**Machine Learning on Agriculture Dataset**



**Date:- 27-Feb-21 Author- Vikas Ojha**

**Problem Definition**

Context:

In the present world, concept of Machine Learning is widely used in all the sectors to determine upcoming instances. One of such field is Agriculture. We can use in all the process of Agriculture from seed selection as per soil type, soil preparation, seed breeding along with water feed management.

Machine Learning Technique emerged as smart way in field of Agriculture for Smart farming that makes agriculture more efficient and effective with its high precision Algorithms.

Content:

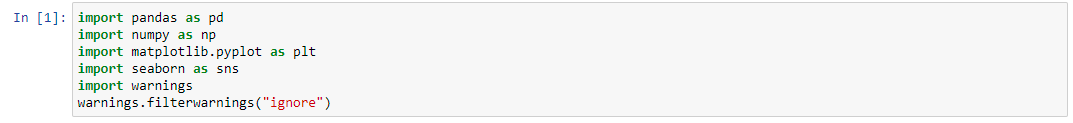
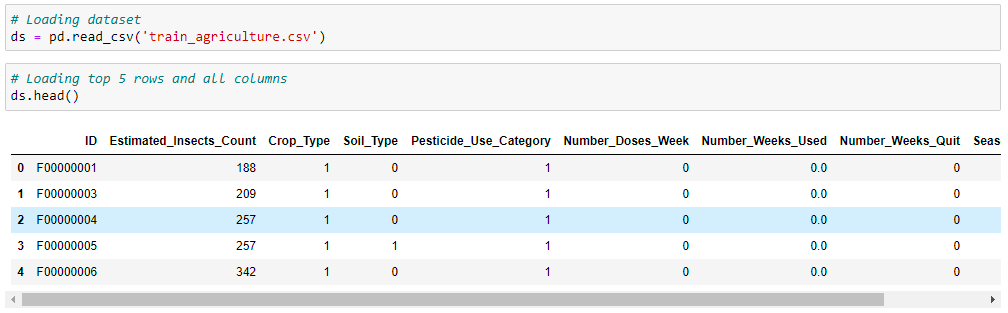
The farming task or farmers job is not an easy task. They have to work hard on the agricultural field with high strength and determination. Initially they prepare the land for sowing the seed into it. Once they sow the seed into land, they have to work day and night to make sure that he gets maximum output (Harvest) at the end of season. Availability of water, soil fertility, protecting crops from rodents, timely use of pesticides & other useful chemicals and nature are some important factors that ensure good harvest at the end of season. As all these factors are not possible for a farmer to control but the amount and frequency of pesticide can be controlled.

Pesticides are also very important as they protect the crop from damage. If we use pesticides more than the required, it may damage the entire crop which means dead crop and even can spoil the soil. The data on which we are going to work is based on crops harvested by various farmers at the end of season.

We need to predict the outcome of harvest season, weather the crop would be healthy(alive) or damaged by pesticides or by any other reason.

**Data Analysis**

Before we start the analysis of data, we need to import the supporting libraries such as pandas, numpy, matplotlib, seaborn, etc. These libraries are essentially required to perform task on our dataset.

1. *Support Libraries:*
2. *Loading Dataset:*

Now, as we have loaded our dataset, it was found that we have 2 different files for train and test data. Initially we have to work on train data and prepare our model to look which model is giving best accuracy. Then we will save our model in. pkl form.

Then we will load test data and will perform all the tasks as performed on train data. Then we will import our saved model and we will fit the test data into our model to get our result.

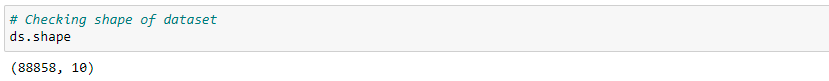
Dataset contains 10 columns namely ID, Estimated Insects Count, Crop Type, Soil Type, Pesticide Use Category, Number Doses Week, Number Weeks Used, Number Weeks Quit, Season and Crop Damage

Following are the details of each columns: -

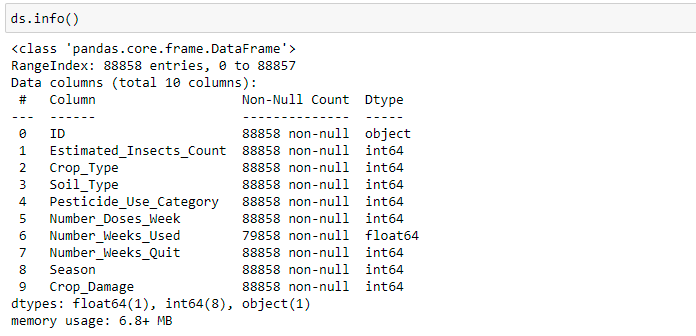
|  |  |  |
| --- | --- | --- |
| ***Sr No.*** | ***Column*** | ***Description*** |
| *1* | **ID** | *Unique Identification Number given to each crop.* |
| *2* | **Estimated Insects Count** | *Estimated Insects Count on the crop*  *(Values between 150 – 4097)* |
| *3* | **Crop Type** | *Category of the Crop*  *(****0*** *- Rabi /* ***1*** *- Kharif)* |
| *4* | **Soil Type** | *Category of the soil in which the crop was grown. (****0*** *-* Alluvial */* ***1*** *- Black-Cotton)* |
| *5* | **Pesticide Use Category** | *Category of the Pesticide Used on the crop*  *(****1*** *- Insecticides /* ***2*** *- Bactericides /* ***3*** *- Herbicides)* |
| *6* | **Number Doses Week** | *Number of Doses given to the crop*  *(Values between 0 – 95)* |
| *7* | **Number Weeks Used** | *Number of Weeks for which the pesticide*  *(Values between 0 – 67)* |
| *8* | **Number Weeks Quit** | *Survival duration of the crops in weeks.*  *(Values between 0 – 50)* |
| *9* | **Season** | *The season in which the crop was grown*  *(****1*** *- Summer/* ***2*** *- Monsoon/* ***3*** *- Winter)* |
| *10* | **Crop Damage** | *Classification of the amount of damage to the crop. (****0*** *- Minimal /* ***1*** *- Partial /* ***2****- Significant Damage)* |

In this project we will use different exploratory data analysis techniques. Before we proceed further, we need to understand more about dataset. Let’s understand the variables to which we are playing with.

