**ML in Agriculture**

Subtopics:

• Introduction

• Problem Definition

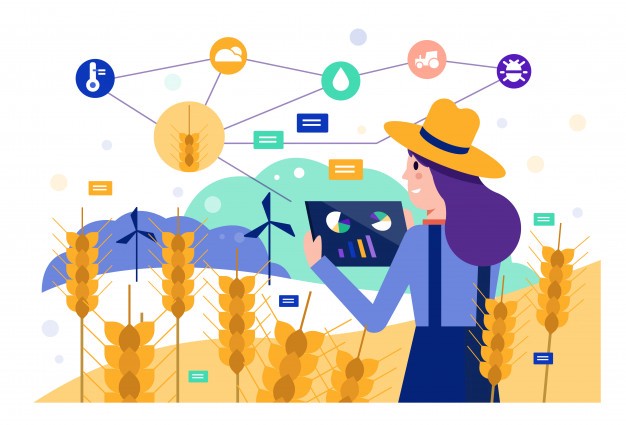
• Data Analysis

• EDA Concluding Remarks

• Pre-Processing Pipeline

• Building ML Model

• Concluding Remarks



1. **INTRODUCTION**

Machine Learning is everywhere, it is used almost in every field to have better insights and output of particular project/ work. From knowing the behavior of customers coming in the shop to predicting the GDP of the country, it is widely used with other AI tools. Similarly, in the very ancient field called farming AI is been used from the soil preparation, seeds breeding and water feed measurement. Farmers are using AI and ML techniques to have better understanding and better results of their harvests.

Agriculture is how humans started to mold nature according to their needs. Humanity may have come a long way since then, changing the ways and methods of farming but the objective remains the same; improving the quality of crops.

With the continuous increase in global population and land becoming scarcer, there is a need for farming to become more efficient with fewer resources. The AI and ML tools are much helpful in providing better solutions to get more quality crops than ever. Farmers are in a better position to decide on various factors of farming like which crop to harvest in which season, how much amount of water to be feed to particular crop or in a particular season or soil, decision on usage of pesticides etc.

In this article, I will be using a use case of Agriculture dataset to show how we can harness the power of ML and apply it on real world problems in the field of agriculture. Here, we will predict the harvest as per the usage of pesticide.

**Dataset (Link of dataset used for your hands on practise)**

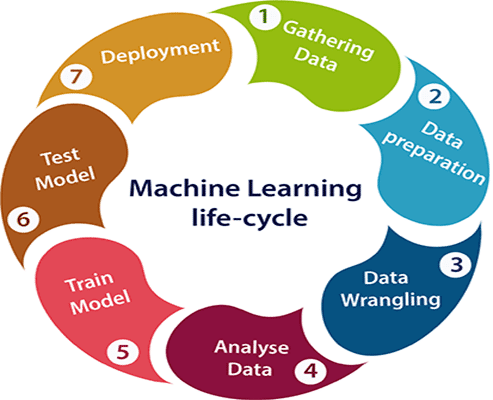
**Training set**

<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/train_agriculture.xlsx>

**Testing set**

<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/test_agriculture.xlsx>

Pesticides are also special, because while 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.

****

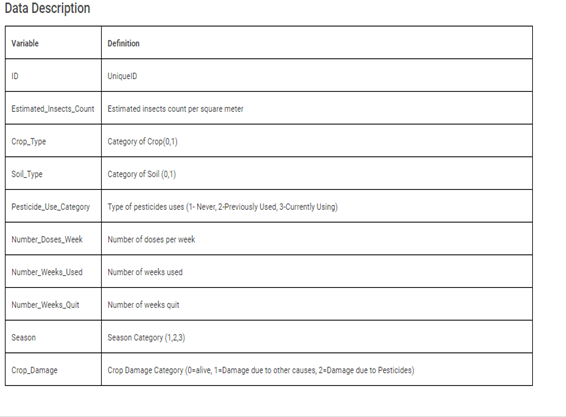
**Life cycle of Machine Learning is well explained above, we will be using the same standard steps in our dataset to get the valued predictions.**

