CROP RELATED PATTERN AND CROP DISEASE DETECTION USING DATA MINING AND DENSENET-121

A PROJECT REPORT

Submitted by

SAMUNDAR SINGH

(2019272033)

A report for the phase-I of the project
submitted to the Faculty of

INFORMATION SCIENCE AND TECHNOLOGY

in partial fulfilment
for the award of the degree of

MASTER
of
COMPUTER APPLICATIONS

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600 025
APRIL 2022

BONAFIDE CERTIFICATE

Certified that this project report titled CROP RELATED PATTERN AND CROP DISEASE DETECTION USING DATA MINING AND DENSENET-121 is the bona fide work of Samundar Singh who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE: CHENNAI
DATE:

DR. R. GEETHA RAMANI
PROFESSOR
PROJECT GUIDE
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

DR. S. SRIDHAR
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

India is an agriculture-based economy where climate and other environmental changes have become a major threat in the agricultural field. Machine learning is an essential approach for achieving partial and effective solution for this problem. One of the significant issues in the industry is lacking an accurate way to identify the best crop that can be grown with the available soil fertility in a particular land, also If the crops are not feed the required amount of water, the crop yields are adversely affected, by over-irrigation or under-irrigation. Another aspect is identification of the plant diseases which is the key to preventing the losses in the yield and quantity of the agricultural product.

Crop produce prediction: This module predicts the produce of the crop on basis of area and location. There are 15 types of crops whose data is collected from 1966 to 2017 almost all the district of the India. In this module linear regression is model is used to predict the produce of crop on basis of area and location provided.

Fertilizer and Irrigation prediction: This module predict the various amount of fertilizer and amount of water to be used in irrigation. There are 13 crops whose data is collected from 1966 to 2017 almost all district of India. In this module after analysis of various model, the random forest module is used to predict both the fertilizer and irrigation

Plant Disease Prediction: This module aims to predict the plant disease on basis of the image of leaf. The proposed algorithm is to optimize the information from the resources available to us for the betterment of the result without any complexity. The neural network used for classification is the Dense Convolution Neural Network (DCNN). In this project, a pre-trained neural network model (densenet-121) has been used to classify the 29 different diseases for 7 plants (potato, tomato, corn, bell pepper, grape, apple and cherry). In this project, the original image is converted to HSV colour, greyscale colour and lab colour form and then the masked image is generated by thresholding and given to the proposed model for training and classification

In this project a Web application has been developed which consist of the above module. The tech stack used in this project are mentioned below:

- **Frontend:** Angular
- **Backend:** Nodejs, python
- **Libraries:** plotlyjs, scikitlearn, Pandas, Numpy.

ACKNOWLEDGEMENT

I express my deepest sense of gratitude to my supervisor Dr.R.Geetha Ramani, Faculty of Department of Information Science and Technology, College of Engineering, Guindy for her valuable guidance, inspiration and constructive suggestions throughout the period of project work. Moreover, her optimistic attitude, guidance and appreciation was such as to give impact us to our own thoughts and understandings, making us believe that all that accomplish was my own efforts for which I will ever remain obliged to her.

I deeply express my sincere thanks to DR.S. SRIDHAR, Professor and Head of the Department, Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai for extending support.

I would like to express my sincere thanks to the project committee members, DR.SASWATI MUKHERJEE, Professor, DR.M.VIJAYALAKSHMI, Associate Professor, DR.E.UMA, Assistant Professor, Department of Information Science and Technology, Anna University, Chennai for giving their valuable suggestions, encouragement and constant motivation throughout the duration of my project.

Thank you all for your support.

Samundar Singh (2019272033)

MCA (Self-Supporting)

TABLE OF CONTENTS

CHAPTER 1
1. INTRODUCTION

1.1 Crop related pattern (Fertilizer and irrigation):

In India there is need of advances in agriculture sector so that it meets the needs of people. agriculture is considered as the main and the foremost culture practiced in India. Nowadays, modern people don’t have awareness about the cultivation of the crops in a right time and at a right place Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. Crop yield prediction is an important agricultural problem. Each and Every farmer is always trying to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analysing farmer's previous experience on a particular crop. The Agricultural yield is primarily depending on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. The fertilizer and water irrigation plays a major role in the crop production.

1.2 Crop Disease:

Disease detection involves the steps like image acquisition, image pre-processing, image segmentation, feature extraction and classification. Disease on plants results in large reduction in both the standard and the quantity of agricultural products. The study of diseases asks the studies of visually observable patterns on the plants. In most of the cases, the symptoms of a plant disease are seen on the leaves and stem. The plant leaf for the detection of disease is taken into account which shows the disease symptoms. The disease of plant is also highly responsible for the produce of the crop.

1.3 Problem Statement:

The decrease in the production of crop due to lack of knowledge of correct amount of fertilizer that is needed to be used and proper arrangement of the irrigation facility.

Also due to lack of knowledge of plant disease the correct precaution is also not taken. There are possibilities that a particular crop may have more than one type of disease and sometimes the farmers missed this which result in low production of crops although he has taken the correct measure.

1.4 Objective:

The objective of this challenge is to build a machine learning algorithm to correctly classify if a crop is healthy, has stem rust, or has leaf rust with more accuracy.

The machine learning algorithm give accurate early stage prediction about dieses occurring in the crop so that mortality rate of crop yields can be improved. Disease development depends on some conditions like crop susceptibility to disease, viability of pathogen and environment favourability for that crop yield. After detecting the machine learning algorithm will suggest the best possible way to deal with the problem.

1.5 Project Overview

This project is basically dealing with the disease prediction and the crop pattern prediction system. The crop pattern is predicted on basis of the location and crop type. The user has to mentioned his state and district in order to find the amount of fertilizer and irrigation estimation. For fertilizer and irrigation prediction the harvesting area and location detail is taken from user and using Random forest model the output is predicted.

For plant disease prediction the image of leaf is taken from the user and using deep learning the disease is predicted. After the disease prediction It also suggest the remedy or cure for the predicted disease.

1.6 Organization of Report:

Chapter 2: This chapter explains about the literature survey made on the existing system, analysing the problem statements and issues with the existing system.

Chapter 3: This chapter consists System design of the project with its preliminary design and descriptive details about the modules.

Chapter 4: This chapter consists Algorithm and pseudocode related to the models with their outcomes.

Chapter 5: This chapter consists Result and Analysis of the model and generation of summary overview.

CHAPTER 2

2. LITERATURE SURVEY/RELATED WORK

This Chapter explains about the literature survey made on the existing system, analysing the problem statements and issues with the existing system and proposed objectives for the new system. Role of Crop related pattern in produce and effect of disease in crop produce analysis has been conducted by may authors. Some of them are summarised here

2.1 Crop production related pattern.

In recent years, machine learning techniques have been applied in agriculture domain by various researchers. A review on the implementation of different ML techniques in the field of crop yield prediction from soil analysis from past few years is presented as under.

Bhuyar [1] proposed an approach where different classification algorithms such as J48, Naïve Bayes, and Random forest algorithm were applied to soil data set to predict its fertility. J48 algorithm gave better result with an accuracy of 98.17% than other algorithms. This analysis is done only to soil of a particular location.

Rajeshwari and Arunesh [2] performed a comparative analysis of ML algorithms i.e. Naive Bayes, JRIP and J48 for prediction of soil types. The experiments were performed on soil data consisting of 110 samples using data analytics tool R. The experimental results predicted that JRIP algorithm performed better as it gave highest accuracy of 98.18% with kappa statistic of approximate 1.0. The problem with this analysis is that the dataset is taken is very small.

Awasthi and Bansal [3] performed comparative study on two data mining techniques namely Artificial Neural Network and Support Vector Machine with the help of data analytics tool R on soil. ANN was implemented with 7 hidden nodes and this model trained for 73073 steps. It predicted accuracy of 55% with root mean square error is of 15. SVM implemented with Radial basis kernel and it achieved much better results with 74% of accuracy. This model is not ideal since the accuracy is very low

Singh et al.[4] implemented different machine learning techniques in order to predict the category of the rice crop yield based on macro-nutrients and micro-nutrients status in dataset. ANN, Naïve Bayes and KNN are applied on soil data. Decision Tree Classifier and Naïve Bayes Classifier are found to be better models for classifying the soils into categories and in the prediction of yield on the basis of Nutrient status in the soil.

Rajesh Kannan Mega lingam [5] presented IOT sensor unit which is part of the rover system consists of moisture sensor and a temperature sensor, for detection of the moisture and temperature respectively in close proximity of plants.

2.2 Plant Disease Literature Survey

S. S. Sannakki and V. S. Raj purohit, [7], proposed a “Classification of Pomegranate Diseases Based on Back Propagation Neural Network” which mainly works on the method of Segment the defected area and color and texture are used as the features. Here they used neural network classifier for the classification. The main advantage is it Converts to L*a*b to extract chromaticity layers of the image and Categorisation is found to be 97.30% accurate. The main disadvantage is that it is used only for the limited crops

P. R. Rothe and R. V. Kshirsagar [8] introduced a” Cotton Leaf Disease Identification using Pattern Recognition Techniques” which Uses snake segmentation, here Hu’s moments are used as distinctive attribute. Active contour model used to limit the vitality inside the infection spot, BPNN classifier tackles the numerous class problems. The average classification is found to be 85.52%. the problem with this model is the background noise.

Godliver Owomugisha, John A. Quinn, Ernest Mwebaze and James Lwasa, proposed” [10] Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease “Color histograms are extracted and transformed from RGB to HSV, RGB to L*a*b.Peak components are used to create max tree, five shape attributes are used and area under the curve analysis is used for classification. They used nearest neighbors, Decision tree, random forest, extremely randomized tree, Naïve bayes and SV classifier. In seven classifiers extremely, randomized trees yield a very high score, provide real time information provide flexibility to the application.

Aakanksha Rastogi, Ritika Arora and Shanu Sharma [11] Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic”. K-means clustering used to segment the defected area; GLCM is used for the extraction of texture features, Fuzzy logic is used for disease grading. They used artificial neural network (ANN) as a classifier which mainly helps to check the severity of the diseased leaf.

Uan Tian, Chunjiang Zhao, Shenglian Lu and Xinyu Guo [12] SVM-based Multiple Classifier System for Recognition of Wheat Leaf Diseases,” Color features are represented in RGB to HIS, by using GLCM, seven invariant moment are taken as shape parameter. They used SVM classifier which has MCS, used for detecting disease in wheat plant offline.