Now let’s check the shape of our train data.



We can see that our train dataset contains 88858 rows and 10 columns. Now let’s check on the missing data in our dataset.

1. *Dataset Types:*

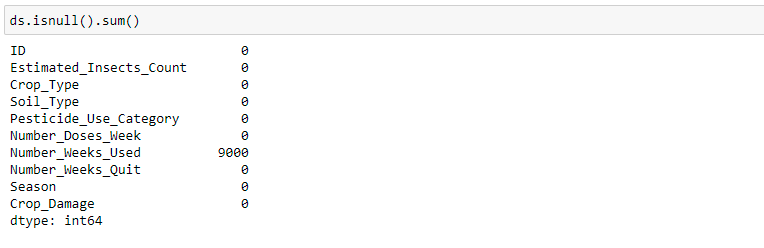
There are 3 different datatypes, namely object, int64 and float64. ID column is of categorical type which we need to convert afterwards using Label Encoder, if needed as machine learning model cannot understand categorical variables.

We will perform the following steps further in this dataset: -

1. Finding missing values in dataset, if any
2. Univariate and Bi Variate Analysis
3. Summary Statistics
4. Correlation Check
5. Skewness check
6. Label Encoding of categorical data
7. Checking and removing outliers, if any
8. Defining x and y variables
9. Removing skewness from x variable
10. Apply standard scaler
11. Apply different Algorithms on model
12. Hyper Parameter tuning using GridsearchCV
13. Saving the best model

**Exploratory Data Analysis**

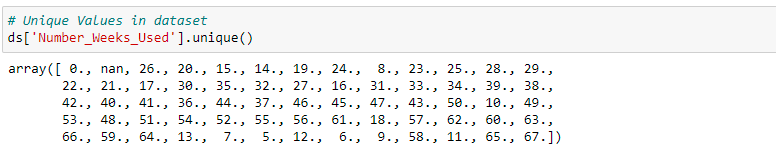
As discussed above, now we will check the missing data in Dataset.



Oh No! Number Weeks Used Column contains 9000 missing data. We need to replace these NaN values with some appropriate data.

**Handling NaN Values**

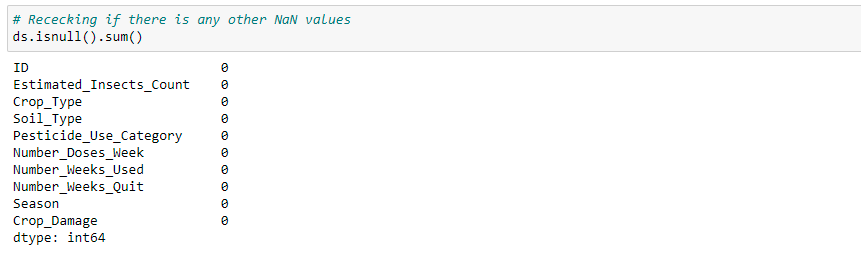
Now let’s check the unique values in Number\_Weeks\_Used column.



There are two ways to handle this NaN values. First, we can replace the NaN values with the mean of the column or secondly, we can fill the missing values with the number which has highest number of occurrences in the column. It is known as Mode. I choose the Mode method to replace the NaN values as it suits the best in present scenario.



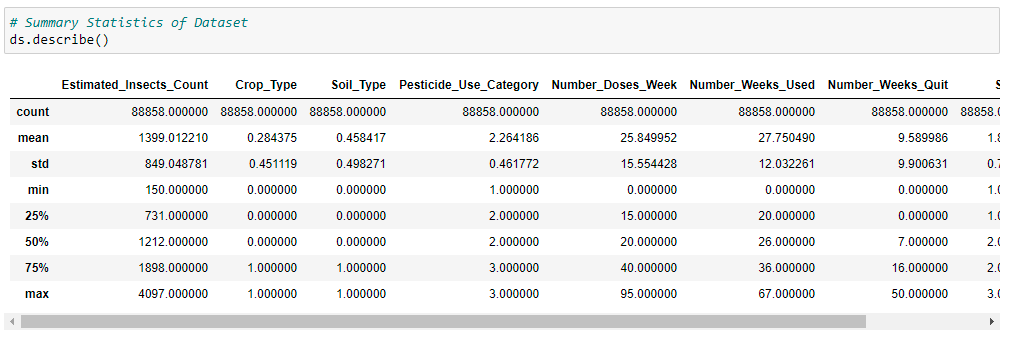
Now we have filled the missing data with mode. Let’s verify whether the missing value have been replaced or not.



There are no missing values in our dataset now. We can now proceed with further analysing our dataset.

**Summary Statistics**

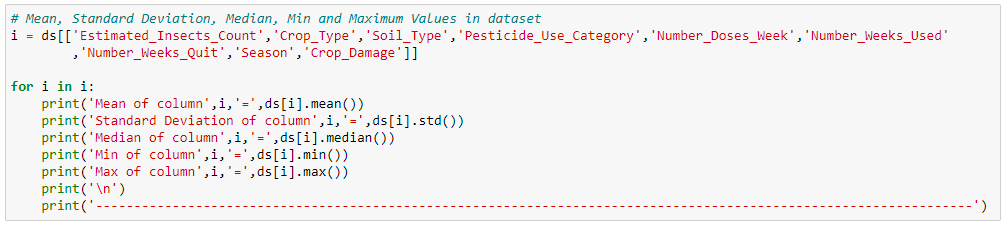
Let’s draw some observations from summary statistics.



Observations:

1. Few columns are highly skewed.
2. Few columns have outliers present, we will check it later.

Now finding mean, standard deviation, median, minimum and maximum values in our dataset.