1. **PROBLEM DEFINITION**

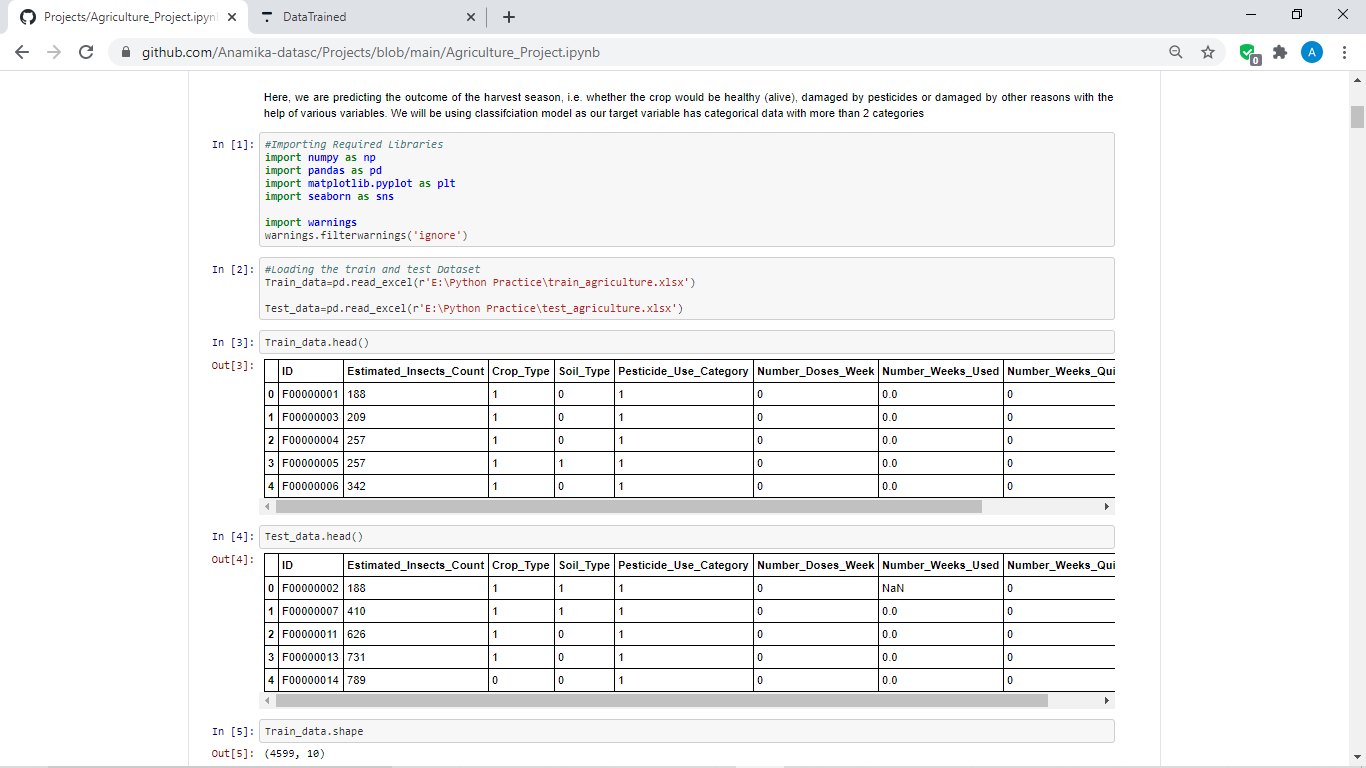
Here, we need to determine the outcome of the harvest season, i.e. whether the crop would be healthy (alive), damaged by pesticides or damaged by other reasons with the help of various variables (refer below table). We will be using **Classification** **Model** as our target variable has categorical data with more than 2 categories.

Here, our target variable is **Crop\_Damage**.

Description of various variables in this dataset are as below:



Before moving ahead with data analysis, let’s import some necessary libraries.



