**Agriculture Dataset**

In recent times, the concept of smart farming has made agriculture more effective and efficient.

Machine Learning has contributed a lot to quantify and understand the agricultural pattern in daily operations and led to smart farming with precision.

In the current dataset of agriculture, we would determine the outcome of the harvest i.e., whether the crop would be healthy (alive), damaged due to pesticides or damaged by other reasons.

We would be dealing with the following attributes and determine the outcome of the harvest.

First, we will train the model with train dataset and then check the score with the test dataset.

The attributes are as follows: -

Data Description: -

ID - Unique ID

Estimated\_Insects\_Count- Estimated Insects count per square meter

Crop\_Type - Category of crop (0,1)

Soil\_Type -Category of soil (0,1)

Pesticide\_Use\_Category - Type of pesticide uses (1-Never,2-Previously used,3-Currently using)

Number\_Doses\_Week - No. of doses per week

Number\_Weeks\_Used - No of weeks used

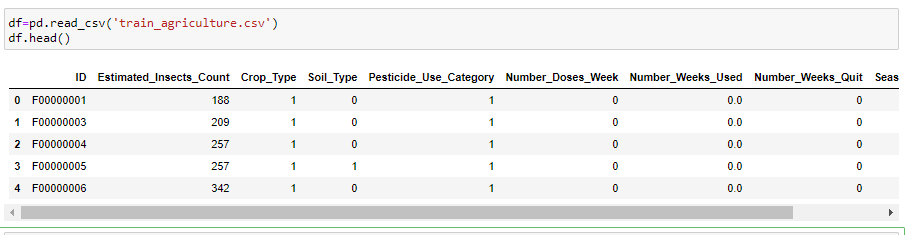
Number\_Weeks\_Quit - No of weeks quit

Season - Season category(1,2,3)

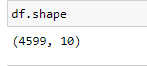
Crop\_Damage -Crop damage category(0=alive,1=Damage due to other causes,2=Damage due to pesticide)

We have imported numpy and pandas for data analysis in jupyter notebook, seaborn and matplotlib for visualization, warnings to ignore the version change and Label Encoder to chage object data to numeric.

First, Imported the dataset into jupyter notebook

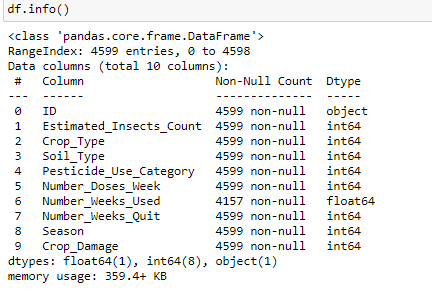


Next checking the shape of the dataset:-



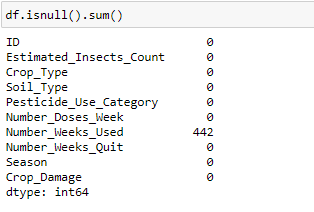
There are 4599 rows and 1o columns

Next, checking the data type-



#Unique ID is object type, Number\_Weeks\_Used is float rest int type

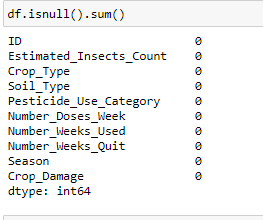
Now, checking if there is any null values or not:-



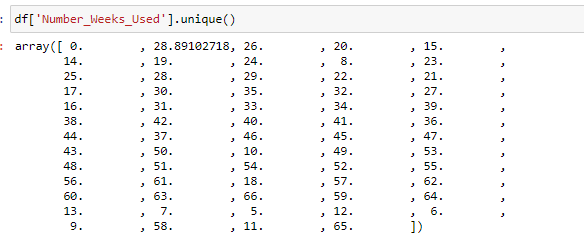
No null values are there except for Numer\_Weeks\_Used and filling the null values with mean.