Aswin Vellaichamy , Akshay Swaminathan , C Varun and Dr. Kalaivani S [13] performed implemented the HSV technique to find the crop disease using leaf using densenet-121 and Achieved Accuracy of 97%

CHAPTER 3

SYSTEM DESIGN

This module consists system design of the project with preliminary design such as overall Architecture diagram and process flow diagram which tells about the modules integration in the project.

System Architecture

The proposed work of the system architecture is shown in Figure 3.1. The proposed system works on Crop Data Set. The target is to get data of the produce, fertilizer, irrigation and disease of the crop.

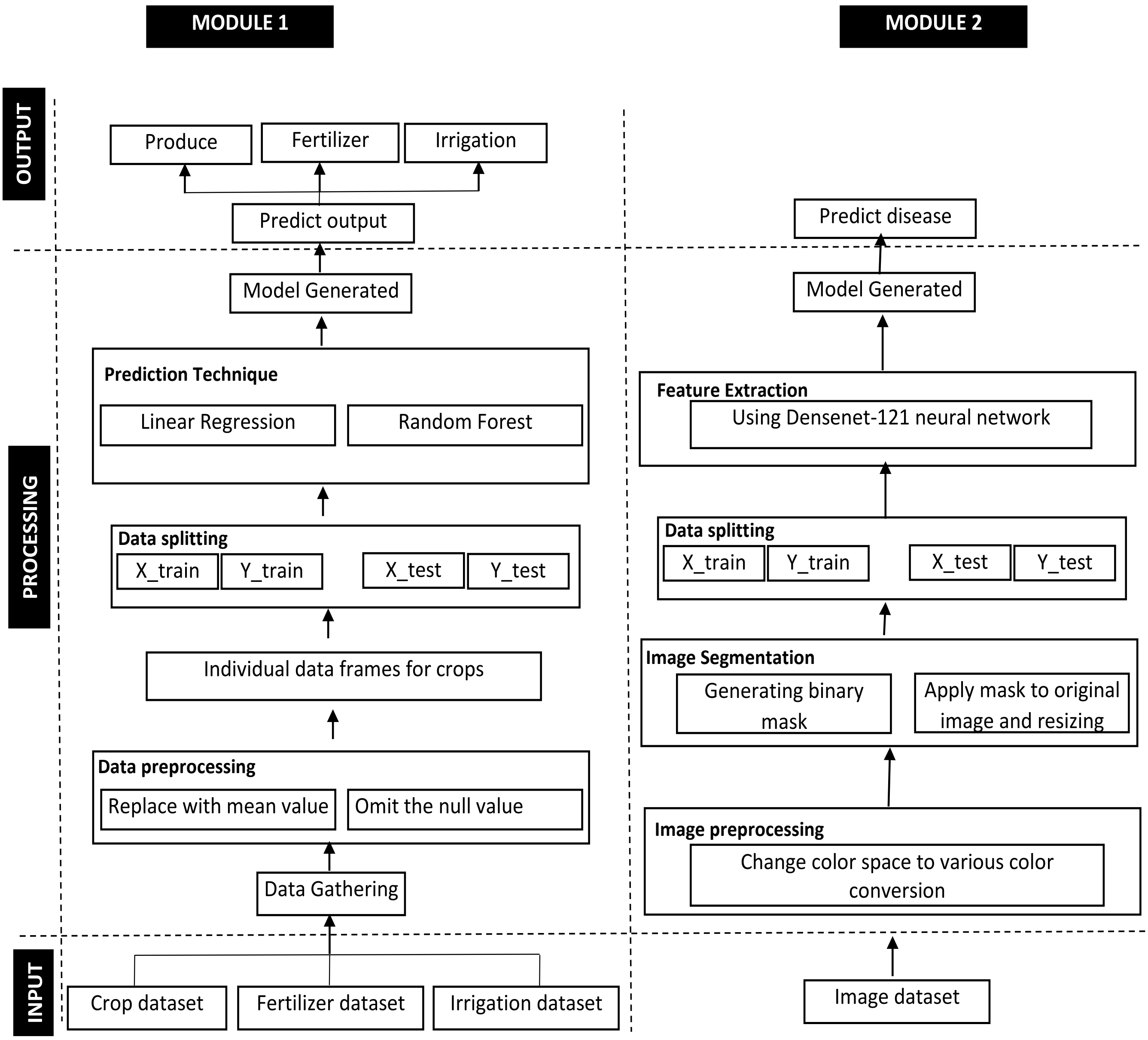


Fig 3.1 System Architecture

3.1 Crop Pattern Prediction

3.1.1 Data Gathering

There are three datasets are used in proposed model these datasets consists of the data regarding the fertilizers, irrigation and crop produce of all district in India. These datasets are collected from year 1966 – 2017. Total 14 type of Commercial crops as well as non-commercial crops. the dataset is collected from ICRISAT website. Further detail of these datasets are given below.

- The produce dataset contains 79 attributes and 16047 rows. These 79 attributes define the basic features of the data which is being forecasted like area, yield and produce. These attributes are divided into area, produce and yield of different crops. In this dataset the target value is the produce of different crop. This produce is dependent on the area.
- The fertilizer dataset contains 20 attributes and 16146 rows. These 20 attributes further divided into 3 main categories of fertilizer like Nitrogen, Phosphate, potash which is used in all district of India in between year 1966 – 2017. The target data that need to be predicted is the composition of Nitrogen, Phosphate, potash on basis of the location (various District of India).
- The irrigation dataset contains 24 attribute and 14506 columns. These attributes define the irrigated area of all the 14 crops that is mentioned in above dataset. The target data is the irrigation amount for different crop this is dependent on the location (various District of India).

At last the data set is divided as training and test data set.

3.1.2 Data Pre-processing

The pre-processing of data is a method for preparing and adapting raw data to a model of learning. This is the first and significant step to construct a machine learning model. Real-world data generally contain noise, missing values and may not be used in an unusable format especially for machine learning models. Data pre-processing needs to be performed in order to purify data and adapt it to the machine learning model of a system which also makes a machine learning model more accurate and efficient. The first thing for data preprocessing is to collect the required data set, and then check the missing values once the data set is imported. Correcting missed values is necessary, or else the data would be difficult to access and maintain. This data is unclean and contain many errors as well as bugs. Correcting missed values is necessary, or else the data would be difficult to access and maintain. Then calculate the mean of the column containing missing values to rectify the missed values, and substitute it with the measured mean. Now, this data set can be used to train a machine learning algorithm to predict values.

3.1.3 Prediction Technique

3.1.3.1 Linear Regression

Linear regression algorithm tries to predict the results by plotting the graph between an independent variable and a dependent variable that are derived from the dataset. It is a general statistical analysis mechanism used to build machine learning models. The general equation for linear regression is

$$Z = a + b*E$$

Where, Z is the dependent variable and E is independent variable.

3.1.3.2 Random Forest

Random Forest Algorithm is used to incorporate predictions from multiple decision trees into a single model. This algorithm uses bagging mechanism to create a forest of decision trees. It the incorporates the predictions from multiple decision trees to give very accurate predictions. The Random Forest algorithm has two steps involved

a) Random forest formation. b) Predict by Random forest classifier generated.

3.1.4 Outcome:

After training the model on above mentioned data then 3 individual model is created whose output are mentioned below:

Output 1: Produce of the crop per hectare

Output 2: Nitrogen, Phosphate, potash amount in kg/hectare

Output 3: irrigation of water on per cubic meter per hectare

3.2 Crop Disease Detection

3.2.1 Image Acquisition

A dataset which consist of various plant disease is taken from Kaggle. This dataset contains 35779 images of size 256 x 256 from Plant Village which contains across 29 labels of diseases for 7 plants namely apple, tomato, corn, bell pepper, grape, cherry and potato. The images used for each category are mentioned in table 3.1 which gives the image count for each category under various crops.

Crop Name	Image category	Number of Images
Apple	Scab Black rot Healthy	630 621 1645
Cherry	Cedar apple rust Powdery mildew Healthy	275 1052 854
Corn	Cercosporin leaf spot Common rust	513 1192

	Northern Leaf Blight Healthy	985 1162
Grape	Black rot Esca (black measles) Leaf blight Healthy	1180 1384 1076 423
Pepper-bell	Bacterial spot Healthy	997 1478
Potato	Early Blight Late Blight Healthy	1000 1000 152
Tomato	Bacterial spot Early blight Late blight Healthy Leaf mold Septoria leaf spot Spider mites Yellow leaf curl virus Target spot	2127 1000 1909 1591 952 1771 1676 5357 1404

Table 3.1 Total no of images for each class of disease

3.2.2 Image Pre-Processing:

The image colour is converted from RGB to other Form. The R, G and B values are divided by 255 to change the range of values from 0-255 to 0-1

$$R' = R/255$$

$$G' = G/255$$

$$B' = B/255$$

$$C_{max} = \max(R', G', B')$$

$$C_{min} = \min(R', G', B')$$

$$\Delta = C_{max} - C_{min}$$

Hue calculation:

$$H = \begin{cases} 60^\circ \times \frac{G' - B'}{\Delta} \bmod 6, C_{max} = R' \\ 60^\circ \times \frac{B' - R'}{\Delta} + 2, C_{max} = G' \\ 60^\circ \times \frac{R' - G'}{\Delta} + 4, C_{max} = B' \end{cases}$$

Saturation calculation:

$$S = \begin{cases} 0 & , if\ C_{max} = 0 \\ \frac{\Delta}{C_{max}} & , if\ C_{max} \neq 0 \end{cases}$$

Value calculation:

$$V = C_{max}$$

Where, R, G and B represent values of red, green and blue in original image and H, S and V represent hue, saturation and value after processing the original image.

3.2.3 **Image Segmentation**

Generating binary mask via thresholding image is defined as:

$$g\,(x, y) = 1,\ T1 \leq f\,(x, y) < T2$$

$$0,\ otherwise$$

Where x and y represent the coordinal values of the pixels in image g and f. Thresholding is done where T1 is equal to 0.130 and T2 at 0.600 for H-values alone. These values are chosen after statistical analysis of images and their histograms. After this step there are still discrepancies in the image due to noise and shadows, which makes the binary map imperfect. To overcome this, the gaps is filled by morphological reconstruction. After this binary mask is created. The area of the binary mask is highlighted by a red line along with the identification of the centroid of the image.

