes	1.000	1.000	1.000	37
cotton	1.000	1.000	1.000	30
coffee	1.000	1.000	1.000	21
coconut	1.000	1.000	1.000	23
chickpea	1.000	1.000	1.000	23
blackgram	0.778	1.000	0.875	28
banana	1.000	1.000	1.000	24
apple	1.000	1.000	1.000	24

REFRENCE

1. https://www.kaggle.com/code/theeyeschico/crop-analysis-and-prediction/notebook
2. https://d1wqtxts1xzle7.cloudfront.net/55495855/IRJET-V4I12179-with-cover-page-v2.pdf