Then rechecking the null values: -

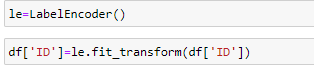


Now checking the unique values of week and converting it into integer: -

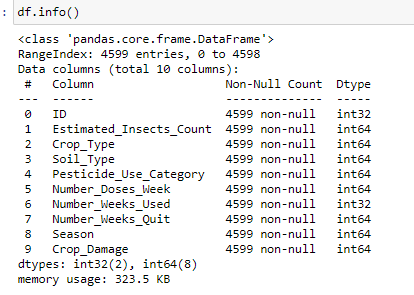




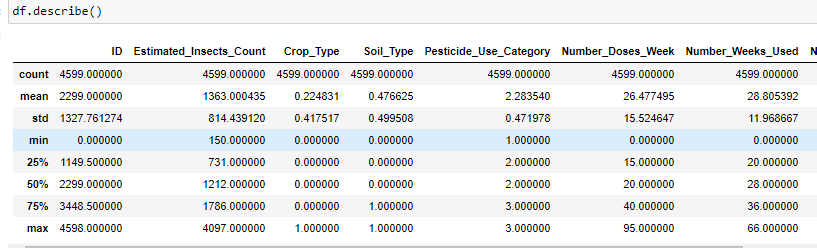
Now, changing the ID into numeric through LabelEncoder



Re-checking the data-type and now all are in integer.



Now, checking the variance of the dataset through describe function.

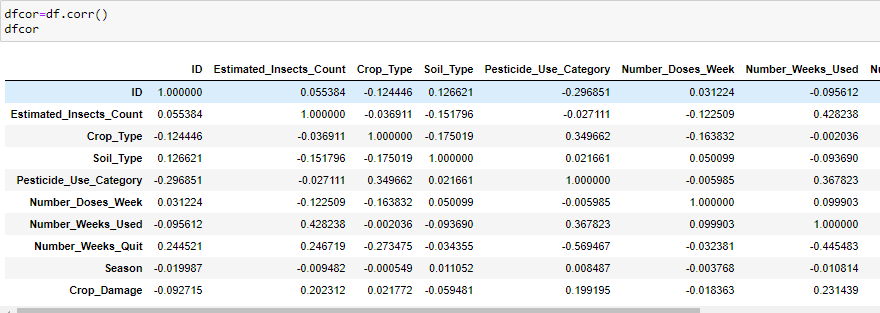


#There is not much difference between mean and median except for ID,Estimated\_Insects\_Count

#The dataset is skewed

#Stand deviation is high in the columns - ID,Estimated\_Insects\_Count,Number\_Doses\_Week,Number\_Weeks\_Used and Number\_Weeks\_Quit

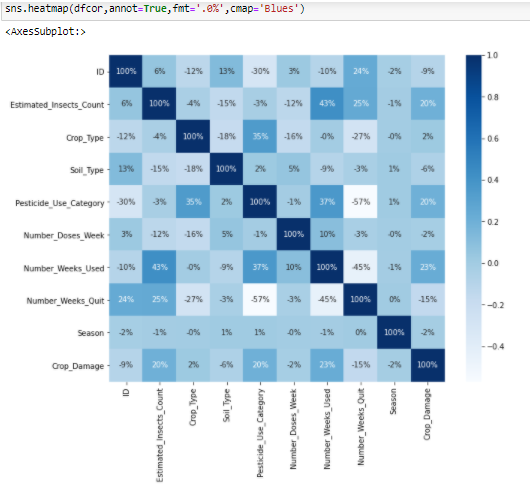
Now, checking the correlation between each columns:-