The obtained binary mask is then used to make masked image by,

Masked image = binary mask. * Original image

Here "*" represents normal dot multiplication of corresponding values that are present in binary mask and original image These have been displayed the sample outputs of images obtained in each of the above-described steps below. These masked images are then resized to 64 to reduce time consumption for neural network training.

3.2.4 **Feature Extraction with DCNN**

Dense Net is a convolutional neural organization where each layer is associated with any non-subsequent layers that are more profound in the organization, that is, the principal layer is associated with the second, third, fourth etc., the subsequent layer is associated with the third, fourth, fifth, etc. This is done to empower greatest data stream between the layers of the organization. To protect the feed-forward nature, each layer acquires contributions from every one of the past layers and gives its own element guides to every one of the layers which will come after it. So, the 'i'th' layer has 'I' data sources and comprises of highlight guides of all its first convolutional blocks. Its own element maps are given to all the following 'I-I' layers. The number 121 in densenet-121 is computed as follows.

DenseNet-121: 5 + [2 x (6 + 12 + 24 + 16)]

5– Convolution and Pooling Layer

3– Transition Layers (6, 12, 24)

1– Classification Layer (16)

2– Dense Block (1 × 1 and 3 × 3conv)

3.2.5 **Outcome:**

As an output the model predict the type of disease related to the crop.

CHAPTER 4
ALGORITHM AND PSEUDO CODE

This section explains in detail the various modules in the system. Each module includes the input for the module, process flow for the module and output for the module in detail.

4.1 Crop Produce Prediction Using Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. This Algorithm give the accuracy of 78 to 93 percent ranging on different crop. Machine learning, more specifically the field of predictive modelling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explain ability. In applied machine learning we will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends. As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm. Linear Regression for Machine Learning more specifically the field of predictive modelling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explain ability. In applied machine learning we will borrow, reuse and steal It is both a statistical algorithm and a machine learning algorithm. He reason is because linear regression has been around for so long (more than 200 years). It has been studied from every possible angle and often each angle has a new and different name. Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Linear Regression Algorithm to find the crop produce

1: Start
2: import linear regression from sklearn
3: take variable lr = Linear Regression(normalize=True)
4: get the X train = train df(area)
5: Get the Y train = train df(produce)
6: Get the x test = test df. drop (produce, axis=1). copy()
7: fit in variable lr.fit (x train, Y train)
8: lr. predict = lr. predict (X test)
9: lr. accuracy = round (lr. score(X train,Y train) * 100,2)
10: print lr. accuracy

OUTPUT: Got Accuracy of 78 to 93 %. Depending on type of crop.

4.2 Crop fertilizer and irrigation Prediction using Random Forest

Random Forest Classifier- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and out putting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

Random Forest Algorithm to find the irrigation and fertilizer value

---

1: start
2: import RandomForestRegressor from sklearn.
3: take variable tree1,tree2 = RandomForestRegressor
4: split the irrigation dataframe according to the crop.
5: split the fertilizer dataframe according to the district.
6: get the x1train = train df (district name)
7: get the x2train = train df (district)
8: Get the y1 train = train df(fertilizer names)
9: Get the y2train = train df(crop irrigation data)
10: create the test data for both the data frame also using train_test_split feature
11: fit in variable tree1.fit(x1train, y1train)
12: fit in variable tree2.fit(x2train, y2train)
13: output1 = tree1.predict(x1 test)
14: output2 = tree2.predict(x2.test)
15: outpu1.accuracy = print round(tree1.score(X train,Y train) * 100,2)
16: outpu2.accuracy = print round(tree2.score(X train,Y train) * 100,2)

OUTPUT: Got Accuracy of 84 % to 93 %.in predicting irrigation amount, depending on type of crop.

4.3 Crop disease prediction on basis of Densenet -121

The problems arise with CNNs when they go deeper. This is because the path for information from the input layer until the output layer (and for the gradient in the opposite direction) becomes so big, that they can get vanished before reaching the other side. The dense net requires parameters than an equivalent traditional CNN, as there is no need to learn redundant feature maps, it combines the previous layer output with a future layer. In Dense net the layer is connected densely to all layer. Details of the DenseNet-121 is following: 5—convolution and pooling layers, 3—transition layers (6,12,24), 1—Classification layer (16) and 2—dense block (1 × 1 and 3 × 3 conv).
Generally, traditional CNNs calculate the output layers (*l*th) using a non-linear transformation H_l (.) to the output of the previous layer X_{l-1}

$$X_l = H_l(X_{l-1}).$$

DenseNets do not really sum up the layer output functionality maps with the inputs but concatenate them. DenseNet offers an easy communication model for improving information flow between layers: the *l*th layer receives inputs from the features of all previous levels: The equation is then again transformed into

$$X_l = H_l\left[(X_0, X_1, X_2, \ldots, X_{l-1})\right],$$

where [X_0, X_1, X_2,…, X_{l-1},] is a single tensor formed by the concatenation of the output maps of previous layers. Out of the functions, $H_l(.)$ represents a non-linear transformation function. This function consists of three major operations, batch normalization (BN), activation (ReLU) and pooling and convolution (CONV)

Dense net Algorithm to find the disease of plant.

1: start
2: image Acquisition.
3: Convert the RGB to HSV, lab colour and greyscale format.
4: Generate Binary Mask after filling gaps.
5: Apply mask to the original image and Resizing it.
6: import Densenet 121 from keras.application .
7: get xtrain = np.zeros((train.shape[0], IMAGE_SIZE, IMAGE_SIZE, 3))
8: Normalise the data by dividing it 255.
9: get ytrain=train['DiseaseID'].values.
10: ytrain = to_categorical(y_train, num_classes=15).
11: xtest=pick an image and normalize it and feed to model.
12: use IimageDataGenerator and fit the imane into it.
13: get Accuracy by model.fit_generator function

OUTPUT: Got Accuracy of 93.56%.

CHAPTER 5
RESULT AND ANALYSIS

This section includes the result and analysis for each model along with the summary and graphical analysis.

5.1 Crop pattern Model Analysis.

5.1.1. Dataset collection of Crop production

The produce dataset contains 79 attributes and 16149 rows. These 79 attributes define the basic features of the data which is being forecasted like area, yield and produce. These attributes are divided into area, produce and yield of different crops. In this dataset the target value is the produce of different crop. This produce is dependent on the area.

The figure 5.1 is referred to as the Produce dataset in the csv file.

```
In [4]: df1
```

Out[4]:

	Dist Code	Year	State Name	Dist Name	RICE AREA (1000 ha)	RICE PRODUCTION (1000 tons)	RICE YIELD (Kg per ha)	WHEAT AREA (1000 ha)	WHEAT PRODUCTION (1000 tons)	WHEAT YIELD (Kg per ha)	...	SUGARCANE YIELD (Kg per ha)	COTTON AREA (1000 ha)	COTTON PRODUCTION (1000 tons)	COTTC YIEL (Kg p h
0	1	1966	Chhattisgarh	Durg	548.00	185.00	337.59	44.00	20.00	454.55	...	1777.78	0.0	0.0	0
1	1	1967	Chhattisgarh	Durg	547.00	409.00	747.71	50.00	26.00	520.00	...	1500.00	0.0	0.0	0
2	1	1968	Chhattisgarh	Durg	556.30	468.00	841.27	53.70	30.00	558.66	...	1000.00	0.0	0.0	0
3	1	1969	Chhattisgarh	Durg	563.40	400.80	711.40	49.40	26.50	536.44	...	1900.00	0.0	0.0	0
4	1	1970	Chhattisgarh	Durg	571.60	473.60	828.55	44.20	29.00	656.11	...	2000.00	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
16141	917	2013	Jharkhand	Singhbhum	267.06	579.70	2170.67	1.53	1.85	1209.15	...	0.00	0.0	0.0	0
16142	917	2014	Jharkhand	Singhbhum	256.33	586.63	2288.57	5.36	6.65	1240.67	...	0.00	0.0	0.0	0
16143	917	2015	Jharkhand	Singhbhum	263.21	264.71	1005.70	1.99	1.82	914.57	...	0.00	0.0	0.0	0
16144	917	2016	Jharkhand	Singhbhum	224.05	319.01	1423.84	0.38	0.83	2167.98	...	0.00	0.0	0.0	0
16145	917	2017	Jharkhand	Singhbhum	386.91	669.97	1731.62	0.00	0.00	0.00	...	0.00	0.0	0.0	0

16146 rows × 79 columns

5.1.2 Dataset collection of Fertilizers

The fertilizer dataset contains 20 attributes and 16047 rows. These 20 attributes further divided into 3 main categories of fertilizer like Nitrogen, Phosphate, potash which is used in all district of India in between year 1966 – 2017. The target data that need to be predicted is the composition of Nitrogen, Phosphate, potash on basis of the location (various District of India).

The figure 5.2 is referred to as the fertilizer dataset in the csv file.

```
In [3]: fertilizerdataframe
Out[3]:
```

	Dist Code	Year	State Code	State Name	Dist Name	NITROGEN CONSUMPTION (tons)	NITROGEN SHARE IN NPK (Percent)	NITROGEN PER HA OF NCA (Kg per ha)	NITROGEN PER HA OF GCA (Kg per ha)	PHOSPHATE CONSUMPTION (tons)	PHOSPHATE SHARE IN NPK (Percent)	PHOSPHATE PER HA OF NCA (Kg per ha)	PHOSPHAT PER HA C GCA (Kg p h
0	1	1966	14	Chhattisgarh	Durg	1375	80.74	1.38	1.27	292	17.15	0.29	0.
1	1	1967	14	Chhattisgarh	Durg	1516	78.79	1.50	1.17	207	10.76	0.20	0.
2	1	1968	14	Chhattisgarh	Durg	3042	85.84	2.95	2.27	458	12.92	0.44	0.
3	1	1969	14	Chhattisgarh	Durg	4131	82.39	3.99	3.08	847	16.89	0.82	0.
4	1	1970	14	Chhattisgarh	Durg	4594	67.92	4.44	3.32	2024	29.92	1.96	1.
...	...	...	...	...	...	...	...	...	...	...	...	...	
16042	917	2013	15	Jharkhand	Singhbhum	3796	89.72	14.43	12.81	392	9.26	1.49	1.
16043	917	2014	15	Jharkhand	Singhbhum	2925	89.61	11.64	10.24	326	9.99	1.30	1.
16044	917	2015	15	Jharkhand	Singhbhum	3367	77.63	14.31	11.70	815	18.79	3.46	2.
16045	917	2016	15	Jharkhand	Singhbhum	2016	81.85	9.93	8.41	427	17.34	2.10	1.
16046	917	2017	15	Jharkhand	Singhbhum	2588	60.06	-1.00	-1.00	1600	37.13	-1.00	-1.

