**Application of Machine Learning in Agriculture-:**

The world population is expected to reach 9.3 billion by the year 2050 from the current 7.3 billion. The Food and Agriculture Organization (FAO) predicts the growth of agriculture to be increased by 70% to serve the projected demand (“The Future of Agriculture”, 2016). There is a great need to increase the production of crops with limited available resources such as land, water, and fertilizers to cater to the needs of this increasing population. Agriculture, as classified by Lance Donny, has evolved from small scale labour intensive stage (until the 1920s) to industrial stage (from 1920 till 2010) with the improvement in heavy machinery and seed sciences, and it's paving its way towards the third stage or “Ag 3.0”, which involves making data-driven decisions using the info obtained from external sources (Lohr, 2015). Donny claims that data-driven decisions in agriculture provides higher productivity, practices sustainability and even helps to provide transparency to consumers wanting to know more about their food.

This blog discusses the increasing scope of data science in modern agriculture. It firstly gives an account on the need for data science in agriculture, followed by a discussion of opportunities, the issues that can arise in its implementation. After that, the impacts of data-driven agriculture are highlighted followed by impacts assessment and conclusion.

**Introduction-:**

Recently we have observed the emerging concept of smart farming that makes agriculture more efficient and effective with the help of high-precision algorithms. The mechanism that drives it is Machine Learning — the scientific field that gives machines the ability to learn without being strictly programmed. It has emerged together with big data technologies and high-performance computing to create new opportunities to unravel, quantify, and understand data intensive processes in agricultural operational environments.

Machine learning is everywhere throughout the whole growing and harvesting cycle. It begins with a seed being planted in the soil — from the soil preparation, seeds breeding and water feed measurement — and it ends when neural networks pick up the harvest determining the ripeness with the help of computer vision.

### **The Toxic Pesticides**

Though, many of us don't appreciate much, but a farmer's job is real test of endurance and determination. Once the seeds are sown, he works days and nights to make sure that he cultivates a good harvest at the end of season. A good harvest is ensured by several factors such as availability of water, soil fertility, protecting crops from rodents, timely use of pesticides & other useful chemicals and nature. While a lot of these factors are difficult to control for, the amount and frequency of pesticides is something the farmer can control.

Pesticides are also special, because they protect the crop with the right dosage. But, if you add more than required, they may spoil the entire harvest. A high level of pesticide can deem the crop dead / unsuitable for consumption among many outcomes. This data is based on crops harvested by various farmers at the end of harvest season. To simplify the problem, you can assume that all other factors like variations in farming techniques have been controlled for.

**Problem Statement-:**

To determine the outcome of the harvest season, i.e. whether the crop would be healthy (alive), damaged by pesticides or damaged by other reasons.

**Introduction to Dataset-:**

The Test dataset and Training dataset are given differently for the analysis and description are given below-:

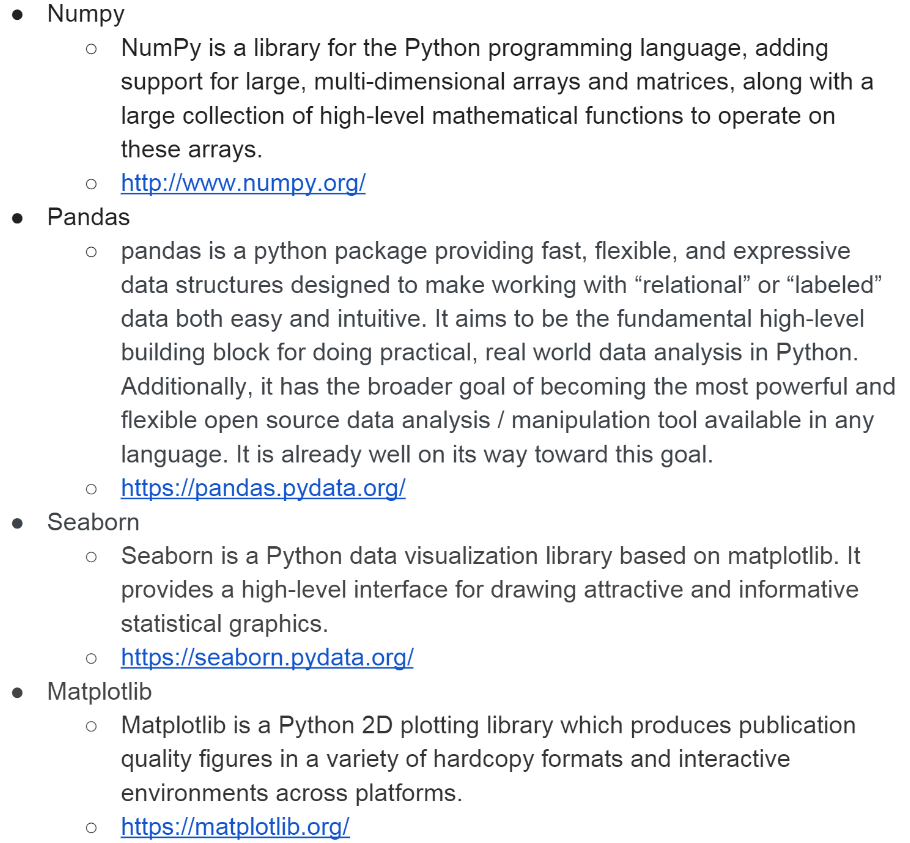


**EDA-:**

Exploratory Data Analysis (EDA) is a pre-processing step to explore the data. There are numerous methods and steps in performing EDA including visualization, wrangling, cleaning,etc, however, most of them are specific, focusing on either visualization or distribution , and are incomplete . Therefore, here, I will take you through step-by-step to understand, explore, and extract the information from the data to answer the questions or assumptions. There are no structured steps or methods to follow, however, this project will provide insights on EDA.

**Brief Introduction to the used libraries-:**

As discussed above, we are going to use the following libraries to perform different operations on the data.