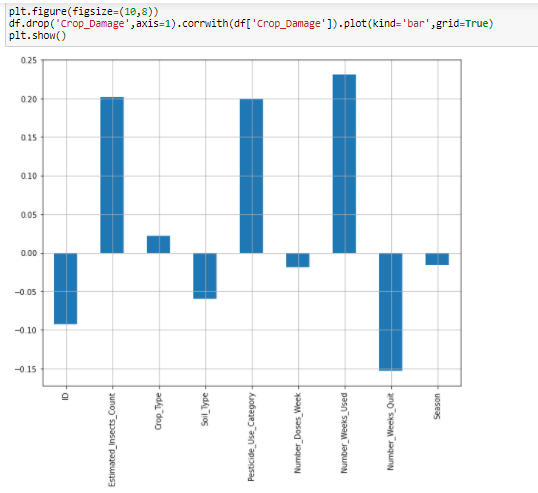
#Estimated\_Insects\_Count,Pesticide\_Use\_Category,Number\_Weeks\_Used are good correlation with the column crop damage

#ID, No of weeks quit,soil type has negative correlation which is below graphically represented

Graphically, representing the correlation through heatmap.



Now, finding the correlation with the target column -Crop Damage with other columns: -

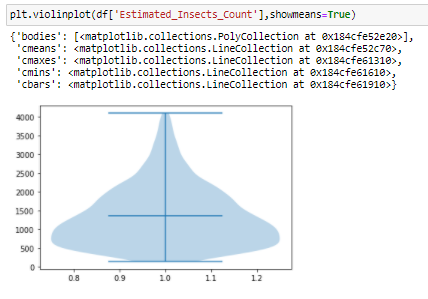


Crop damage has high positive correlation with the columns – estimated insect count, Pesticide use and number of weeks used.

Got good negative correlation with the column Number of weeks quit, soil type and ID

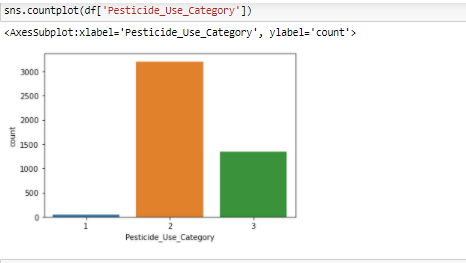
**Now Visualizing the dataset for better understanding**

First checking the univariate analysis: -



#Maximum 'Estimated\_Insects\_Count' is between 500 and 1000

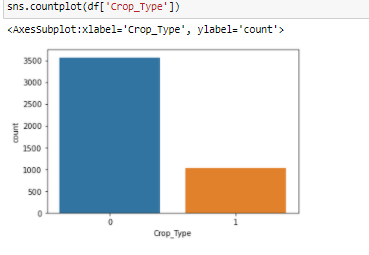
Second, checking the pesticide used category:-



#Pesticide\_Use\_Category - Type of pesticide uses (1-Never,2-Previously used,3-Currently using)

# Never used pesticide is the lowest,previously used is highest.

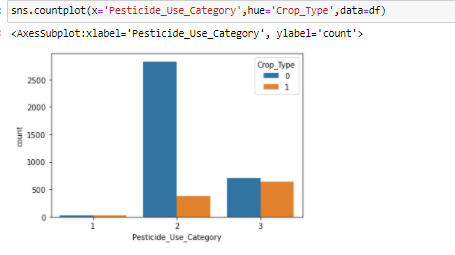
Third, checking the Crop Type using countplot



#Crop type 0 is much higher to crop type 1

**Bivariate Analysis**

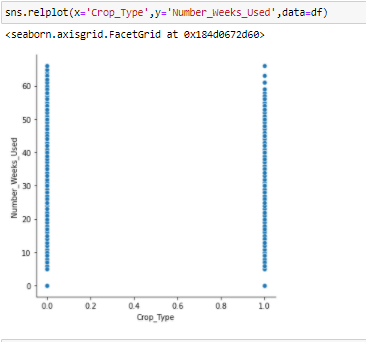
**Checking the relation between pesticide use category and crop type**



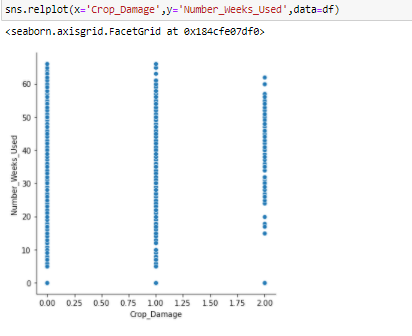
#crop type was maximum when pesticide was used , being used pesticide(currently) cropy type growith is more or less equal