Observations:

Mean of column Estimated\_Insects\_Count = 1399.0122104931463

Standard Deviation of column Estimated\_Insects\_Count = 849.0487806670102

Median of column Estimated\_Insects\_Count = 1212.0

Min of column Estimated\_Insects\_Count = 150

Max of column Estimated\_Insects\_Count = 4097

----------------------------------------------------------------------------------------------------------------

Mean of column Crop\_Type = 0.2843750703369421

Standard Deviation of column Crop\_Type = 0.4511188091517885

Median of column Crop\_Type = 0.0

Min of column Crop\_Type = 0

Max of column Crop\_Type = 1

----------------------------------------------------------------------------------------------------------------

Mean of column Soil\_Type = 0.4584167998379437

Standard Deviation of column Soil\_Type = 0.4982706408302739

Median of column Soil\_Type = 0.0

Min of column Soil\_Type = 0

Max of column Soil\_Type = 1

----------------------------------------------------------------------------------------------------------------

Mean of column Pesticide\_Use\_Category = 2.2641855544801817

Standard Deviation of column Pesticide\_Use\_Category = 0.4617715815834632

Median of column Pesticide\_Use\_Category = 2.0

Min of column Pesticide\_Use\_Category = 1

Max of column Pesticide\_Use\_Category = 3

----------------------------------------------------------------------------------------------------------------

Mean of column Number\_Doses\_Week = 25.849951608183844

Standard Deviation of column Number\_Doses\_Week = 15.55442776539594

Median of column Number\_Doses\_Week = 20.0

Min of column Number\_Doses\_Week = 0

Max of column Number\_Doses\_Week = 95

----------------------------------------------------------------------------------------------------------------

Mean of column Number\_Weeks\_Used = 27.750489545116928

Standard Deviation of column Number\_Weeks\_Used = 12.032260928714667

Median of column Number\_Weeks\_Used = 26.0

Min of column Number\_Weeks\_Used = 0.0

Max of column Number\_Weeks\_Used = 67.0

----------------------------------------------------------------------------------------------------------------

Mean of column Number\_Weeks\_Quit = 9.589986270228904

Standard Deviation of column Number\_Weeks\_Quit = 9.900631115270183

Median of column Number\_Weeks\_Quit = 7.0

Min of column Number\_Weeks\_Quit = 0

Max of column Number\_Weeks\_Quit = 50

----------------------------------------------------------------------------------------------------------------

Mean of column Season = 1.8969591933196785

Standard Deviation of column Season = 0.7013221327434505

Median of column Season = 2.0

Min of column Season = 1

Max of column Season = 3

----------------------------------------------------------------------------------------------------------------

Mean of column Crop\_Damage = 0.19056247045848432

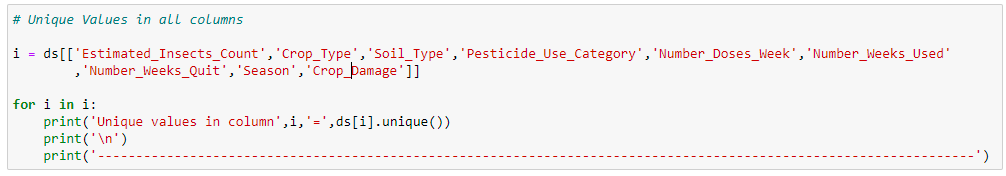
Standard Deviation of column Crop\_Damage = 0.4542150681956372

Median of column Crop\_Damage = 0.0

Min of column Crop\_Damage = 0

Max of column Crop\_Damage = 2

We will have a look in the unique values in all the columns.



Observations:

Unique values in column Estimated\_Insects\_Count = [ 188 209 257 342 448 577 731 1132 1212 1575 1785 2138 2401 2999 3516 3895 4096 150 151 168 232 231 256 283 312 311 375 411 410 488 489 531 626 625 677 732 789 790 851 850 915 916 984 1056 1296 1297 1385 1478 1576 1678 1786 1898 2015 2016 2139 2267 2402 2542 2541 2688 2687 2840 3164 3165 3336 3337 3702 3896 169 3515 4097]

----------------------------------------------------------------------------------------------------------------

Unique values in column Crop\_Type = [1 0]

----------------------------------------------------------------------------------------------------------------

Unique values in column Soil\_Type = [0 1]

----------------------------------------------------------------------------------------------------------------

Unique values in column Pesticide\_Use\_Category = [1 3 2]

----------------------------------------------------------------------------------------------------------------

Unique values in column Number\_Doses\_Week = [ 0 30 40 60 20 25 45 50 15 10 5 35 70 55 80 90 85 65 75 95]

----------------------------------------------------------------------------------------------------------------

Unique values in column Number\_Weeks\_Used = [ 0. 20. 26. 15. 14. 19. 24. 8. 23. 25. 28. 29. 22. 21. 17. 30. 35. 32. 27. 16. 31. 33. 34. 39. 38. 42. 40. 41. 36. 44. 37. 46. 45. 47. 43. 50. 10. 49. 53. 48. 51. 54. 52. 55. 56. 61. 18. 57. 62. 60. 63. 66. 59. 64. 13. 7. 5. 12. 6. 9. 58. 11. 65. 67.]

----------------------------------------------------------------------------------------------------------------

Unique values in column Number\_Weeks\_Quit = [ 0 1 9 4 13 3 14 6 2 12 15 10 8 11 7 5 17 18 16 20 21 19 22 24 23 27 25 28 26 29 30 31 33 32 36 37 34 39 44 41 35 47 46 40 42 38 43 45 49 50 48]

----------------------------------------------------------------------------------------------------------------

Unique values in column Season = [1 2 3]

----------------------------------------------------------------------------------------------------------------