**The outline for EDA are as follows-:**

**1.Import and analyze to know the data**

a.Importing required libraries

b.Importing Dataset

c.Knowledge of data(Meet & Greet with dataset)

**2. Data Analysis**

a.Check the data type

b. Check for the data characters mistakes

c. Check for missing values and replace them

d. Check for duplicate rows

e. Statistics summary

**3. EDA Concluding Remarks**

a.graphs between target variable and features

b.boxplot,and others

c.count of crops alive

d.pairplot, heatmap, correlation map

e.distibution and relationship distplot

**4. Pre-processing Pipeline**

a. Outliers and how to remove them

b. Checking and removing skewness

**5. Building Machine Learning Models**

a.Selection of best model

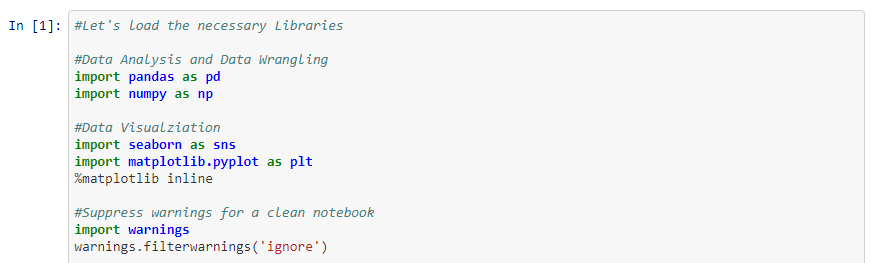
b.Loading test data and tarain model

c.Saving the result and model.

**6. Concluding Remarks**

**Import and analyze to know the data**

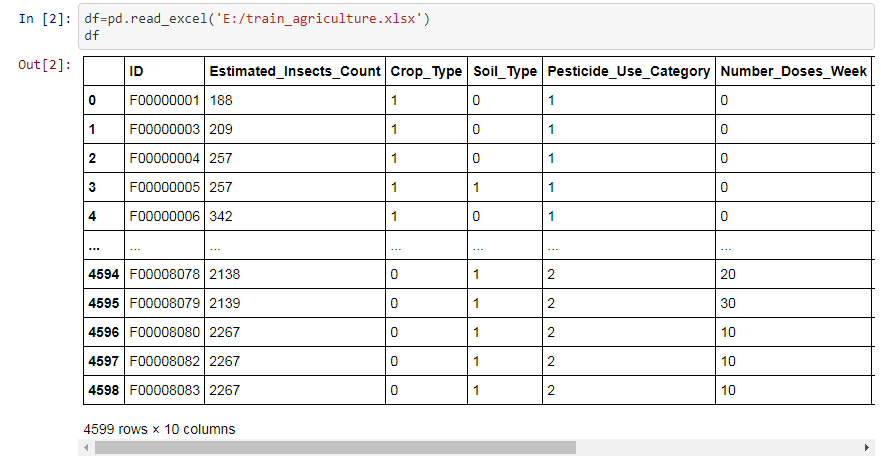
**1.Importing required libraries**

****

Required libraries are imported

**2.Loading dataset**

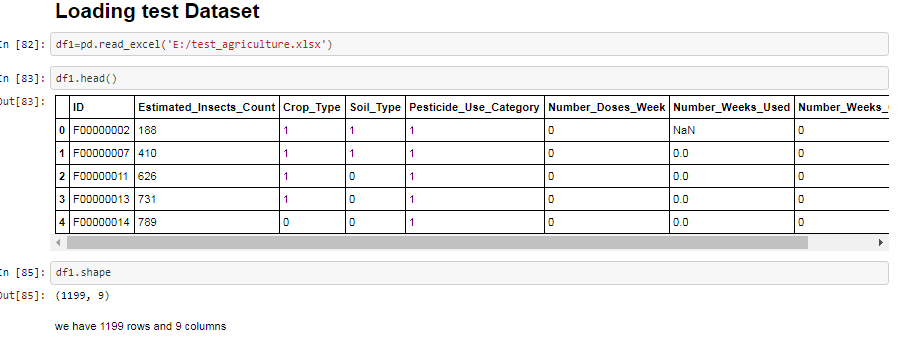
**a.Train Dataset**

****

We have loaded the train dataset for analysis. Later we will upload a test dataset to test the

Model

**b.Test Dataset**



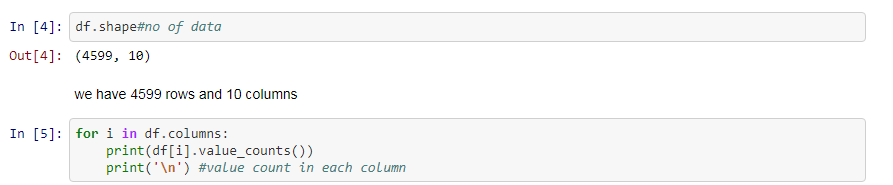
Here, we have 1199 rows and 9 columns.

.

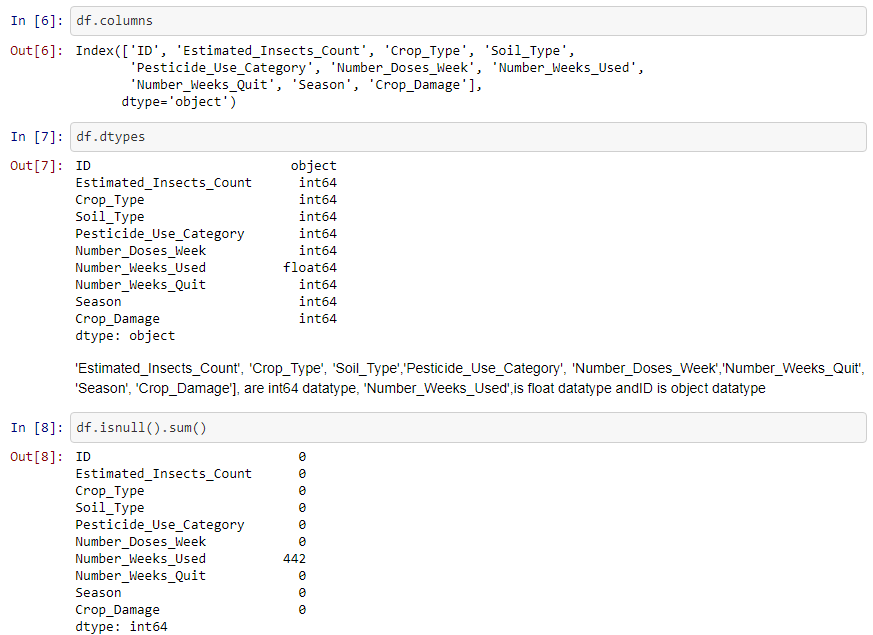
**3.Knowledge of data(Meet & Greet with dataset)**

In this section , we will explore the data and try to find out it’s characteristics Collecting Basic Information about the Data.It shows the number of rows, number of columns, data types information, Memory usage, number of null values in each column.