16047 rows × 20 columns

5.1.3 Dataset collection of Irrigation

The irrigation dataset contains 24 attribute and 14506 columns. These attributes define the irrigated area of all the 14 crops that is mentioned in above dataset. The target data is the irrigation amount for different crop this is dependent on the location (various District of India).

The figure 5.3 is referred to as the Irrigation dataset in the csv file.

```
In [5]: irrigationdataframe
Out[5]:
```

	Dist Code	Year	State Name	Dist Name	RICE IRRIGATED AREA (1000 ha)	WHEAT IRRIGATED AREA (1000 ha)	KHARIF SORGHUM IRRIGATED AREA (1000 ha)	RABI SORGHUM IRRIGATED AREA (1000 ha)	SORGHUM IRRIGATED AREA (1000 ha)	PEARL MILLET IRRIGATED AREA (1000 ha)	...	PIGEONPEA IRRIGATED AREA (1000 ha)	MINOR PULSES IRRIGATED AREA (1000 ha)	PULSES IRRIGATED AREA (1000 ha
0	1	1966	Chhattisgarh	Durg	73.30	0.50	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0
1	1	1967	Chhattisgarh	Durg	100.70	0.60	0.0	0.0	0.0	0.0	...	0.00	0.10	0.1
2	1	1968	Chhattisgarh	Durg	124.60	0.90	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0
3	1	1969	Chhattisgarh	Durg	130.80	0.80	0.0	0.0	0.0	0.0	...	0.00	0.10	0.1
4	1	1970	Chhattisgarh	Durg	131.30	1.00	0.0	0.0	0.0	0.0	...	0.00	0.10	0.1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
14501	917	2012	Jharkhand	Singhbhum	11.31	1.32	0.0	0.0	0.0	0.0	...	0.01	0.01	0.0
14502	917	2013	Jharkhand	Singhbhum	9.26	1.46	0.0	0.0	0.0	0.0	...	0.00	0.11	0.1
14503	917	2014	Jharkhand	Singhbhum	11.25	2.46	0.0	0.0	0.0	0.0	...	0.00	0.09	0.1
14504	917	2015	Jharkhand	Singhbhum	6.45	1.88	0.0	0.0	0.0	0.0	...	0.00	0.34	0.3
14505	917	2016	Jharkhand	Singhbhum	1.40	0.38	0.0	0.0	0.0	0.0	...	0.00	0.05	0.0

14506 rows × 24 columns

5.1.4 Data Pre-processing

Real-world data generally contain noise, missing values and may not be used in an unusable format especially for machine learning models. This is the first and significant step to construct a machine learning model. Data pre-processing needs to be performed in order to purify data and adapt it to the machine learning model of a system which also makes a machine learning model more accurate and efficient.

5.1.5 Data Exploration

The figure 5.4 , 5.5, 5.6 is showing the detailed information of dataset including attributes, their data type, null or not null and count. After having a look at the dataset, certain information about the data was explored

```
In [4]: cropdataframe.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16146 entries, 0 to 16145
Data columns (total 79 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Dist Code                                 16146 non-null  int64
 1   Year                                      16146 non-null  int64
 2   State Name                                16146 non-null  object
 3   Dist Name                                 16146 non-null  object
 4   RICE AREA (1000 ha)                       16146 non-null  float64
 5   RICE PRODUCTION (1000 tons)               16146 non-null  float64
 6   RICE YIELD (Kg per ha)                    16146 non-null  float64
 7   WHEAT AREA (1000 ha)                      16146 non-null  float64
 8   WHEAT PRODUCTION (1000 tons)              16146 non-null  float64
 9   WHEAT YIELD (Kg per ha)                   16146 non-null  float64
 10  KHARIF SORGHUM AREA (1000 ha)             16146 non-null  float64
 11  KHARIF SORGHUM PRODUCTION (1000 tons)     16146 non-null  float64
 12  KHARIF SORGHUM YIELD (Kg per ha)          16146 non-null  float64
 13  RABI SORGHUM AREA (1000 ha)               16146 non-null  float64
 14  RABI SORGHUM PRODUCTION (1000 tons)       16146 non-null  float64
 15  RABI SORGHUM YIELD (Kg per ha)            16146 non-null  float64
 16  SORGHUM AREA (1000 ha)                    16146 non-null  float64
 17  SORGHUM PRODUCTION (1000 tons)            16146 non-null  float64
 18  SORGHUM YIELD (Kg per ha)                 16146 non-null  float64
 19  PEARL MILLET AREA (1000 ha)               16146 non-null  float64
 23  MAIZE PRODUCTION (1000 tons)              16146 non-null  float64
 24  MAIZE YIELD (Kg per ha)                   16146 non-null  float64
 25  FINGER MILLET AREA (1000 ha)              16146 non-null  float64
 26  FINGER MILLET PRODUCTION (1000 tons)      16146 non-null  float64
 27  FINGER MILLET YIELD (Kg per ha)           16146 non-null  float64
 28  BARLEY AREA (1000 ha)                     16146 non-null  float64
 29  BARLEY PRODUCTION (1000 tons)             16146 non-null  float64
 30  BARLEY YIELD (Kg per ha)                  16146 non-null  float64
 31  CHICKPEA AREA (1000 ha)                   16146 non-null  float64
 32  CHICKPEA PRODUCTION (1000 tons)           16146 non-null  float64
 33  CHICKPEA YIELD (Kg per ha)                16146 non-null  float64
 34  PIGEONPEA AREA (1000 ha)                  16146 non-null  float64
 35  PIGEONPEA PRODUCTION (1000 tons)          16146 non-null  float64
 36  PIGEONPEA YIELD (Kg per ha)               16146 non-null  float64
 37  MINOR PULSES AREA (1000 ha)               16146 non-null  float64
 38  MINOR PULSES PRODUCTION (1000 tons)       16146 non-null  float64
 39  MINOR PULSES YIELD (Kg per ha)            16146 non-null  float64
 40  GROUNDNUT AREA (1000 ha)                  16146 non-null  float64
 41  GROUNDNUT PRODUCTION (1000 tons)          16146 non-null  float64
 42  GROUNDNUT YIELD (Kg per ha)               16146 non-null  float64
 43  SESAMUM AREA (1000 ha)                    16146 non-null  float64
 44  SESAMUM PRODUCTION (1000 tons)            16146 non-null  float64
 45  SESAMUM YIELD (Kg per ha)                 16146 non-null  float64
 46  RAPESEED AND MUSTARD AREA (1000 ha)       16146 non-null  float64
 47  RAPESEED AND MUSTARD PRODUCTION (1000 tons) 16146 non-null  float64
 48  RAPESEED AND MUSTARD YIELD (Kg per ha)    16146 non-null  float64
 49  SAFFLOWER AREA (1000 ha)                  16146 non-null  float64
 50  SAFFLOWER PRODUCTION (1000 tons)          16146 non-null  float64
 51  SAFFLOWER YIELD (Kg per ha)               16146 non-null  float64
 52  CASTOR AREA (1000 ha)                     16146 non-null  float64
 53  CASTOR PRODUCTION (1000 tons)             16146 non-null  float64
 54  CASTOR YIELD (Kg per ha)                  16146 non-null  float64
 55  LINSEED AREA (1000 ha)                    16146 non-null  float64
 56  LINSEED PRODUCTION (1000 tons)            16146 non-null  float64
 57  LINSEED YIELD (Kg per ha)                 16146 non-null  float64
 58  SUNFLOWER AREA (1000 ha)                  16146 non-null  float64
```

```
 59  SUNFLOWER PRODUCTION (1000 tons)            16146 non-null  float64
 60  SUNFLOWER YIELD (Kg per ha)                 16146 non-null  float64
 61  SOYABEAN AREA (1000 ha)                     16146 non-null  float64
 62  SOYABEAN PRODUCTION (1000 tons)             16146 non-null  float64
 63  SOYABEAN YIELD (Kg per ha)                  16146 non-null  float64
 64  OILSEEDS AREA (1000 ha)                     16146 non-null  float64
 65  OILSEEDS PRODUCTION (1000 tons)             16146 non-null  float64
 66  OILSEEDS YIELD (Kg per ha)                  16146 non-null  float64
 67  SUGARCANE AREA (1000 ha)                    16146 non-null  float64
 68  SUGARCANE PRODUCTION (1000 tons)            16146 non-null  float64
 69  SUGARCANE YIELD (Kg per ha)                 16146 non-null  float64
 70  COTTON AREA (1000 ha)                       16146 non-null  float64
 71  COTTON PRODUCTION (1000 tons)               16146 non-null  float64
 72  COTTON YIELD (Kg per ha)                    16146 non-null  float64
 73  FRUITS AREA (1000 ha)                       16146 non-null  float64
 74  VEGETABLES AREA (1000 ha)                   16146 non-null  float64
 75  FRUITS AND VEGETABLES AREA (1000 ha)        16146 non-null  float64
 76  POTATOES AREA (1000 ha)                     16146 non-null  float64
 77  ONION AREA (1000 ha)                        16146 non-null  float64
 78  FODDER AREA (1000 ha)                       16146 non-null  float64
dtypes: float64(75), int64(2), object(2)
memory usage: 9.7+ MB
```