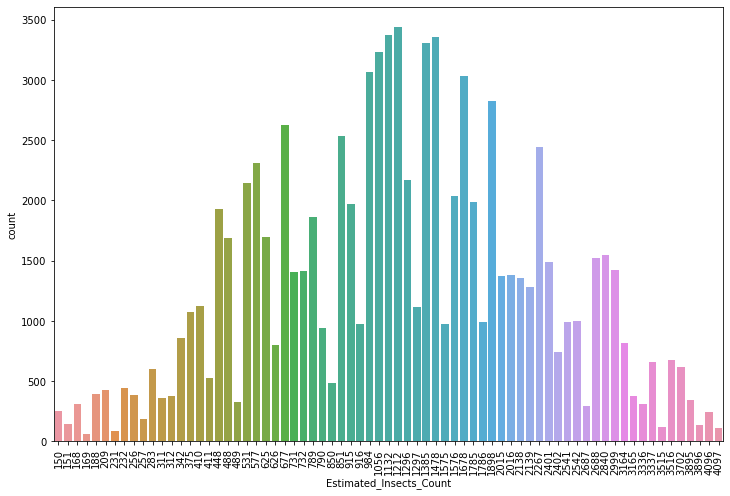
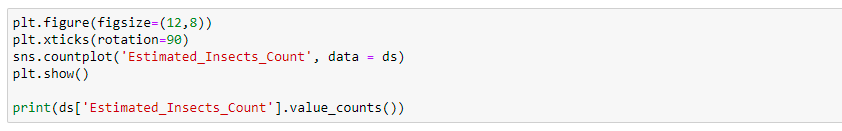
Unique values in column Crop\_Damage = [0 1 2]

We have seen all the unique values in all the columns.

**Univariate Analysis**

## We will now plot the countplot of all the columns of draw some useful insights from dataset .

Estimated Insects Count

****

Observations:

Following are the top 5 insects counts.

1212 - 3438

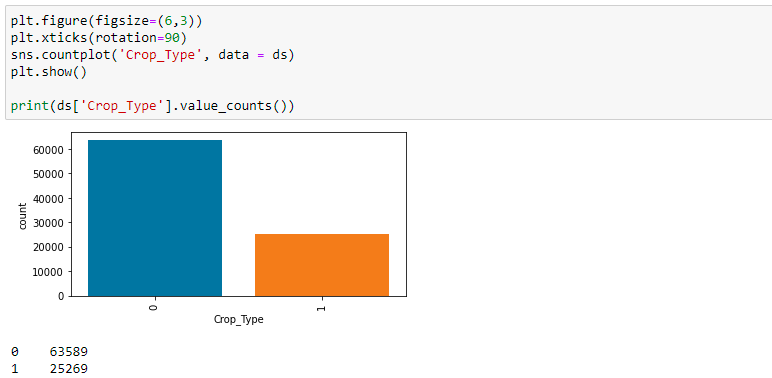
1132 - 3370

1478 - 3358

1385 - 3309

1056 - 3235

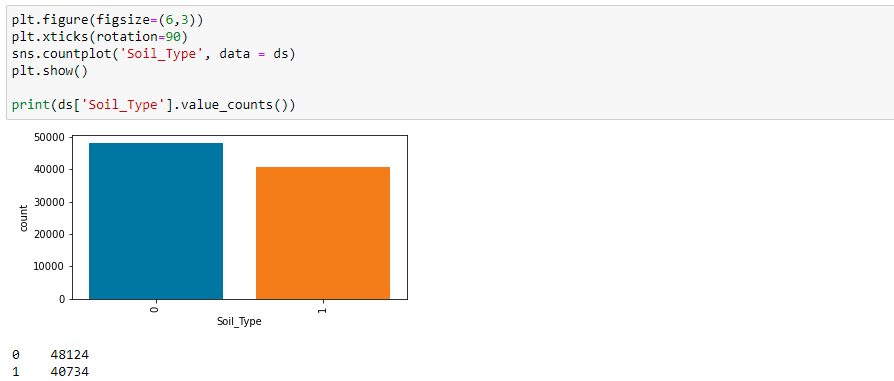
Crop Type



Observations:

1. 0 is Rabi crop and 1 is Kharif Crop
2. Rabi crop has the maximum number of instances. As per dataset 71.5 % data are Rabi Crop
3. Rest are Kharif Crop.

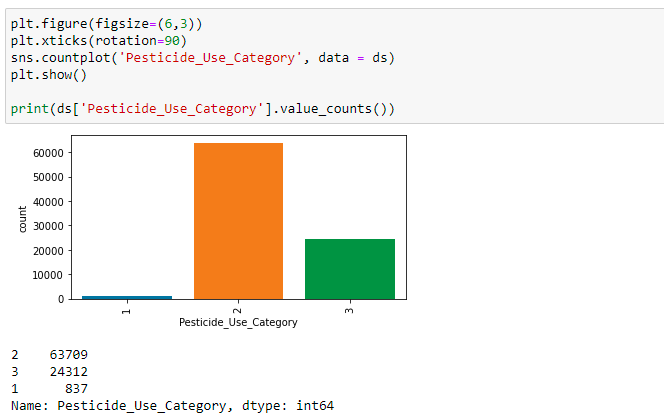
Soil Type



Observations:

1. 0 is Alluvial Soil and 1 is Black Cotton Soil.
2. As per dataset 54% of Soil is Alluvial.

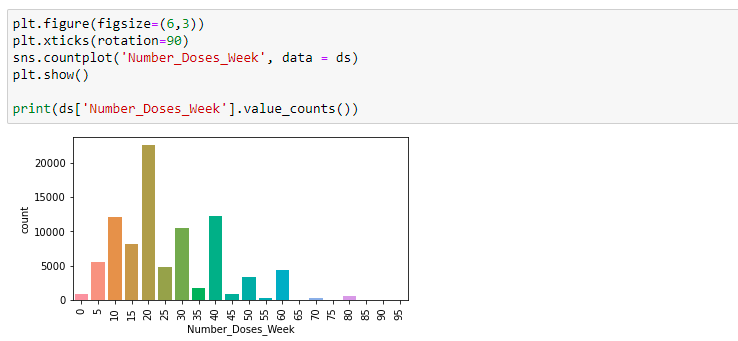
Pesticide Use



Observations:

1. 1 – Insecticides, 2 – Bactericides and 3 – Herbicides
2. From plot we can say that Bactericides are mostly used on all types of Crops.
3. Insecticides are very often used.