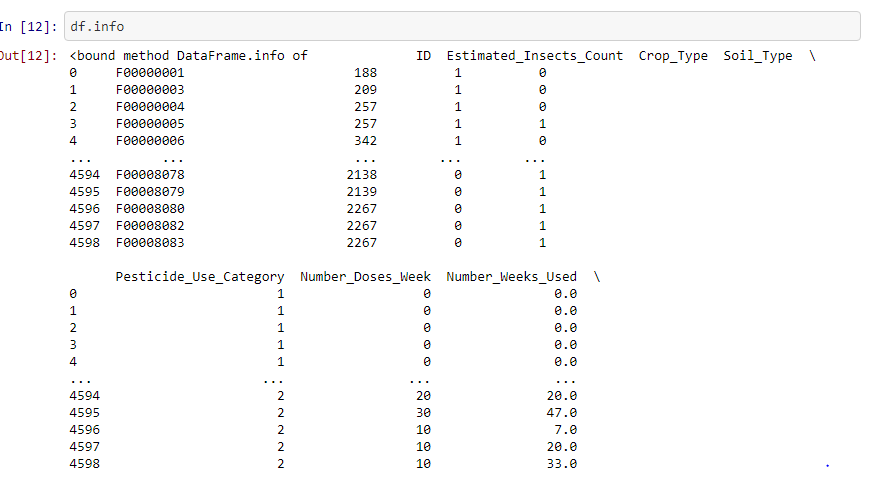
**a.Train Data**



It shows 4599 rows and 10 columns.

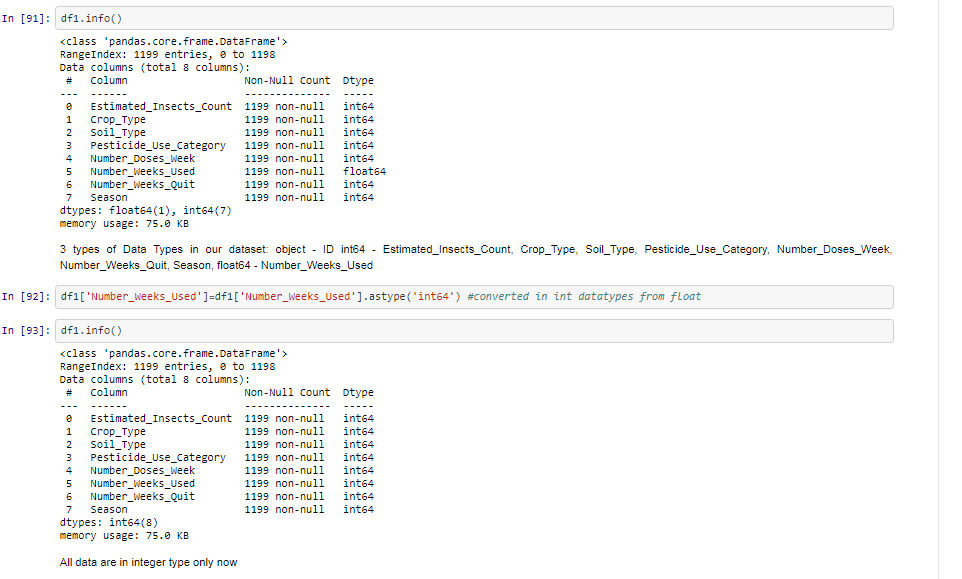


This shows the column name ,the datatypes and null values present in one of the column there were 442 null-values.



This gives the information of the entire data and memories allocation of data.

**b.Test Dataset**



**2.Data Analysis**

**a.Check the data type.**

**b.The variables types** are

Binary & Continuous:Estimated\_Insects\_Count ,Crop\_Type,Soil\_Type,Pesticide\_Use\_Category,Number\_Doses\_Week Number\_Weeks\_Used ,Number\_Weeks\_Quit,Season,Crop\_Damage

Categorical: ID

**c.Is the type of variable correctly classified by python ? Let’s get to know the data type.**

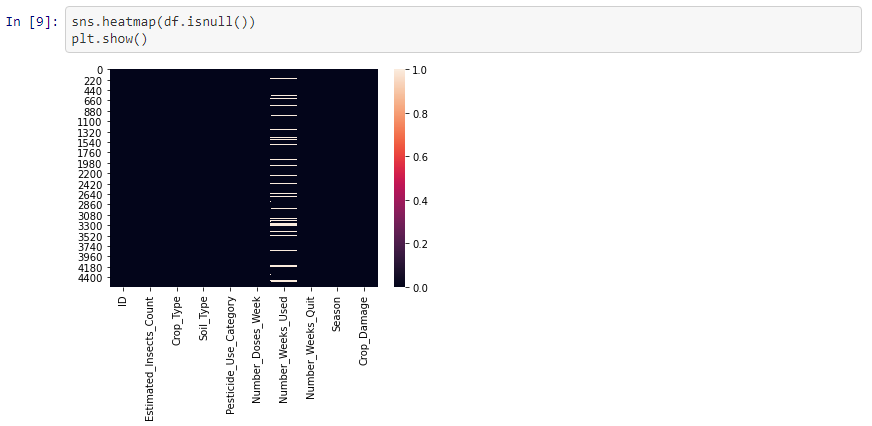
**d.Check for missing values and replace them.**

**e.Treating NaN**

We need to check the presence of the missing values and need to replace them with mean, median and mode of data accordingly. Sometimes we have null values in the form of 0 , so we need to convert them to NaN and then replace them accordingly. The missing values can be removed also but it should be less than 5 percent of the whole dataset.and visualize the missing values using Heatmap. The missing values are represented by the horizontal lines. This map provides an informative way of visualizing the missing values located in each column, and to see whether there is any correlation between missing values of different columns.