#crop growth is minimal when pesticide not used

**Now, checking the correlation between Crop Type and Number of Weeks used through relation plot**



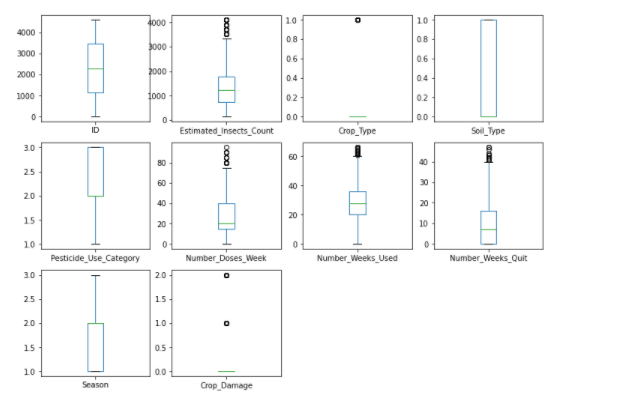
#Both crop type has been uniformly used in correlation to each other



#Crop\_Damage -Crop damage category(0=alive,1=Damage due to other causes,2=Damage due to pesticide)

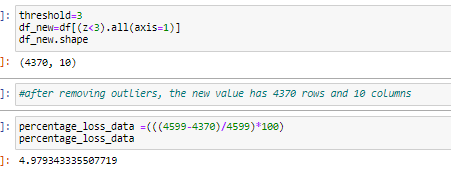
#in relation to number of weeks used crop alive rate and damaged is more or less same

**Now checking the outliers in the dataset:-**



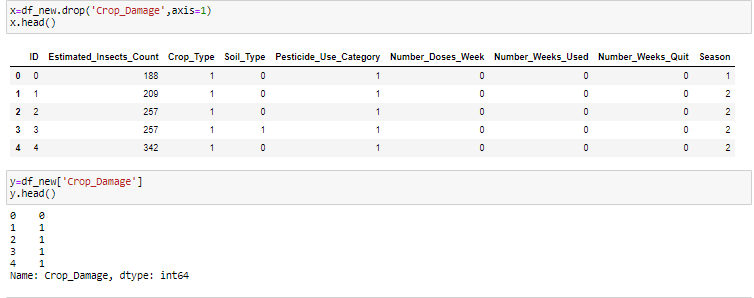
Estimated insect count,number doses week, number weeks used, number weeks quit has good amount of outliers.

**Next, removing the outliers from the dataset through zscore:-**



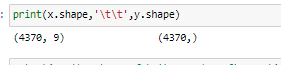
The percentage loss in dataset is 4.97 which we can go ahead with.

**Splitting the dataset into x and y for Machine Learning purpose: -**



The target column being the crop damage, rest being the featured columns.

Checking the shape of x and y



There are 4370 rows and 9 columns for x and for y 1 column and same rows.

**Now, scaling the dataset through power\_transform:-**

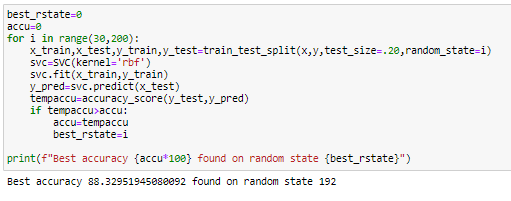


**Next, scaling the data to bring into a uniform range**



**Importing various Classifier algorithms to check the accuracy, confusion\_matrix,classification\_report,cross validation score.**

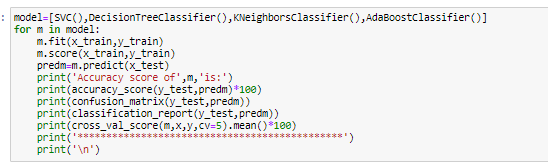
**from sklearn.model\_selection importing train\_test\_split & GridSearchCV to split the dataset and hyper tune the model.**