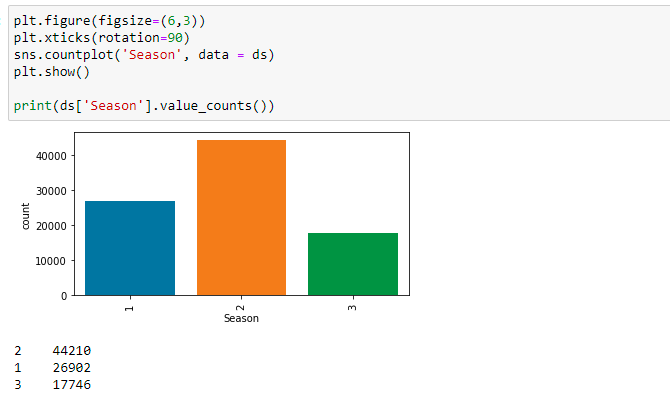
Number Doses Week



Observations:

1. Maximum 20 doses are given on crop of approx. 22000 instances.
2. Minimum is Zero.

Season



Observations:

1. There are three types of seasons in dataset namely Summer, Winter and Monsoon
2. Here 1 is Summer, 2 is Monsson and 3 is Winter
3. Monsoon instances are higher as compared to other seasons. It is ~ 50% in terms of instances.
4. Second most preferred season is Summer.

Crop Damage

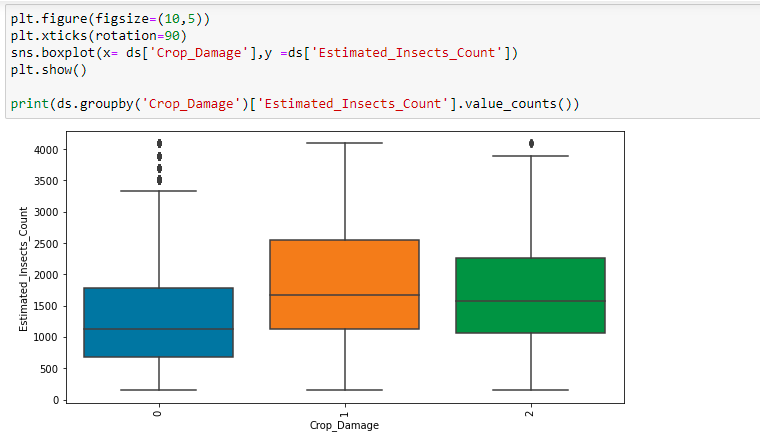


Observations:

1. There are three reasons for crop damage namely Minimal Damage, Partial Damage and Significant damage.
2. Here 0 is Minimal Damage, 1 is Partial Damage and 2 is Significant damage.
3. The instances od Minimal damage is more which comprises of ~84%.

**Bivariate Analysis with Target Variable.**

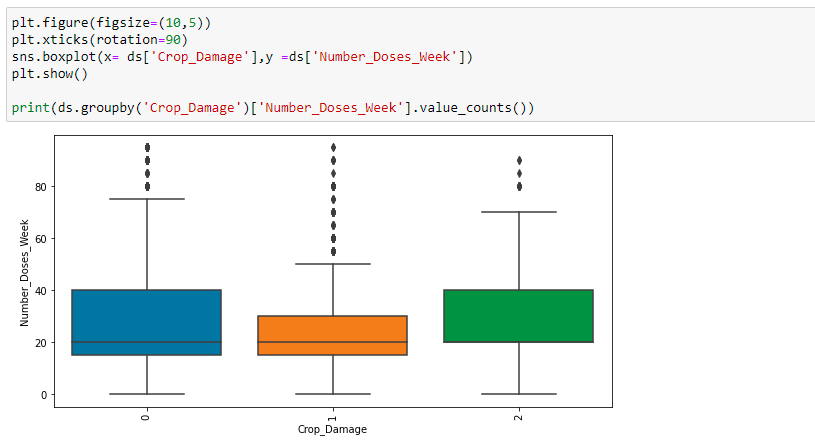
Crop Damage-Estimated Insects Count



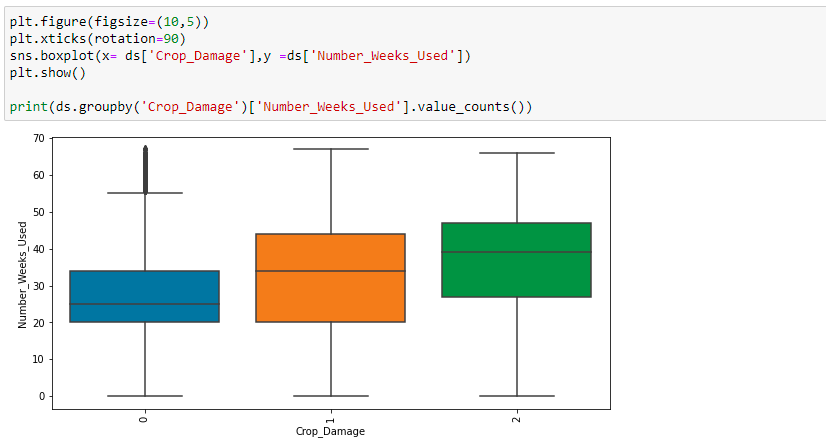
Observations:

1. 1 - Partial Damage has highest occurrence with highest estimated insects count.