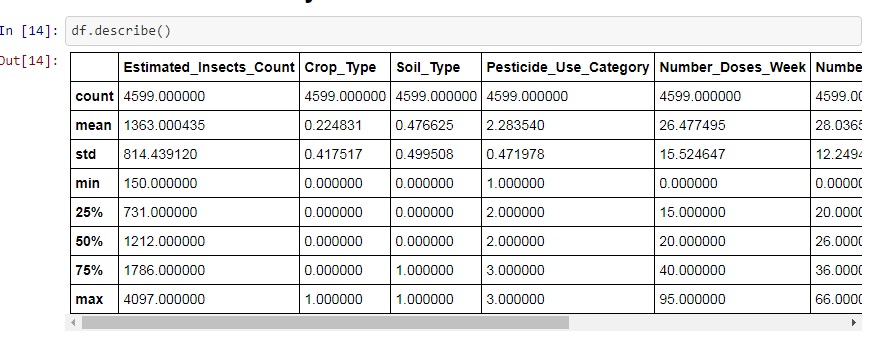
Changed datatype.





Heatmap of null values showing horizontally.

**f.Statistical Summary**



**observations-:**

1.Mean for all the columns is greater than the median so it means our data is positively skewed except for the Season columns as it is negatively skewed.

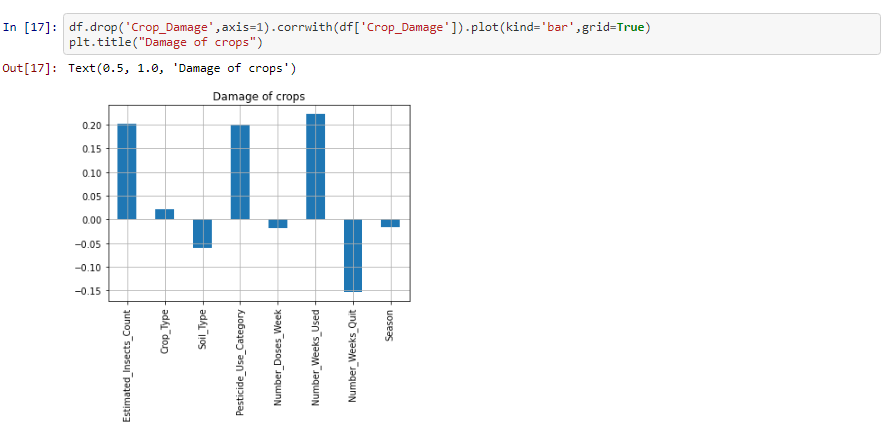
2.75% and 100% have large difference means Estimated\_Insects\_Count, Number\_Doses\_Week, Number\_Weeks\_Used, Number\_Weeks\_Quit have outliers present in the dataset.

3.Estimated\_Insects\_count have the highest standard deviation it means the data is spread throuth out and is not clustered around the mean.

**3. EDA Concluding Remarks**

**a.graphs between target variable and features**

**Correlation of features with Target variable**



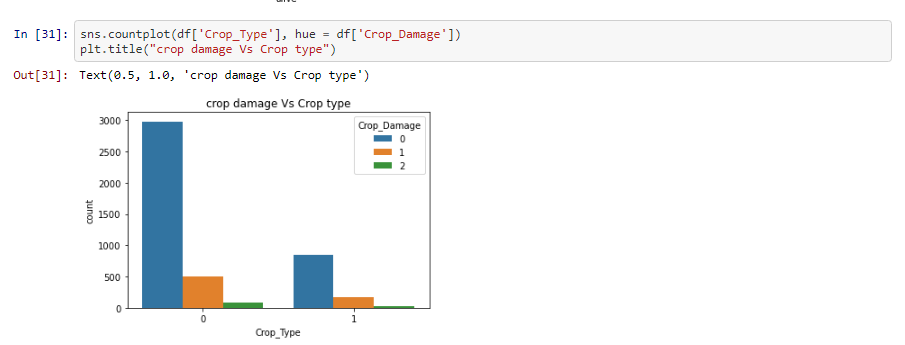
Observation-:

1. 'Estimated\_Insects\_Count', 'Crop\_Type', 'Pesticide\_Use\_Category', 'Number\_Weeks\_Used' have positive correlation with 'Crop\_Damage'.

2. 'Soil\_Type', ,Number\_Doses\_Week', 'Number\_Weeks\_Quit', 'Season' have negative correlation

with 'Crop\_Damage'

**Crop type and damage count**

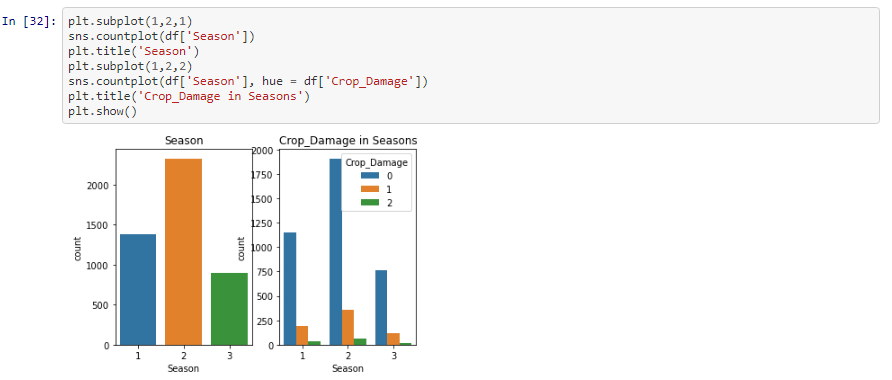


**Observations-:**

1.Type 0 and crop type1 have very huge differnce in the crop that are alive.

2. Crop type 0 are more damaged due to other reasons as compared to type 1 crop 1

**In which season the crop type are more damage are shown below-:**



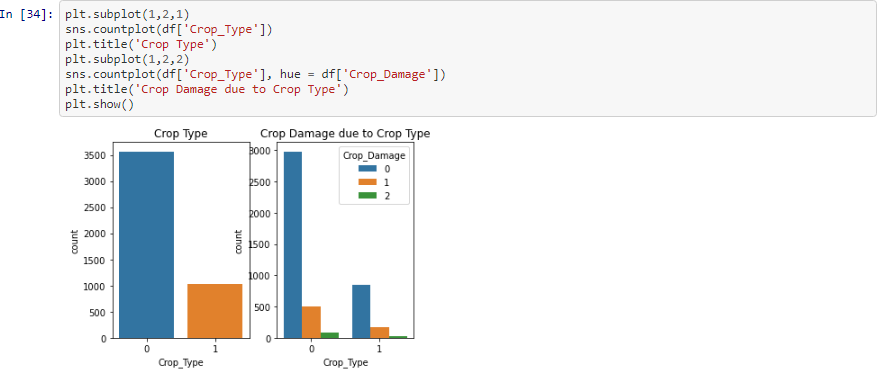
The crop type 0 is damaged in season 2 more than other.