figure 5.4 Dataset Information of Crop produce.

```
In [5]: irrigationdataframe.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14506 entries, 0 to 14505
Data columns (total 24 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Dist Code                                      14506 non-null  int64
 1   Year                                           14506 non-null  int64
 2   State Name                                     14506 non-null  object
 3   Dist Name                                      14506 non-null  object
 4   RICE IRRIGATED AREA (1000 ha)                  14506 non-null  float64
 5   WHEAT IRRIGATED AREA (1000 ha)                 14506 non-null  float64
 6   KHARIF SORGHUM IRRIGATED AREA (1000 ha)        14506 non-null  float64
 7   RABI SORGHUM IRRIGATED AREA (1000 ha)          14506 non-null  float64
 8   SORGHUM IRRIGATED AREA (1000 ha)               14506 non-null  float64
 9   PEARL MILLET IRRIGATED AREA (1000 ha)          14506 non-null  float64
 10  MAIZE IRRIGATED AREA (1000 ha)                 14506 non-null  float64
 11  FINGER MILLET IRRIGATED AREA (1000 ha)         14506 non-null  float64
 12  BARLEY IRRIGATED AREA (1000 ha)                14506 non-null  float64
 13  CHICKPEA IRRIGATED AREA (1000 ha)              14506 non-null  float64
 14  PIGEONPEA IRRIGATED AREA (1000 ha)             14506 non-null  float64
 15  MINOR PULSES IRRIGATED AREA (1000 ha)          14506 non-null  float64
 16  PULSES IRRIGATED AREA (1000 ha)                14506 non-null  float64
 17  GROUNDNUT IRRIGATED AREA (1000 ha)             14506 non-null  float64
 18  SESAMUM IRRIGATED AREA (1000 ha)               14506 non-null  float64
 19  LINSEED IRRIGATED AREA (1000 ha)               14506 non-null  float64
 20  SUGARCANE IRRIGATED AREA (1000 ha)             14506 non-null  float64
 21  COTTON IRRIGATED AREA (1000 ha)                14506 non-null  float64
 22  FRUITS AND VEGETABLES IRRIGATED AREA (1000 ha) 14506 non-null  float64
 23  FODDER IRRIGATED AREA (1000 ha)                14506 non-null  float64
dtypes: float64(20), int64(2), object(2)
memory usage: 2.7+ MB
```

figure 5.5 Dataset Information of Irrigation.

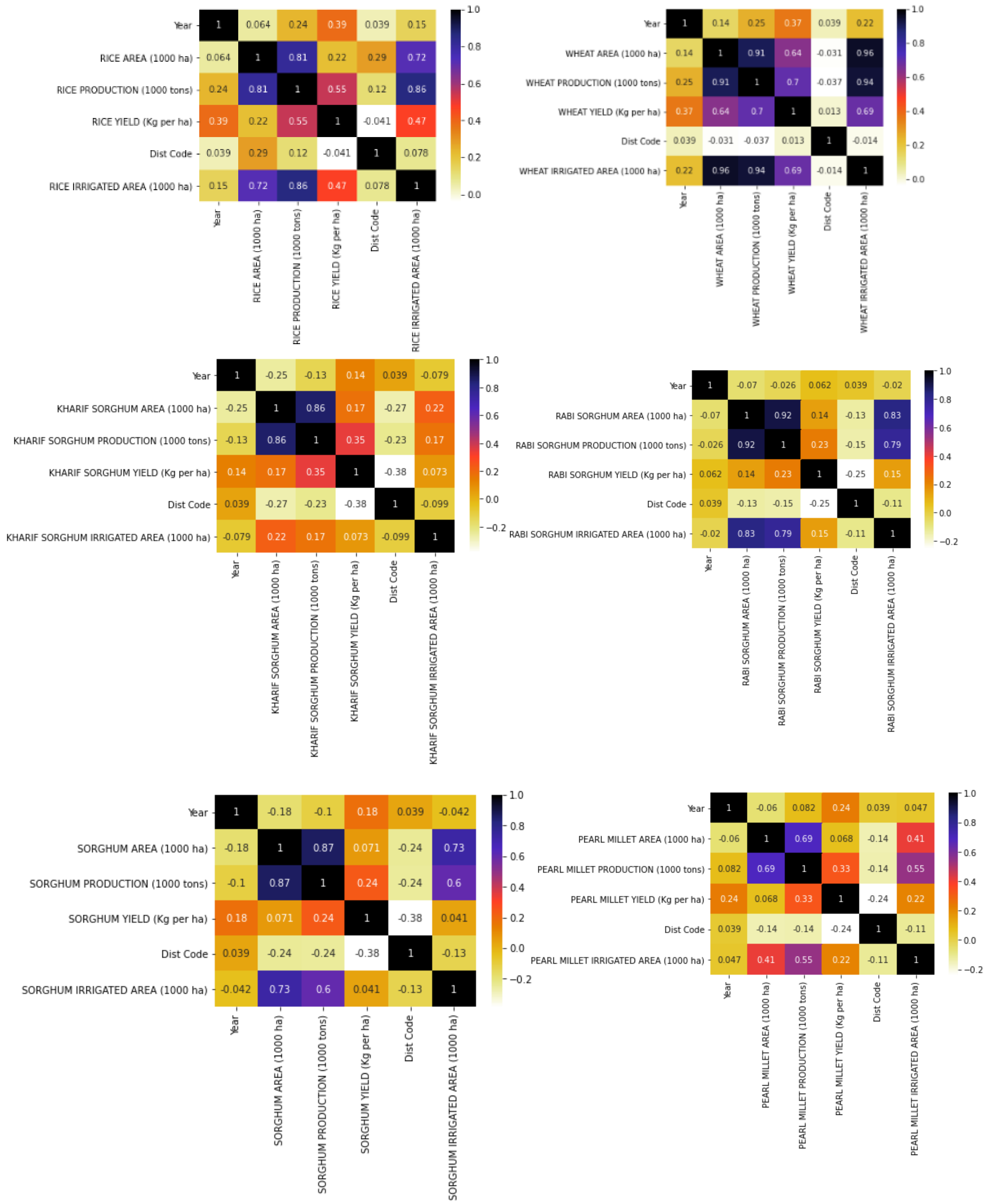
```
In [6]: fertilizerdataframe.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16047 entries, 0 to 16046
Data columns (total 20 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Dist Code                             16047 non-null  int64
 1   Year                                  16047 non-null  int64
 2   State Code                            16047 non-null  int64
 3   State Name                            16047 non-null  object
 4   Dist Name                             16047 non-null  object
 5   NITROGEN CONSUMPTION (tons)           16047 non-null  int64
 6   NITROGEN SHARE IN NPK (Percent)       16047 non-null  float64
 7   NITROGEN PER HA OF NCA (Kg per ha)    16047 non-null  float64
 8   NITROGEN PER HA OF GCA (Kg per ha)    16047 non-null  float64
 9   PHOSPHATE CONSUMPTION (tons)          16047 non-null  int64
 10  PHOSPHATE SHARE IN NPK (Percent)      16047 non-null  float64
 11  PHOSPHATE PER HA OF NCA (Kg per ha)   16047 non-null  float64
 12  PHOSPHATE PER HA OF GCA (Kg per ha)   16047 non-null  float64
 13  POTASH CONSUMPTION (tons)             16047 non-null  int64
 14  POTASH SHARE IN NPK (Percent)         16047 non-null  float64
 15  POTASH PER HA OF NCA (Kg per ha)      16047 non-null  float64
 16  POTASH PER HA OF GCA (Kg per ha)      16047 non-null  float64
 17  TOTAL CONSUMPTION (tons)              16047 non-null  int64
 18  TOTAL PER HA OF NCA (Kg per ha)       16047 non-null  float64
 19  TOTAL PER HA OF GCA (Kg per ha)       16047 non-null  float64
dtypes: float64(11), int64(7), object(2)
memory usage: 2.4+ MB
```

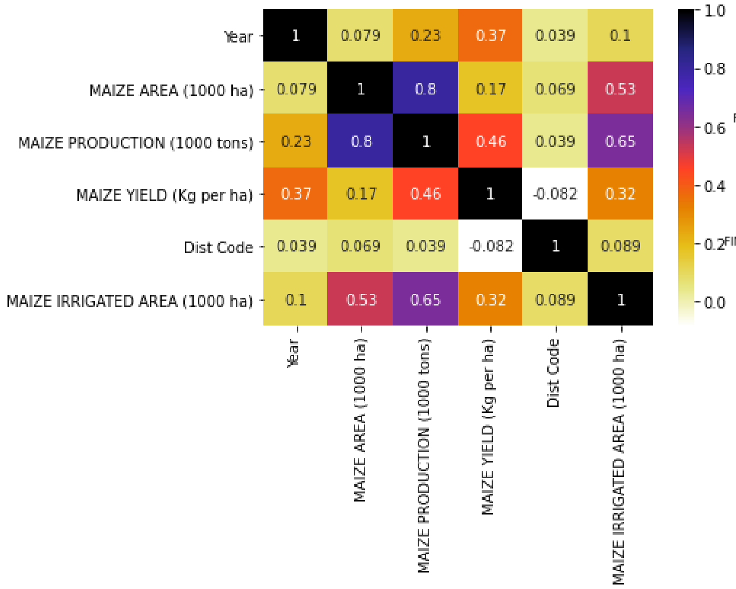
figure 5.6 Dataset Information of Fertilizer.