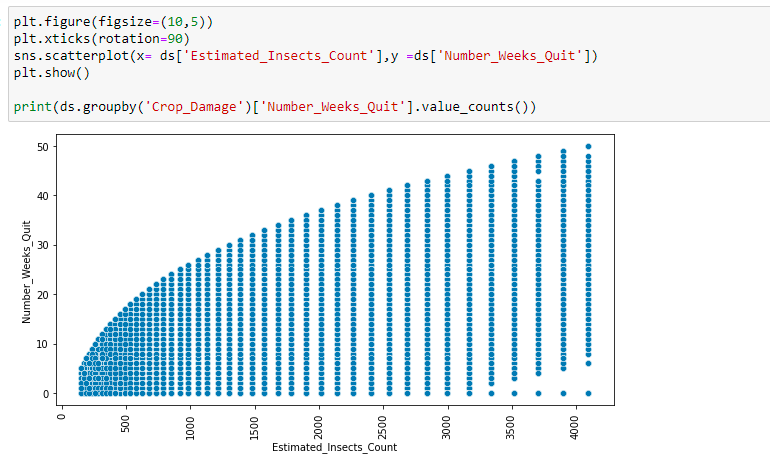
Crop Damage – Number Doses Week



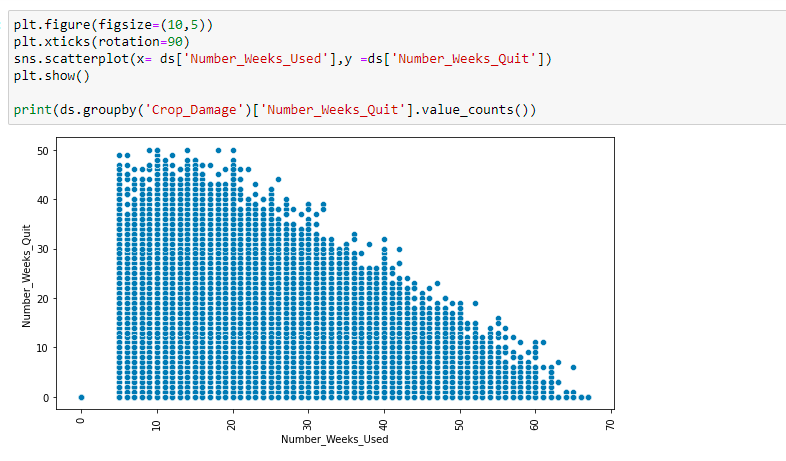
Crop Damage – Number Week Used



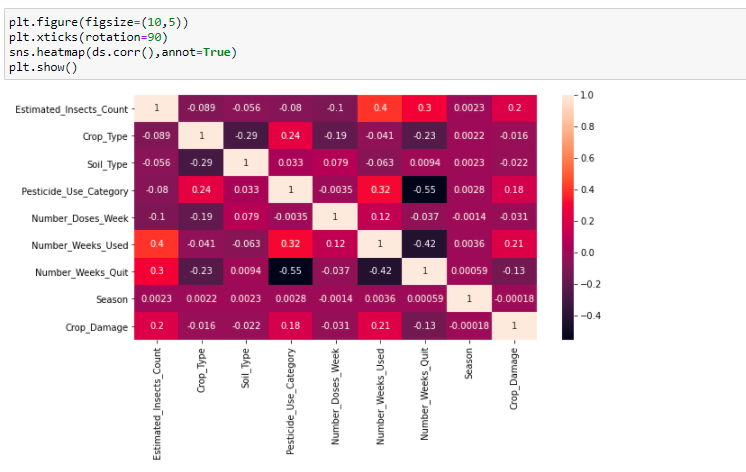
Crop Damage – Number Week Quit

Estimated Insects Count-Number Weeks Quit

Number Weeks Used – Number Weeks Quit



**Correlation Heatmap**

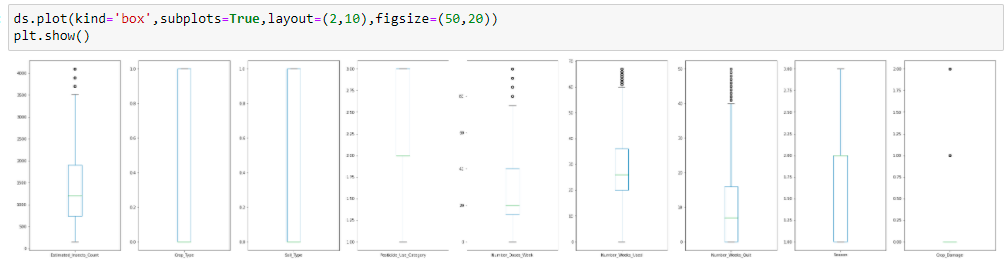


Observations:

1) Column Estimated Insects Count, Number Weeks Used, Pesticide Use Category are highly correlated with target (Crop Damage)

2) Other columns are negatively correlated with target.

**Checking Outliers**



Observations:

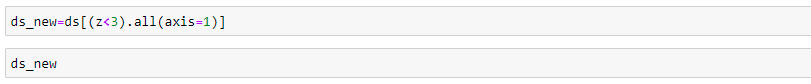
Column Estimated Insects Count, Number Doses week, Number weeks used, Number weeks quit, crop damage contains outliers.

**Removing Outliers**

Now let’s check whether we need to remove the outliers or keep the outliers

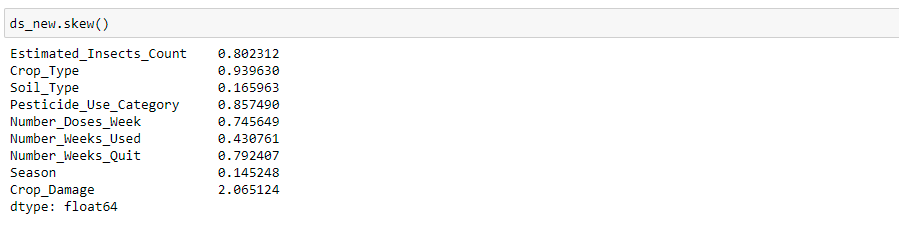






It was observed that the shape of new database was reduced to 84853, 10. Approximate loss of data is 4.5 %. So, we can remove the outliers.

**Skewness Check**



Observations:

As Crop damage skewness is = 2 and all other are within range, so we will not use any method for removal of skewness.