**The effect of soil type in crop life**.



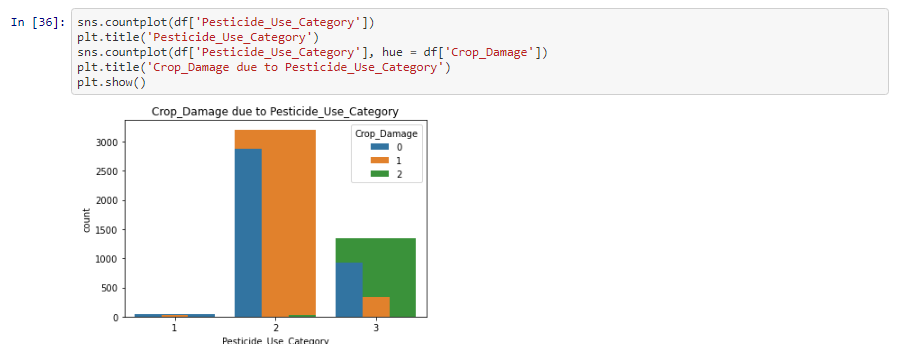
There is a very less effect of soil type on the crop damage.

**Crop type effect on damage of crop.**



The crop type has a huge effect on the crop. As crop type 0 is more alive as compaed to the crop type 1.

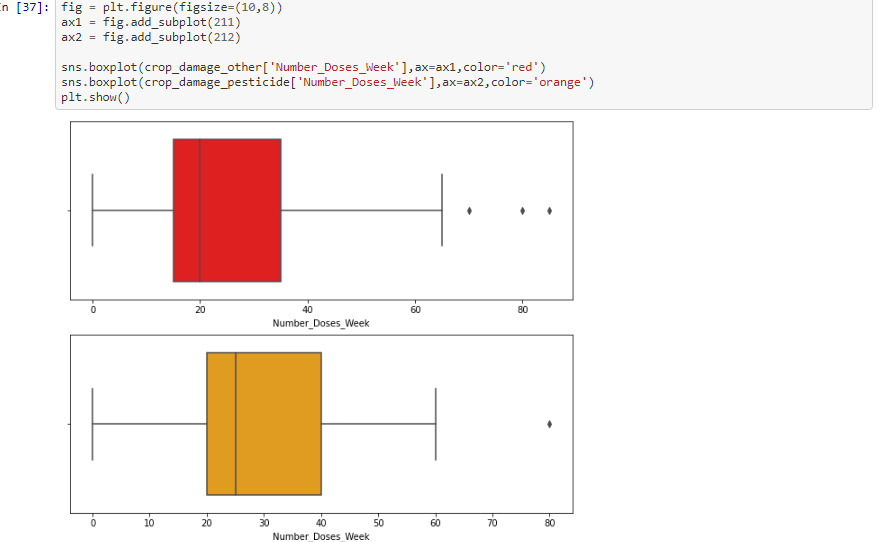
**Pesticide use**



Pesticide used catagory type has more better effect as type 2 pesticide has more crop alive than compared to ther pesticide category

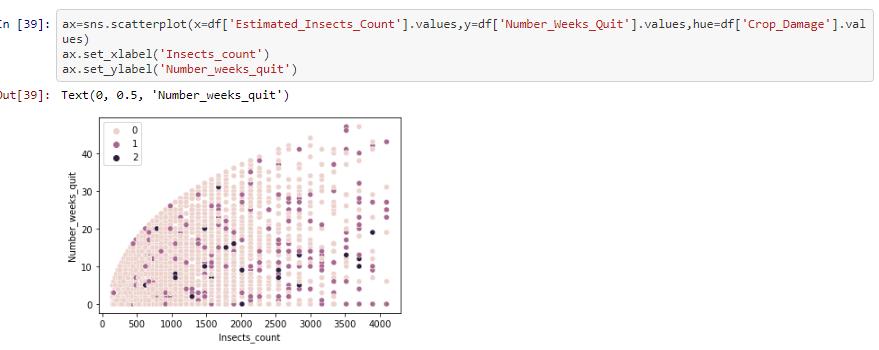
**b.boxplot,and others**

**No. of doses per week impact on crop damage**

****

Crop damaged by other reason is positively skewed and crop damaged by pesticide is more of a normal distribution. Even though they both have some outliers present.

**Insects impact on crop damage**



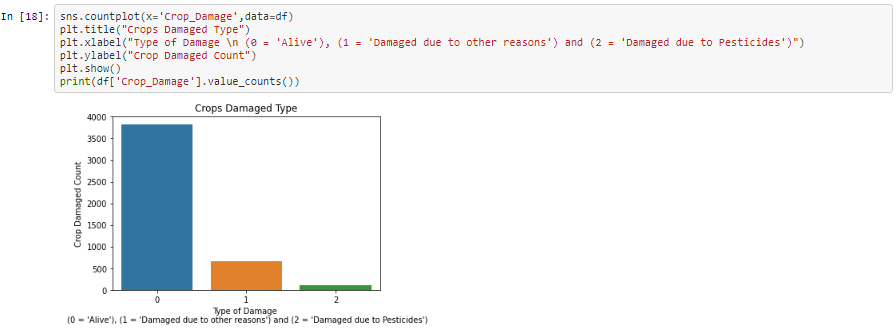
In the number of weeks with no pesticide increase, insect count also increases. when the insect count increases then crop damage due to other reasons are also increased.