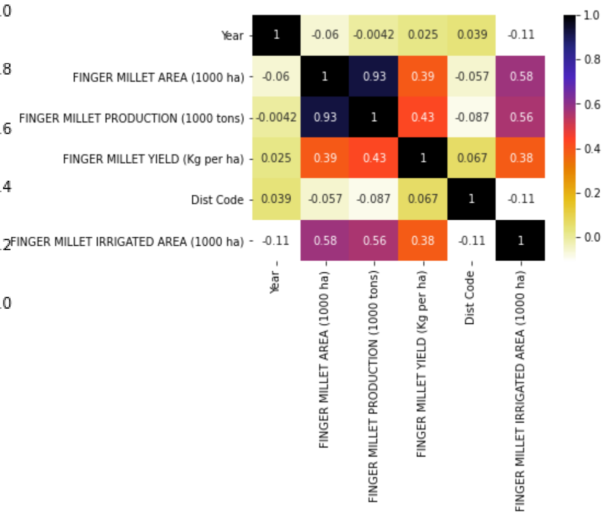
5.1.6 Feature Engineering

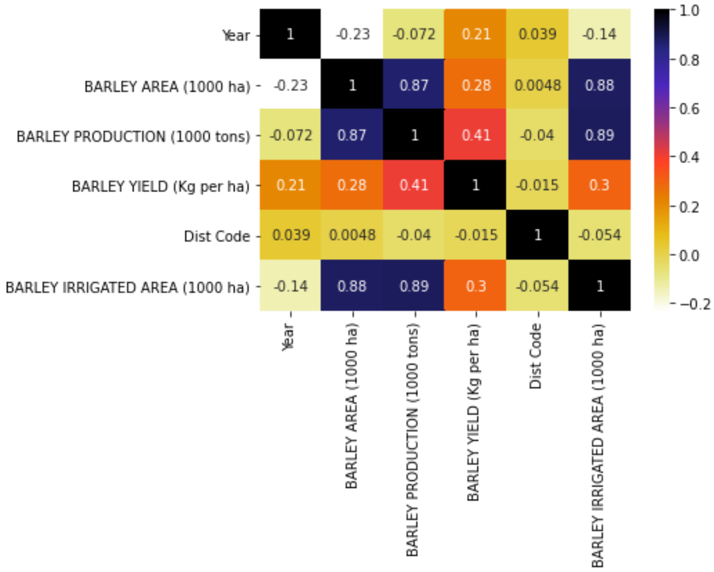
In the data exploration phase, some nuances in the dataset were discovered. So, in this phase, all of the nuances were resolved and made the data appropriate for modelling purpose. The data is splitted on basis of crops. After analysing the data we can find that the Area is strongly corelated to the production and the irrigation is also corelated to the production

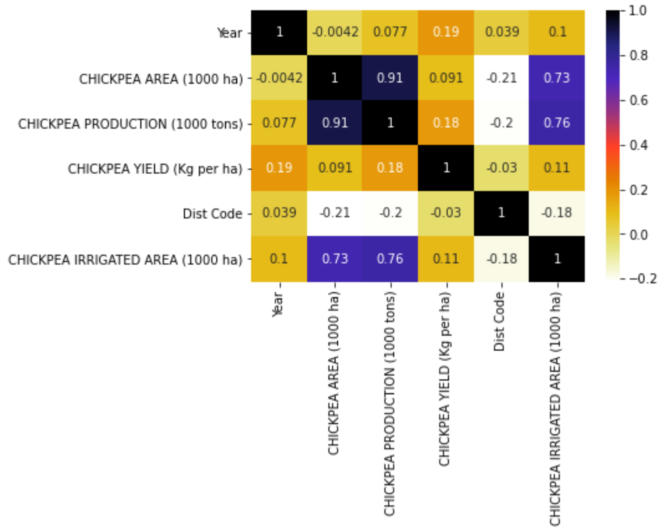
fig. 5.7 correlation between area, produce, yield, irrigation of each crop.

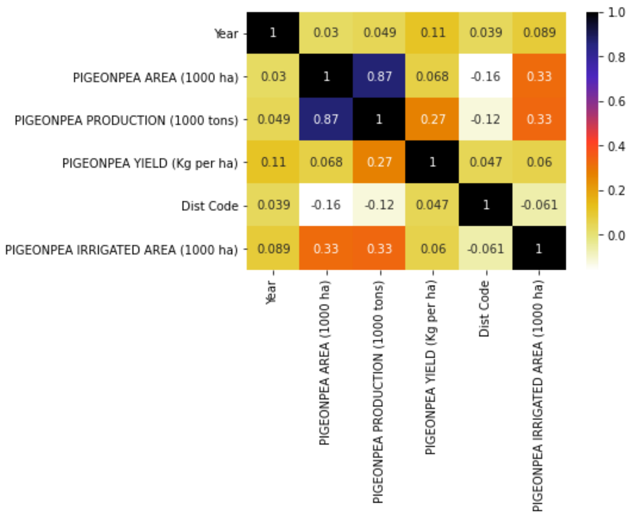


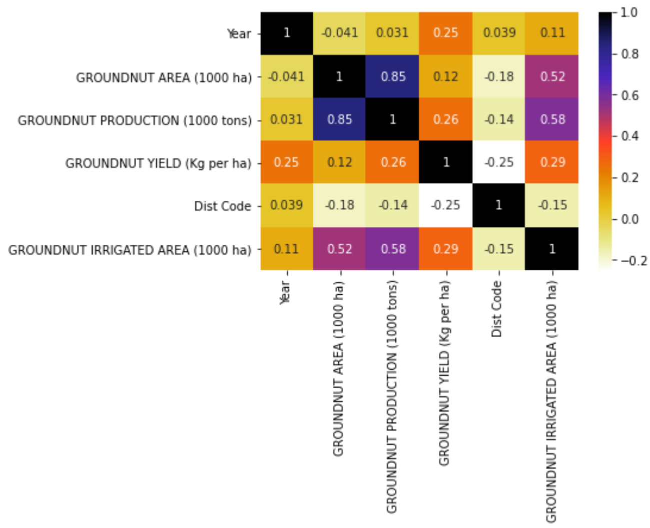

Year 1 0.079 0.23 0.37 0.039 0.1
MAIZE AREA (1000 ha) 0.079 1 0.8 0.17 0.069 0.53
MAIZE PRODUCTION (1000 tons) 0.23 0.8 1 0.46 0.039 0.65
MAIZE YIELD (Kg per ha) 0.37 0.17 0.46 1 -0.082 0.32
Dist Code 0.039 0.069 0.039 -0.082 1 0.089
MAIZE IRRIGATED AREA (1000 ha) 0.1 0.53 0.65 0.32 0.089 1
Year
MAIZE AREA (1000 ha)
MAIZE PRODUCTION (1000 tons)
MAIZE YIELD (Kg per ha)
Dist Code
MAIZE IRRIGATED AREA (1000 ha)
1.0
0.8
0.6
0.4
0.2
0.0

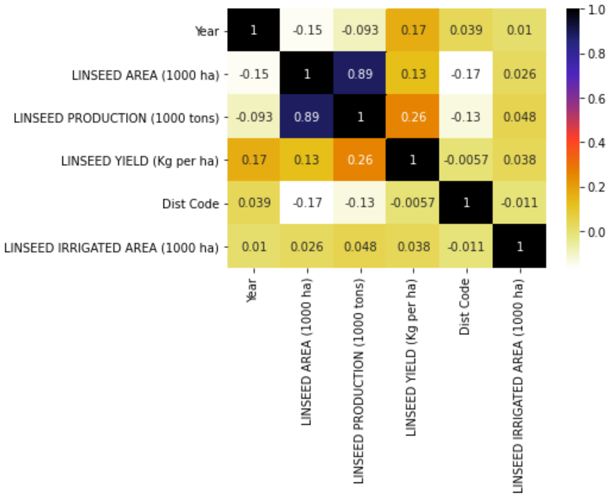

Year 1 -0.06 -0.0042 0.025 0.039 -0.11
FINGER MILLET AREA (1000 ha) -0.06 1 0.93 0.39 -0.057 0.58
FINGER MILLET PRODUCTION (1000 tons) -0.0042 0.93 1 0.43 -0.087 0.56
FINGER MILLET YIELD (Kg per ha) 0.025 0.39 0.43 1 0.067 0.38
Dist Code 0.039 -0.057 -0.087 0.067 1 -0.11
FINGER MILLET IRRIGATED AREA (1000 ha) -0.11 0.58 0.56 0.38 -0.11 1
Year
FINGER MILLET AREA (1000 ha)
FINGER MILLET PRODUCTION (1000 tons)
FINGER MILLET YIELD (Kg per ha)
Dist Code
FINGER MILLET IRRIGATED AREA (1000 ha)
1.0
0.8
0.6
0.4
0.2
0.0


Year 1 -0.23 -0.072 0.21 0.039 -0.14
BARLEY AREA (1000 ha) -0.23 1 0.87 0.28 0.0048 0.88
BARLEY PRODUCTION (1000 tons) -0.072 0.87 1 0.41 -0.04 0.89
BARLEY YIELD (Kg per ha) 0.21 0.28 0.41 1 -0.015 0.3
Dist Code 0.039 0.0048 -0.04 -0.015 1 -0.054
BARLEY IRRIGATED AREA (1000 ha) -0.14 0.88 0.89 0.3 -0.054 1
Year
BARLEY AREA (1000 ha)
BARLEY PRODUCTION (1000 tons)
BARLEY YIELD (Kg per ha)
Dist Code
BARLEY IRRIGATED AREA (1000 ha)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Year 1 -0.0042 0.077 0.19 0.039 0.1
CHICKPEA AREA (1000 ha) -0.0042 1 0.91 0.091 -0.21 0.73
CHICKPEA PRODUCTION (1000 tons) 0.077 0.91 1 0.18 -0.2 0.76
CHICKPEA YIELD (Kg per ha) 0.19 0.091 0.18 1 -0.03 0.11
Dist Code 0.039 -0.21 -0.2 -0.03 1 -0.18
CHICKPEA IRRIGATED AREA (1000 ha) 0.1 0.73 0.76 0.11 -0.18 1
Year
CHICKPEA AREA (1000 ha)
CHICKPEA PRODUCTION (1000 tons)
CHICKPEA YIELD (Kg per ha)
Dist Code
CHICKPEA IRRIGATED AREA (1000 ha)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Year 1 0.03 0.049 0.11 0.039 0.089
PIGEONPEA AREA (1000 ha) 0.03 1 0.87 0.068 -0.16 0.33
PIGEONPEA PRODUCTION (1000 tons) 0.049 0.87 1 0.27 -0.12 0.33
PIGEONPEA YIELD (Kg per ha) 0.11 0.068 0.27 1 0.047 0.06
Dist Code 0.039 -0.16 -0.12 0.047 1 -0.061
PIGEONPEA IRRIGATED AREA (1000 ha) 0.089 0.33 0.33 0.06 -0.061 1
Year
PIGEONPEA AREA (1000 ha)
PIGEONPEA PRODUCTION (1000 tons)
PIGEONPEA YIELD (Kg per ha)
Dist Code
PIGEONPEA IRRIGATED AREA (1000 ha)
1.0
0.8
0.6
0.4
0.2
0.0