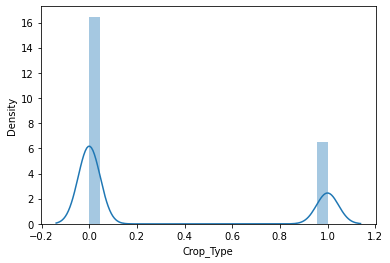
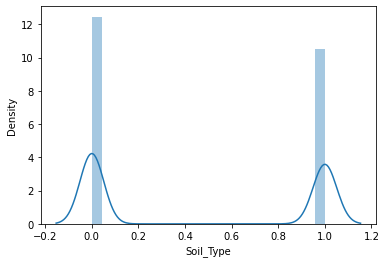
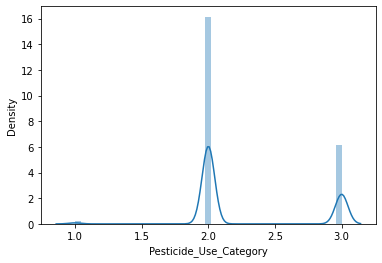
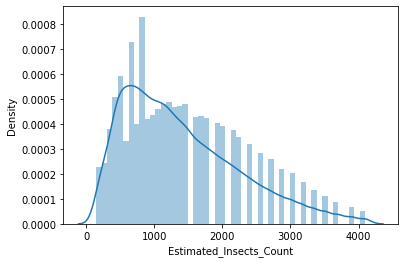
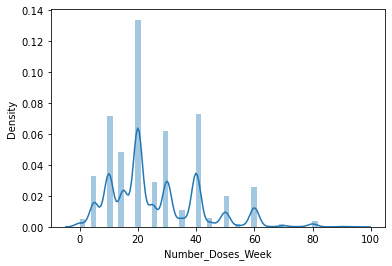
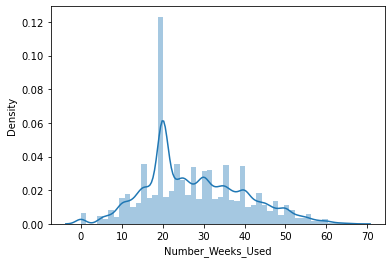
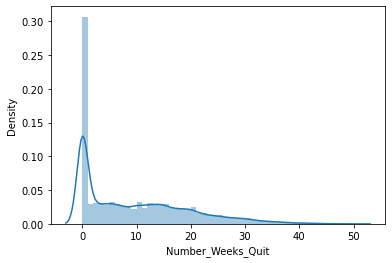
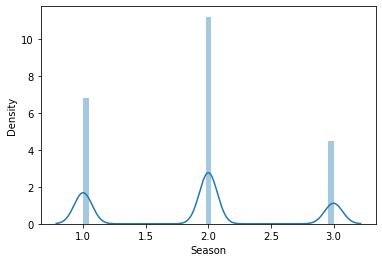
**Dropping Columns**

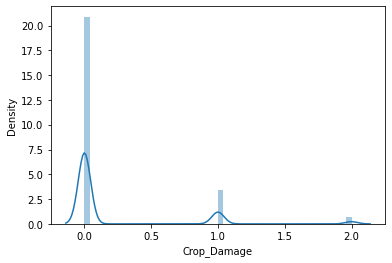
As we can see that the ID column is not required for machine Learning. It the only the identification given to the different crop.



**Distribution Plot**

Let’s plot the distribution plot of columns.

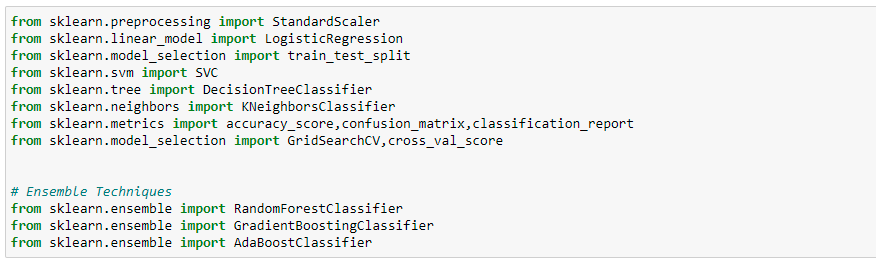




**Building Machine Learning Models**

As Crop damage is our target variable and we can see that the column has three different variables so we will use Classification Model in our Machine Learning process. We will try different models of Classification like Logistic Regression, SVC, Decision Tree Classifier, KNeighbors Classifier, etc. We will also use Ensemble techniques like Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier.

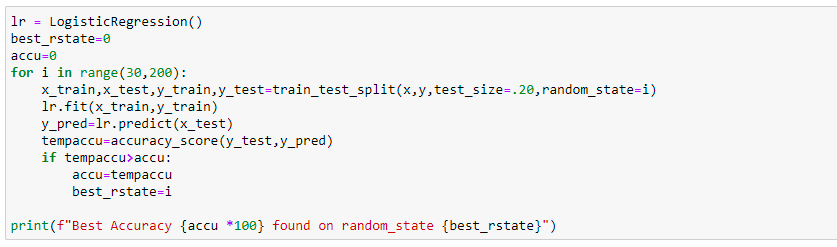
**Importing Machine Learning Algorithms**



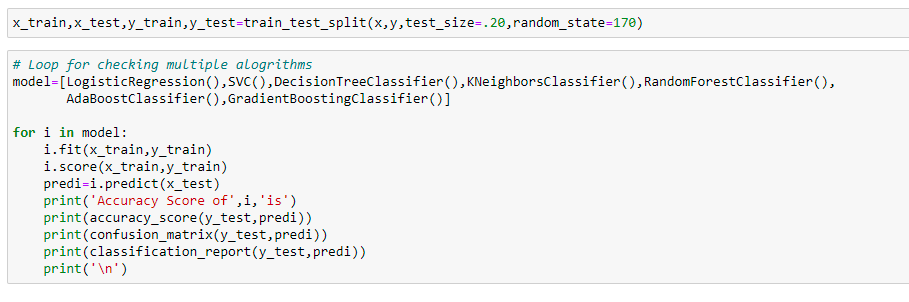
Let’s define x and y variables where y is the target variable. After that standard scaler will be used to bring the dataset in a similar scaled format.



Now we will pass multiple algorithms in a loop to check which model works perfectly on our dataset. Initially we will find the best random state and accuracy. Next step will be to perform cross validation score to control overfitting/underfitting. Then perform Hyper Parameter tuning using GridSearchCV on best model to check if we can further improve our score.



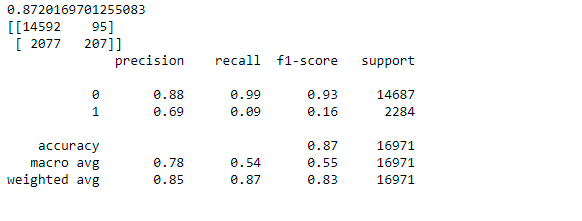
It was observed that the best accuracy of 86.32% at random state 170 was achieved. Now we will define x\_train, x\_test, y\_train, y\_test by taking test size of 20%.