**c.count of crops alive**

**Distribution of crop damage**

****

**Crop damage according to type**



TYPE0 3820

TYPE1 664

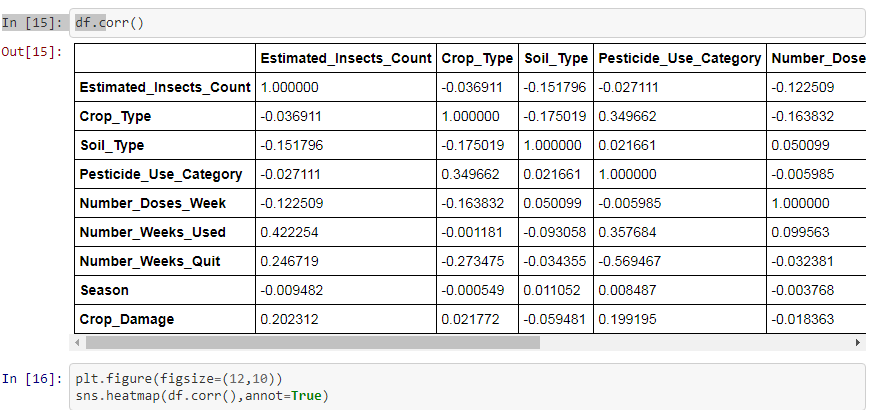
TYPE2 115

Name: Crop\_Damage, dtype: int64

Crop alive is 3820,Crop Damaged by pesticide 115 and crop damaged by other reasons is 664.

**d.pairplot, heatmap, correlation map**

**Correlation map**

****

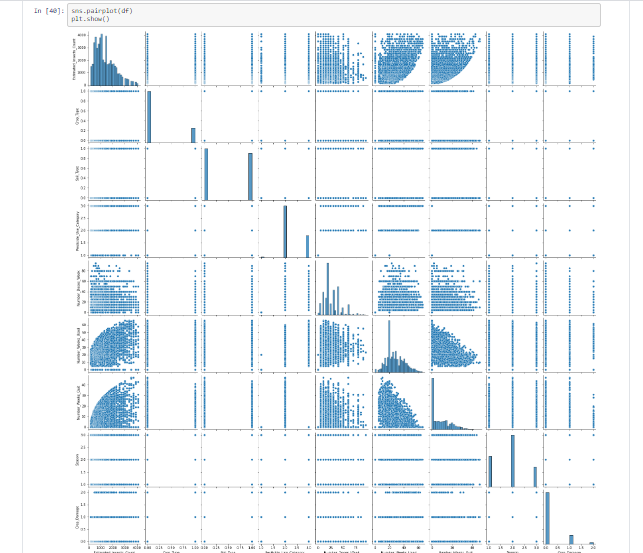
**Observation-:**

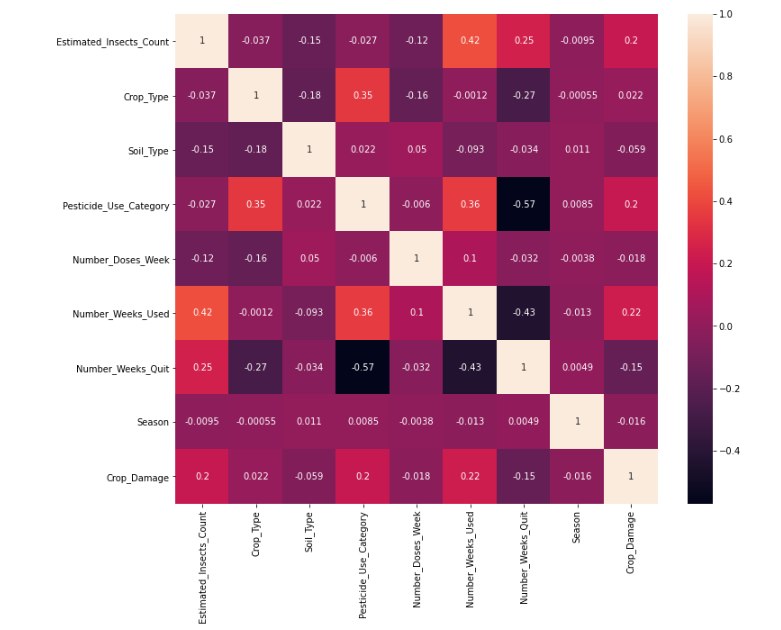
1.Crop\_Type and Pesticide\_Used\_category has some positive correlation.

2.Number\_Weeks\_Used has some positive correlation with the Estimated\_Insetcs\_Count and Pesticide\_Used\_Category.

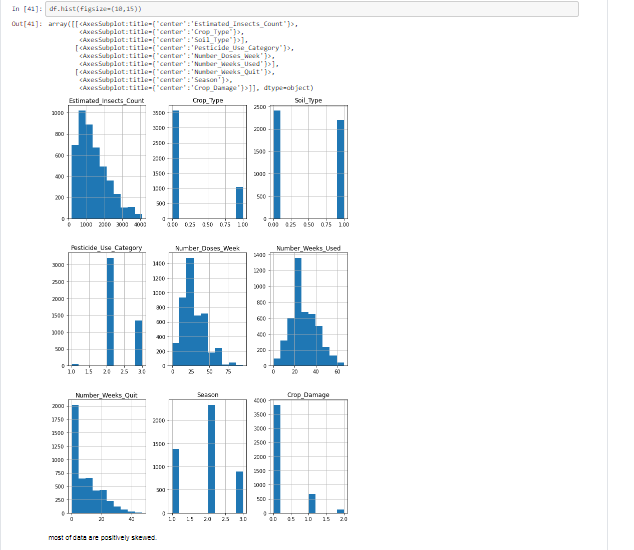
3.Number\_Weeks\_Quit has some negative correlation with Crop\_type and Number\_Weeks\_Used. 4.Number\_Weeks\_Quit has some negative correlation with Pesticide\_Use-Category and Number\_Weeeks\_Used.