Year 1 -0.041 0.031 0.25 0.039 0.11
GROUNDNUT AREA (1000 ha) -0.041 1 0.85 0.12 -0.18 0.52
GROUNDNUT PRODUCTION (1000 tons) 0.031 0.85 1 0.26 -0.14 0.58
GROUNDNUT YIELD (Kg per ha) 0.25 0.12 0.26 1 -0.25 0.29
Dist Code 0.039 -0.18 -0.14 -0.25 1 -0.15
GROUNDNUT IRRIGATED AREA (1000 ha) 0.11 0.52 0.58 0.29 -0.15 1
Year
GROUNDNUT AREA (1000 ha)
GROUNDNUT PRODUCTION (1000 tons)
GROUNDNUT YIELD (Kg per ha)
Dist Code
GROUNDNUT IRRIGATED AREA (1000 ha)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



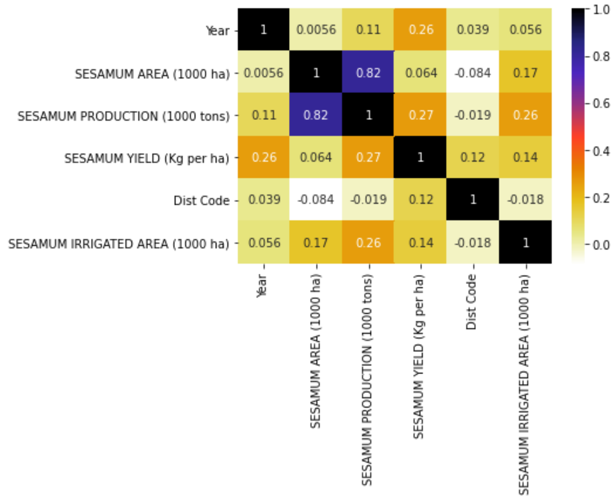
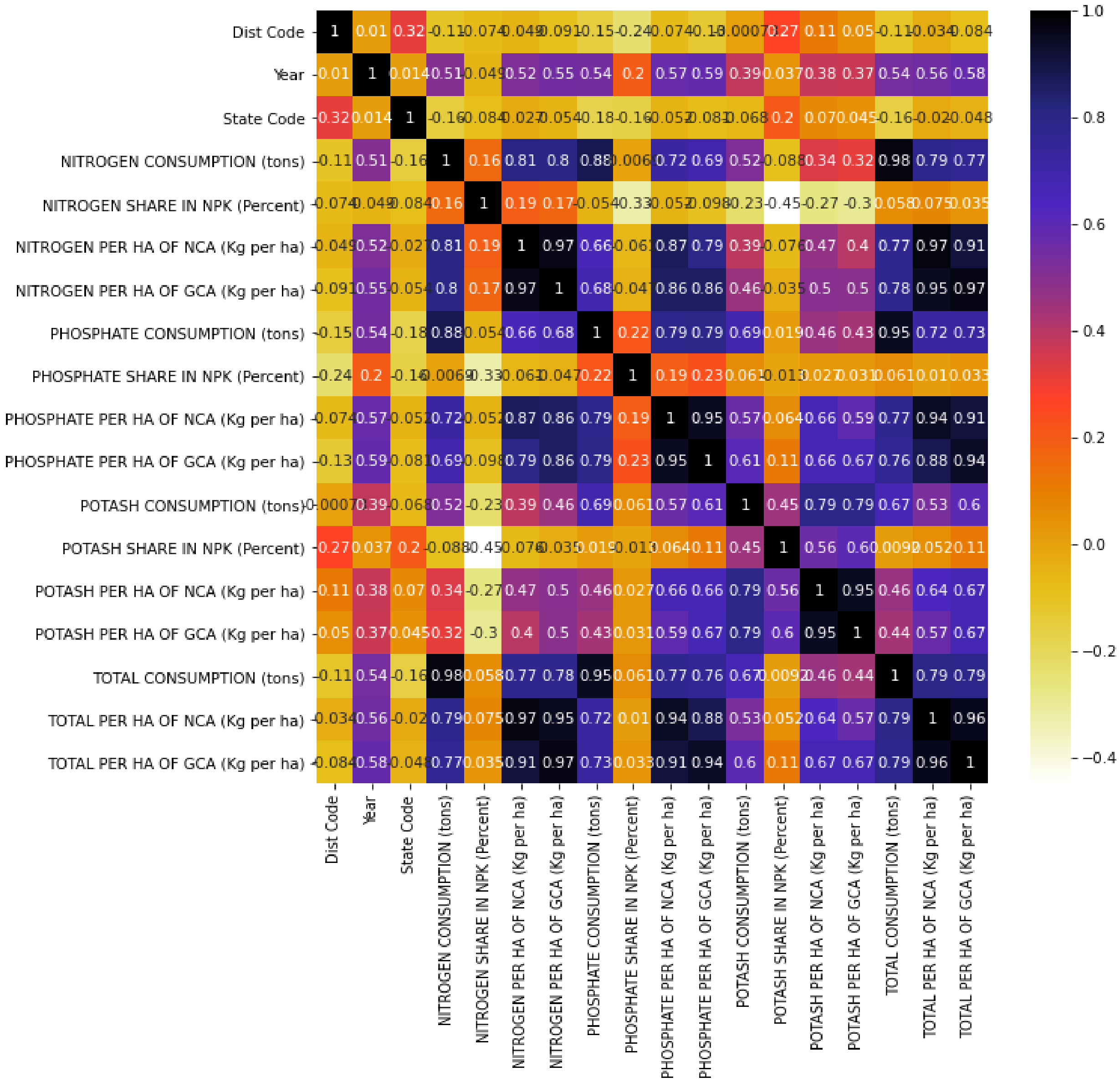


fig. 5.8 Correlation between Fertilizer are shown in below image.



For each and every crop a separate data frame is created which contain the area, yields, production along with the irrigation. It is founded that on average the produce is highly corelated to the area so for the target value of produce the area is fed as input to the model and the irrigation is also related to the produce as well as area so in next irrigation model the model is trained on basis the area ,produce as well as the district.

For the fertilizer data it is found that the district is more suitable on which the various fertilizer amount is predicted so the fertilizer are divided into three sub category such that nitrogen, potash and phosphate and we can predict these data on basis of the district.

5.1.7 Outlier Removal

Outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining. Records having outliers are removed from training dataset. Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results. It increases the error variance and reduces the power of statistical tests. If the outliers are non-randomly distributed, they can decrease normality. Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.

5.1.8 Evaluation Indicators

In order to evaluate the results and performance of different algorithms most popular measures are: MSE, RMSE, Accuracy Score and Cross Validation. These four metrics are explained in the following:

1. Mean Squared Error of an estimator measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

2. **Root Mean Square Error** is calculated as the standard deviation of the expectation mistakes. This mistake is the one which specifies how the data points are arranged about the regression line. Based on how the expectation mistakes are arranged around the regression line the RMSE value will vary.

3. **Model Score** is a inbuilt function in sklearn that calculate how efficient a ml model is working. It is used to compare different model involved in prediction.

4**. Cross Validation** is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data.

5.1.9 Evaluation Results

To achieve the experimental results, each regression model employs a 10-fold cross-validation to appraise the predictive accuracy. In cross-validation, the crops, fertilizer and irrigation dataset are randomly partitioned into 10 subsets and each subset has roughly equal size. 9 subsets of the dataset form the training data and the left-out subset are treated as the test data. The regression techniques form regression models using the training data and the test data is used for measuring the predictive accuracy. This process continues until every subset is treated as test data once.

5.1.9.1 Individual Crop Produce Accuracy Results

CROP NAME	MODEL ACCURACY
RICE	71.40 %
WHEAT	81.82 %
KHARIF SORGUM	71.81 %
RABI SORGUM	79.56 %
PEARL	78.26 %
MAIZE	43.63 %
FINGER	58.85 %
BARLEY	85.29 %
CHICKPEA	76.28 %
PIGEONOPEA	84.03 %
MINOR	75.92 %
GROUNDNUT	56.82 %
SESAMU	93.32 %
RAPESEEED	81.70 %
SAFFLOWER	89.93 %
CASTOR	86.84 %
LINSEED	62.31 %
SUNFLOWER	85.40 %
SOYABEAN	90.99 %
OILSEED	67.75 %
SUGARCANE	73.80 %
COTTON	64.14 %

5.1.9.2 Individual Crop Irrigation Accuracy Results

CROP NAME	MODEL ACCURACY
RICE	87.35 %
WHEAT	82.68 %
KHARIF SORGUM	76.36 %
RABI SORGUM	93.17 %
PEARL	89.28 %
MAIZE	86.17 %
FINGER	91.44 %
BARLEY	84.98 %
CHICKPEA	91.39 %
PIGEONOPEA	59.99 %
MINOR	68.02 %
GROUNDNUT	82.31 %
SESAMU	80.07 %
RAPESEEED	77.93 %
SAFFLOWER	73.98 %
CASTOR	45.11 %
LINSEED	92.28 %
SUNFLOWER	83.56 %
SOYABEAN	70.20 %
OILSEED	71.53 %

5.1.9.3 Fertilizer Accuracy Score

FERTILIZER NAME	MODEL ACCURACY
NITROGEN	80.77 %
PHOSPHORUS	74.27 %
POTASH	71.36 %

5.2 Crop Disease Model Analysis

5.2.1. Dataset collection of Crop Disease

A dataset which consist of various plant disease is taken from Kaggle. This dataset contains 35779 images of size 256 x 256 from Plant Village which contains across 29 labels of diseases for 7 plants namely apple, tomato, corn, bell pepper, grape, cherry and potato. The images used for each category are mentioned in table which gives the image count for each category under various crops.

5.2.2 Data Pre-processing

Real-world image data can contain background noise which decrease the accuracy of model since we trained our model on that background only.in data preprocessing first load the image from the pc to a particular data frame .this data frame can contain the path of each image.

After this process we have to labialized specific type of crop disease with a particular number so that it is easy to train the model.