After passing the above loop for different classification model to find which model gives us best accuracy score.

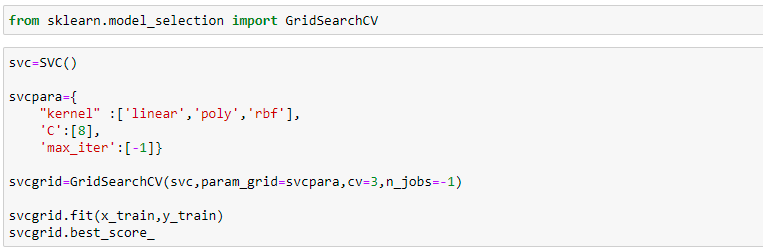
In classification model we will look 3 parameters namely accuracy score, confusion matrix and classification report.

We have observed that the Support Vector Classifier is giving us best accuracy score of 87.20 %. So, we will select it as our best model and we will use Grid SearchCV on Support Vector Classifier model to see if we can further improve our score.

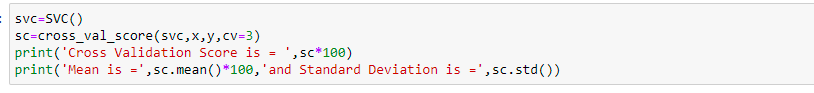


It was found that Support Vector Classifier is giving us accuracy score of 87.20 %.

**Hyper Parameter Tuning using Grid Search CV**



**Cross Validation Score**

****

Cross Validation Score is = [86.49460845 86.55070004 86.52595107]

Mean is = 86.52375318662554 and Standard Deviation is = 0.00022951974562464815

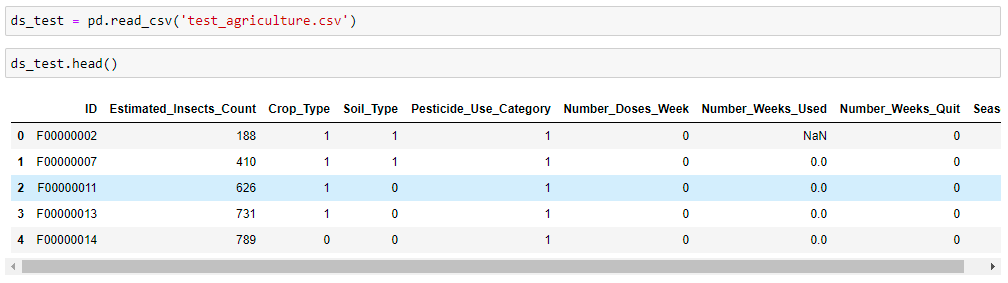
We can see that cross validation is also giving good accuracy.

**Saving Best Model**



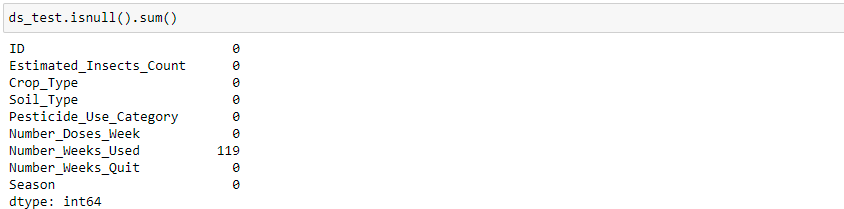
We have saved our model in pickle form and write binary mode. We will load this model later after we perform similar operations on test dataset as done in training dataset. After that we will pass the test dataset on model to get our predictions.

**Loading Test Dataset**



Now we have loaded our test dataset. We can see that our test dataset contains 1199 rows and 9 columns but it doesn’t have crop damage column with is our target. We need to predict this using our saved model.

**Checking Missing Values**



We can see that Number weeks used has 119 missing values. We will fill these values with the mode of column as done in training dataset.

**Handling Missing Values**

****

We have replaced the missing values with the Mode of column.

**Dropping Columns from Test Dataset**

****

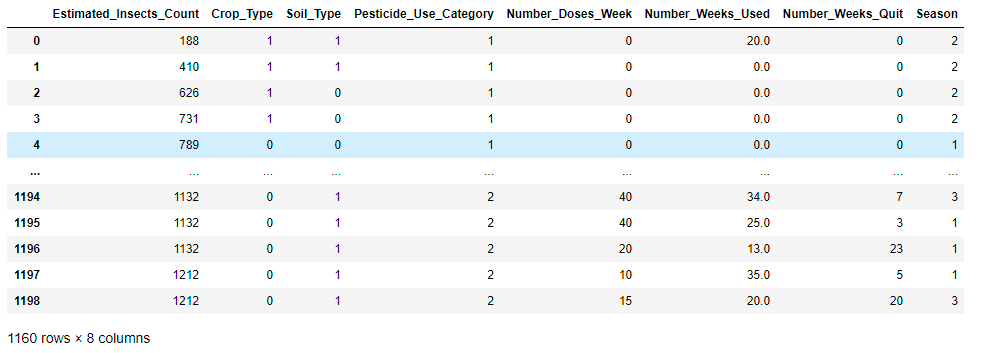
We have dropped the ID column from test dataset as well.

**Removing Outliers from Test Dataset**

****

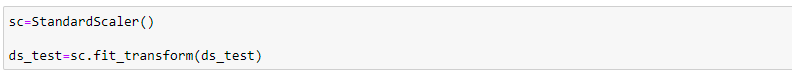
****

****

****

Approx 3% of data is lost after removal of outliers. Now our test data looks like this.

**Scaling Model**

****

Now we have scaled our test dataset. We are now ready to import our saved model to get predicted results.

**Loading Saved Model**

****

**Saving Predicted Results in Dataframe**

****

We have got our predicted results.

**Conclusion**

Our objective in Agriculture dataset is to predict whether the crop damage using Machine Learning. We have found many insights from various plots, summary statistics, etc. Completed EDA on both training and testing dataset.

We have seen that Support Vector Classifier was working perfect on current dataset to predict crop damage. Then after saving our best model using pickle library.

After running passing model on test dataset, we have finally predicted the crop damage from our test data.