**Pairplot**



**Correlation map is below-**

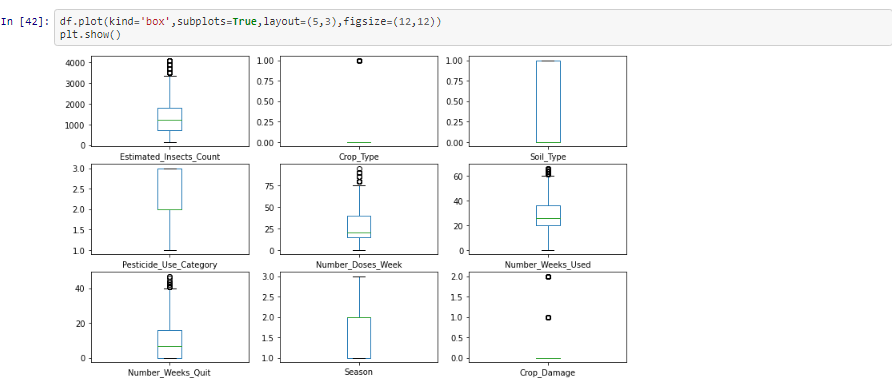
**e.distibution and relationship distplot**



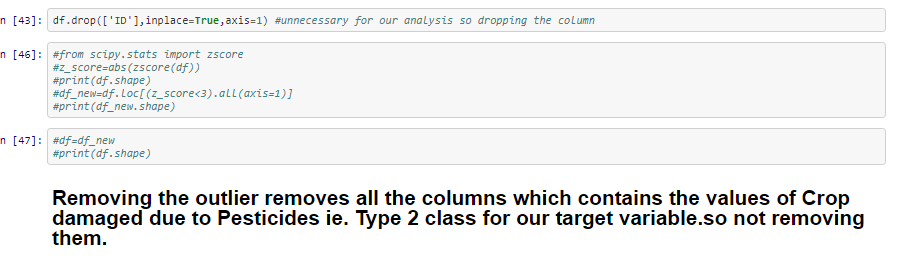
Mostly data are normally distributed.

**4. Pre-processing Pipeline**

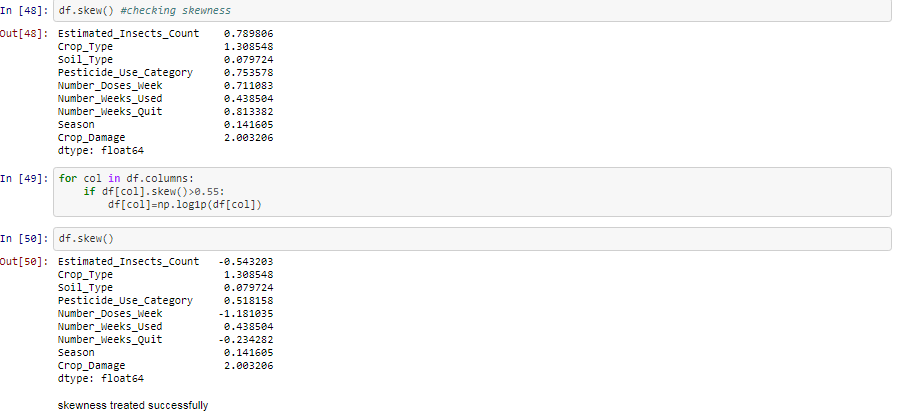
**a. Outliers and how to remove them**



Insects Count, Crop Type, Number Does Week, Number Week Used and Crop Damage have outliers.will be treating the outliers.

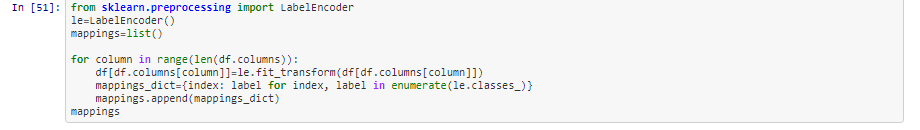


**b. Checking and removing skewness**



Skewness are acceptable upto -0.5 to +0.5 and we can cleary see the skewness was present so we treated them.

**c. Label encoding to encode the data in binary form to be understandable by machine andd the mappings of the data also can be seen.**