In [11]:
```
img_info["labels_integer"] = None
index_labels_integer = img_info.columns.get_loc("labels_integer")
index_species = img_info.columns.get_loc("label")
k = 0
for i in range(len(img_info)):
    if i == 0:
        img_info.iloc[i, index_labels_integer] = k #here, k == 0
    if i > 0:
        if img_info.iloc[i-1, index_species] == img_info.iloc[i, index_species]:
            img_info.iloc[i, index_labels_integer] = k
        else:
            k += 1
            img_info.iloc[i, index_labels_integer] = k
img_info
```

Out[11]:

	image_path	label	labels_integer
0	PlantVillage/Pepper__bell___Bacterial_spot/002...	Pepper__bell___Bacterial_spot	0
1	PlantVillage/Pepper__bell___Bacterial_spot/006...	Pepper__bell___Bacterial_spot	0
2	PlantVillage/Pepper__bell___Bacterial_spot/00f...	Pepper__bell___Bacterial_spot	0
3	PlantVillage/Pepper__bell___Bacterial_spot/016...	Pepper__bell___Bacterial_spot	0
4	PlantVillage/Pepper__bell___Bacterial_spot/016...	Pepper__bell___Bacterial_spot	0
...	...	...	...
20631	PlantVillage/Tomato__Tomato_YellowLeaf__Curl_V...	Tomato__Tomato_YellowLeaf__Curl_Virus	14
20632	PlantVillage/Tomato__Tomato_YellowLeaf__Curl_V...	Tomato__Tomato_YellowLeaf__Curl_Virus	14
20633	PlantVillage/Tomato__Tomato_YellowLeaf__Curl_V...	Tomato__Tomato_YellowLeaf__Curl_Virus	14
20634	PlantVillage/Tomato__Tomato_YellowLeaf__Curl_V...	Tomato__Tomato_YellowLeaf__Curl_Virus	14
20635	PlantVillage/Tomato__Tomato_YellowLeaf__Curl_V...	Tomato__Tomato_YellowLeaf__Curl_Virus	14

fig 5.9 labelling the data

5.2.3 Removing Background Noise from image.

In order to remove the background noise first we select the threshold of colour including the upper limit and lower limit of green colour then we will import the image and convert it from RGB to HSV using CV2 library. Then we generate mask by providing the HSV-image with the threshold Value once the masked is generated then we apply the bitwise and operation to the image so that the thing which are common in the masked and the original image will only show as result.

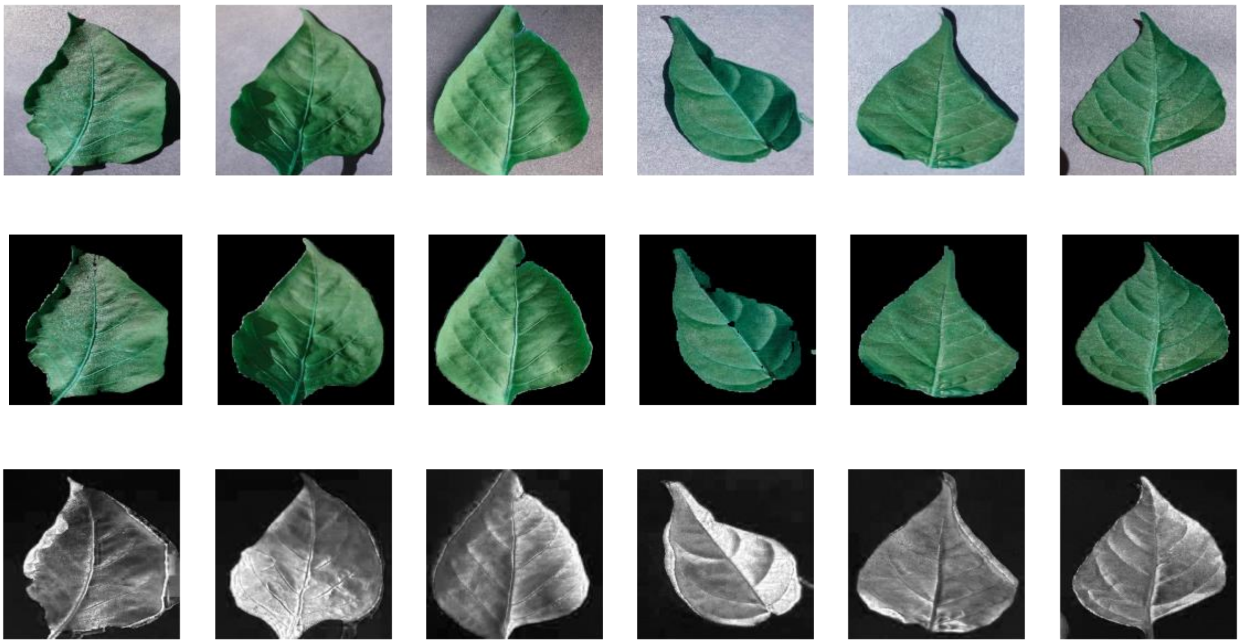


fig 5.10 image in which background noise is removed

5.2.4 FEATURE EXTRACTION WITH DCNN

The neural network used here is a densenet-121 a type of dense convolution neural network. Dense Net (Dense Convolutional Network) is a design that centres around making the deep learning networks go much more profound, and yet making them more effective to prepare, by utilizing more limited associations between the layers. Dense Net is a convolutional neural organization where each layer is associated with any non-subsequent layers that are more profound in the organization, that is, the principal layer is associated with the second, third, fourth etc., the subsequent layer is associated with the third, fourth, fifth, etc

The model summary is given below :

```
Layer (type)                 Output Shape              Param #
=================================================================
input_6 (InputLayer)         [(None, 64, 64, 3)]       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 64, 64, 3)         84
_________________________________________________________________
densenet121 (Functional)     (None, None, None, 1024)  7037504
_________________________________________________________________
global_average_pooling2d_2 ( (None, 1024)              0
_________________________________________________________________
batch_normalization_4 (Batch (None, 1024)              4096
_________________________________________________________________
dropout_4 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               262400
_________________________________________________________________
batch_normalization_5 (Batch (None, 256)               1024
_________________________________________________________________
dropout_5 (Dropout)          (None, 256)               0
_________________________________________________________________
root (Dense)                 (None, 15)                3855
=================================================================
Total params: 7,308,963
Trainable params: 7,222,755
Non-trainable params: 86,208
```

5.2.5 Evaluation Results

The overall accuracy of model is found to be 93.89 % after Iteration of 25 with a loss of 16 %

```
In [24]: final_loss, final_accuracy = model.evaluate(X_val, Y_val)
         print('Final Loss: {}, Final Accuracy: {}'.format(final_loss, final_accuracy))

65/65 [==============================] - 3s 33ms/step - loss: 0.1680 - accuracy: 0.9390
Final Loss: 0.16795822978019714, Final Accuracy: 0.9389534592628479
```

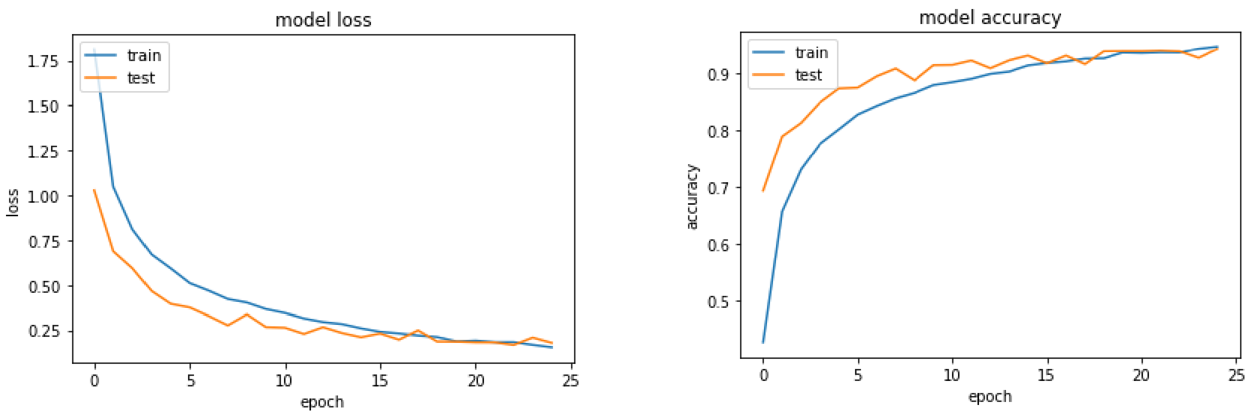


fig 5.11 shows the model accuracy as well as model loss on each iteration.

REFERENCE

1. Bhuyar, V. (2014). Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District. International Journal of Emerging Trends & Technology in Computer Science
2. Rajeshwari, V. & Arunesh, K. (2016). Analyzing Soil Data using Data Mining Classification Techniques. Indian Journal of Science and Technology.
3. Awasthi, N. & Bansal, A. (2017). Application of Data Mining Classification Techniques on Soil Data using R. International Journal of Advances in Electronics and Computer Science
4. Supriya, D. M. (2017). Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach. International Journal of Innovative Research in Computer and Communication Engineering,
5. Rajesh Kannan Megalingam (2020) Irrigation Monitoring and Prediction System Using Machine Learning.
6. Tummala, S., 2021. Classification of Multi Diseases in Apple Plant Leaves`
7. S. Sannakki and V. S. Rajpurohit," Classification of Pomegranate Diseases Based on Back Propagation Neural Network," International Research Journal of Engineering and Technology (IRJET), Vol2 Issue: 02 | May-2015
8. P. R. Rothe and R. V. Kshirsagar," Cotton Leaf Disease Identification using Pattern Recognition Techniques", International Conference on Pervasive Computing (ICPC),2015.
9. Aakanksha Rastogi, Ritika Arora and Shanu Sharma," Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic" 2nd International Conference on Signal Processing and Integrated Networks (SPIN)2015.
10. Godliver Owomugisha, John A. Quinn, Ernest Mwebaze and James Lwasa," Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease ", Preceding of the 1'st international conference on the use of mobile ICT in Africa ,2014.
11. Aakanksha Rastogi, Ritika Arora and Shanu Sharma, "Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic", *International Conference on Signal Processing and Integrated Networks (SPIN)*, 2015.
12. uan Tian, Chunjiang Zhao, Shenglian Lu and Xinyu Guo, "SVM-based Multiple Classifier System for Recognition of Wheat Leaf Diseases", *Proceedings of 2010 Conference on Dependable Computing (CDC'2010)*, November 20–22, 2010.
13. Aswin Vellaichamy S, Akshay Swaminathan , C Varun and Dr. Kalaivani S,(2021) multiple plant leaf disease classification using densenet-121 architecture
14. Kiran Moraye , Aruna Pavate, Suyog Nikam and Smit Thakkar (2021), Crop Yield Prediction Using Random Forest Algorithm for Major Cities in Maharashtra State