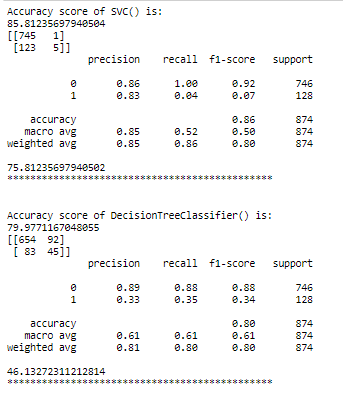


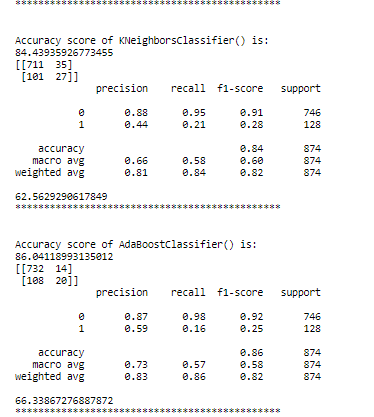
**Finding the best random state 192 with the accuracy score 88 percent.**

**Now, checking the score with different algorithm**

**SVC, DecisionTreeClassifier,KNN,AdaBoostClassfier**

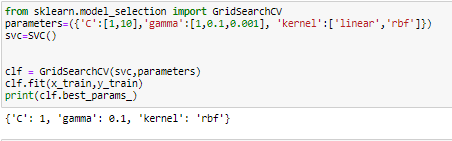






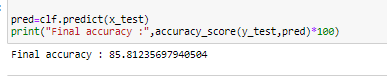
#from the above model SVC is best as the diff between accuracy and cross val is minimum

**Further checking the best parameter of SVC through GridSearchCV**



Gamma is 1 and kernel = rbf

**Predicting the score:-**

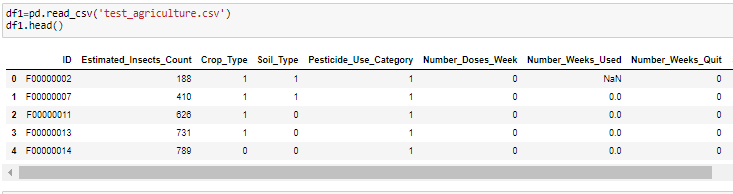


**And saving the model for test dataset**



**Test Dataset:**

**Importing the test dataset:-**

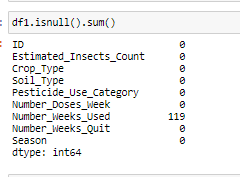


**Checking the shape**



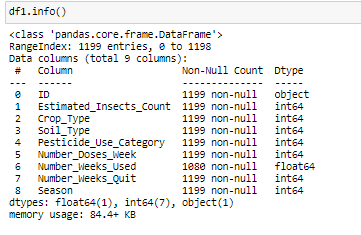
There are 1199 rows and 9 columns

**Checking the null values**



Null values exits in Number\_Weeks\_Used

Checking the data type:-

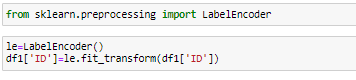


#ID is object datatype, number of weeks used in float, rest all are integer.

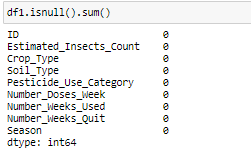
Filling the null values with mean



Converting the ID into numeric through Label Encoder

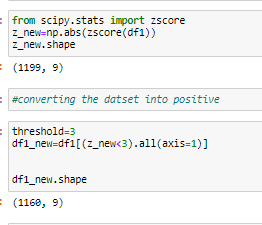


Rechecking the null values



No null values are there

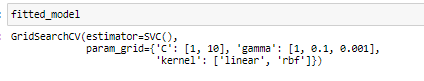
Removing the outliers through zscore



Importing the trained model



Checking the attribute of fitted model



Predicting the score