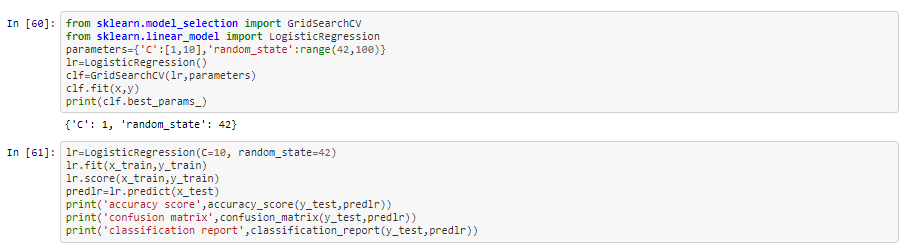


**5. Building Machine Learning Models**

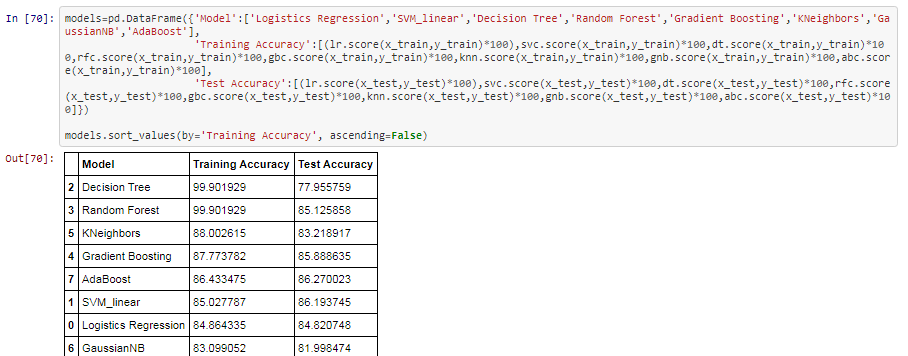
**a.Selection of best model**

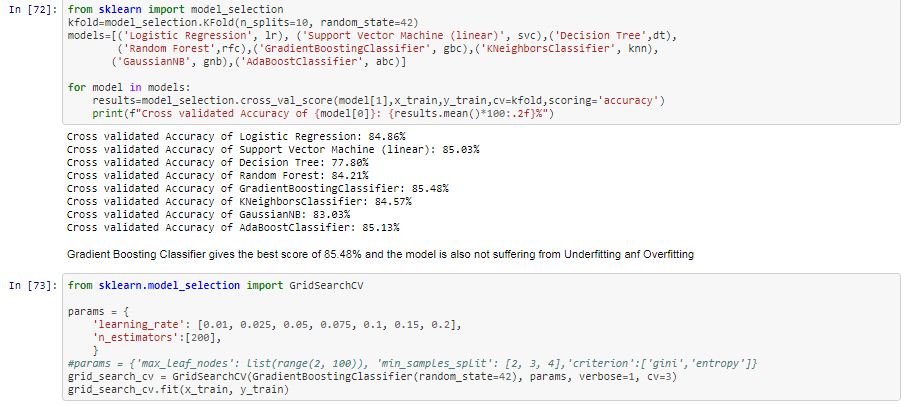


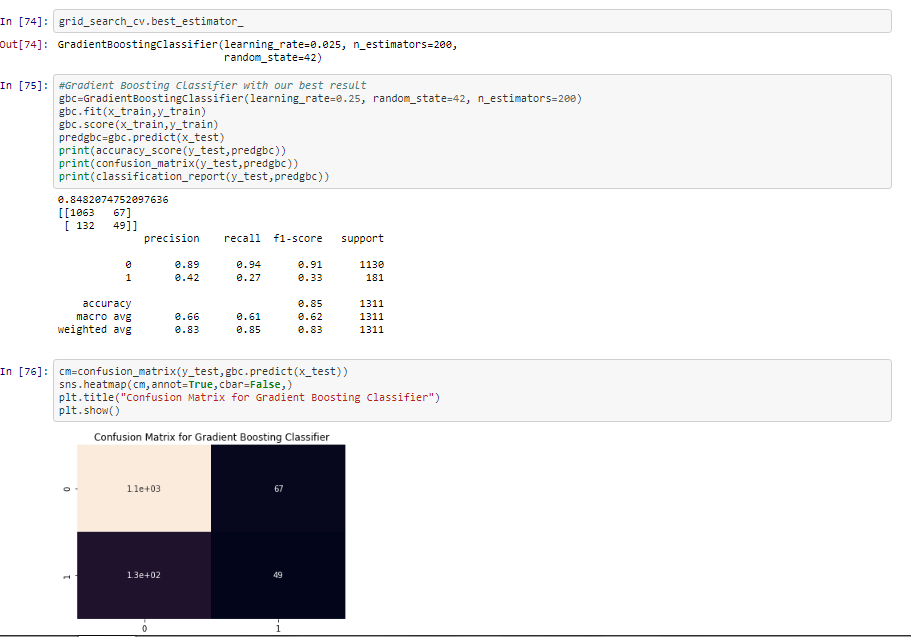
Used GridSearchCV for best paarmeters.



All models and their respective results to choose the best model.

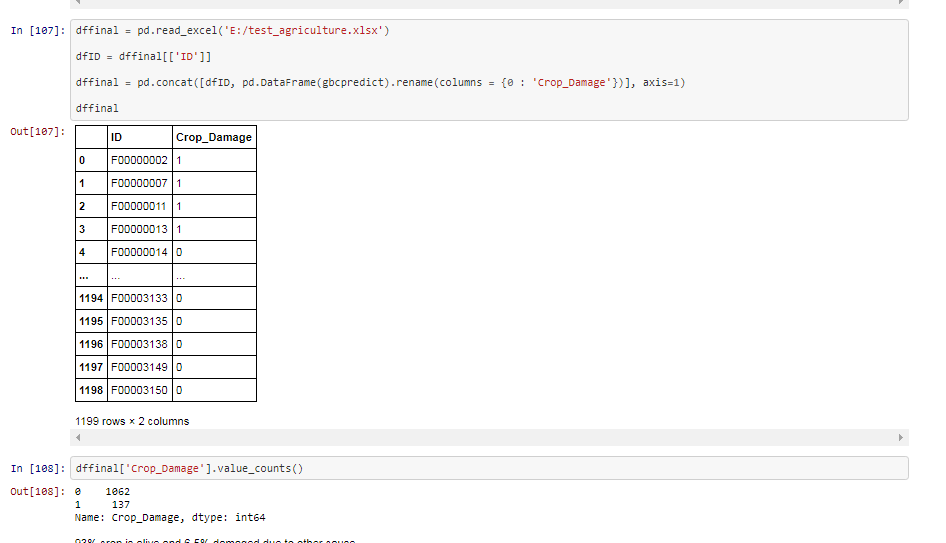






We find that the GNB classifier is the best model for our dataset.

**b. Loading train result and train model for test dataset**



**c. Saving the results**



**6. Concluding Remarks**

Hence we can say from the above that EDA is very important as it helps us in achieving the following :-

1. It helps in detection of mistakes (like missing values and outliers)
2. It determines relationships between explanatory variables
3. Assessing the direction and rough size of relationships between explanatory and outcome variables.
4. It makes our data ready for machine learning algorithm

**Pro-Remarks**

**Impacts of Data Science in Agriculture and Conclusion**

Digital transformation in agriculture has brought about a variety of innovations in the present world. One of such initiatives is MyCrop, an intelligent and self-learning real-time system which takes into account each farmer’s location, weather, and crop data. It uses big data, machine learning, and smartphone technology to deliver information, expertise, and resources to smallholder farmers (Trendov, Varas, & Zeng, 2019).

Large scale farmers and those in the industrial countries are using modern technology for automation in their farms. They have turned their farms into a factory; and have automated every process possible with a heavy amount of computing. Such type of farming is called smart farming (“The Future of Agriculture”, 2016).

At TH Milk facility, Vietnam, data science is being used to regulate the quality of milk production in cows where every cow is tagged with an RFID chip. The milking process is automated with sensors in the suckers that can detect inflammation in the cow’s mammary glands. If inflammation is detected, the machine will stop the milking process and the cow will be marked and checked. In AfiMilk, a similar chip is attached to each goat’s legs that tracks its movement. If the goat does not move for a long period, or if it is showing ambiguous sleeping patterns then it will be checked for illness. (Kshetri, 2016).

With the movement of the world towards digital agriculture, a lot of investment has been poured into it. Numerous research and development are going on to maximize the efficiency of farms. Incorporating new technology will escalate the yields of both small- and large-scale farms. New trends in farming by including data science and analytics will revolutionize the agriculture industry; producing higher-quality foods in a larger quantity sustainably so that the global target of increasing food production by 70% will be met by 2050.

**Happy reading,**

**Vaishali Shukla**