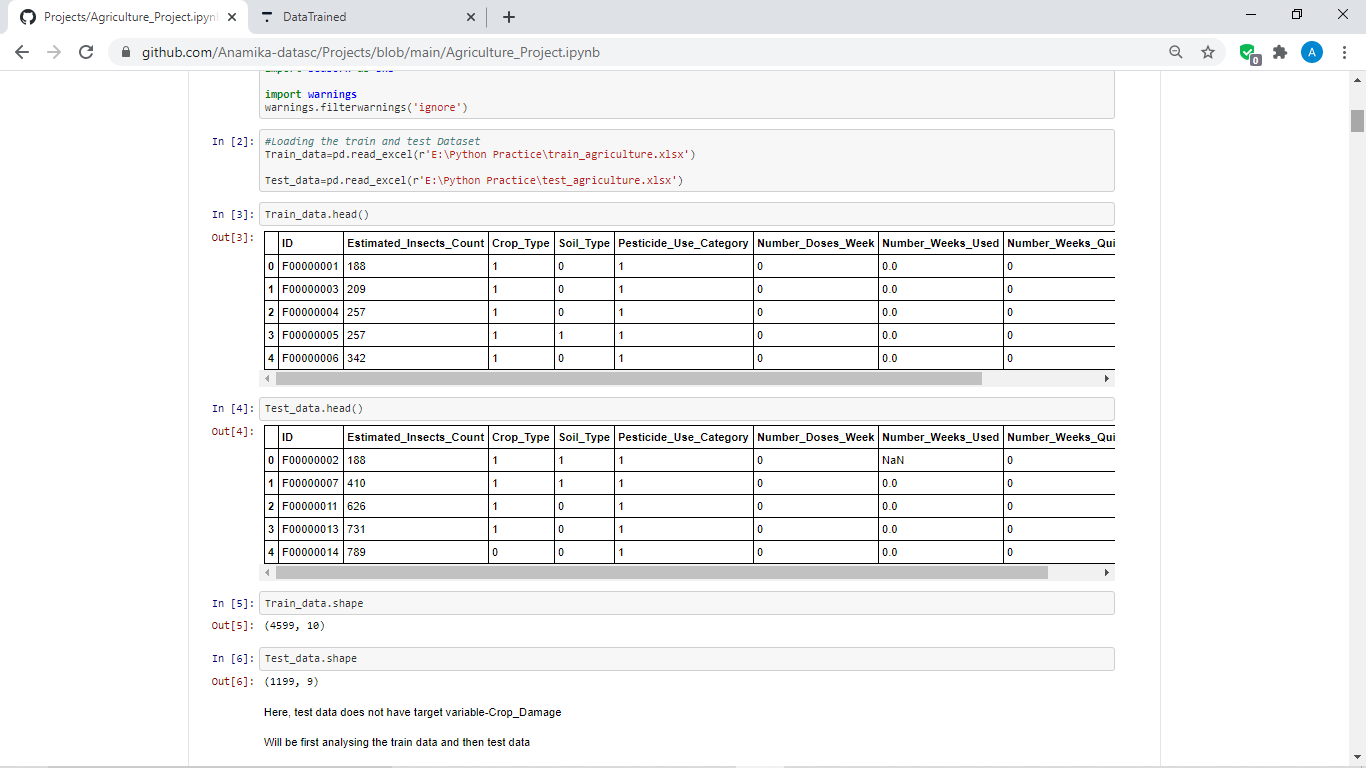
We have imported statistics libraries like **numpy** and **pandas** and visualization libraries like **matplotlib** and **seaborn**. Also, we have imported warnings to avoid warnings that comes after particular codes (However, it’s an optional step).

1. **DATA ANALYSIS**

Now, Let’s do some data analysis to gather more and more information about the data like how many rows and columns are there, is there any null values present in the data, what are the types of data, to check the skewness in the data, to ascertain the outliers and to work on them, to know the correlation of the variables with the target variable, will also be checking the mean, median, min and max range of the variables which will ultimately be telling us the skewness and the presence of outliers in the data.

After knowing all the above insights about the data, will do some data preparation or data pre-processing before passing it to our model to have better interpretations and predictions.

Have done the analysis using following steps:

* After importing the libraries will import data- train and test data both.
* From above, we have observe, the test data does not have target variable-Crop\_Damage which means on the basis of training and testing on train data, we will be predicting the outcome of the test data. In other words, we will be using the trained and most accurate model to determine the outcome in the test data.
* Will be first analysing the train data and then test data also.
* After verifying the Train data, following observations has been noted

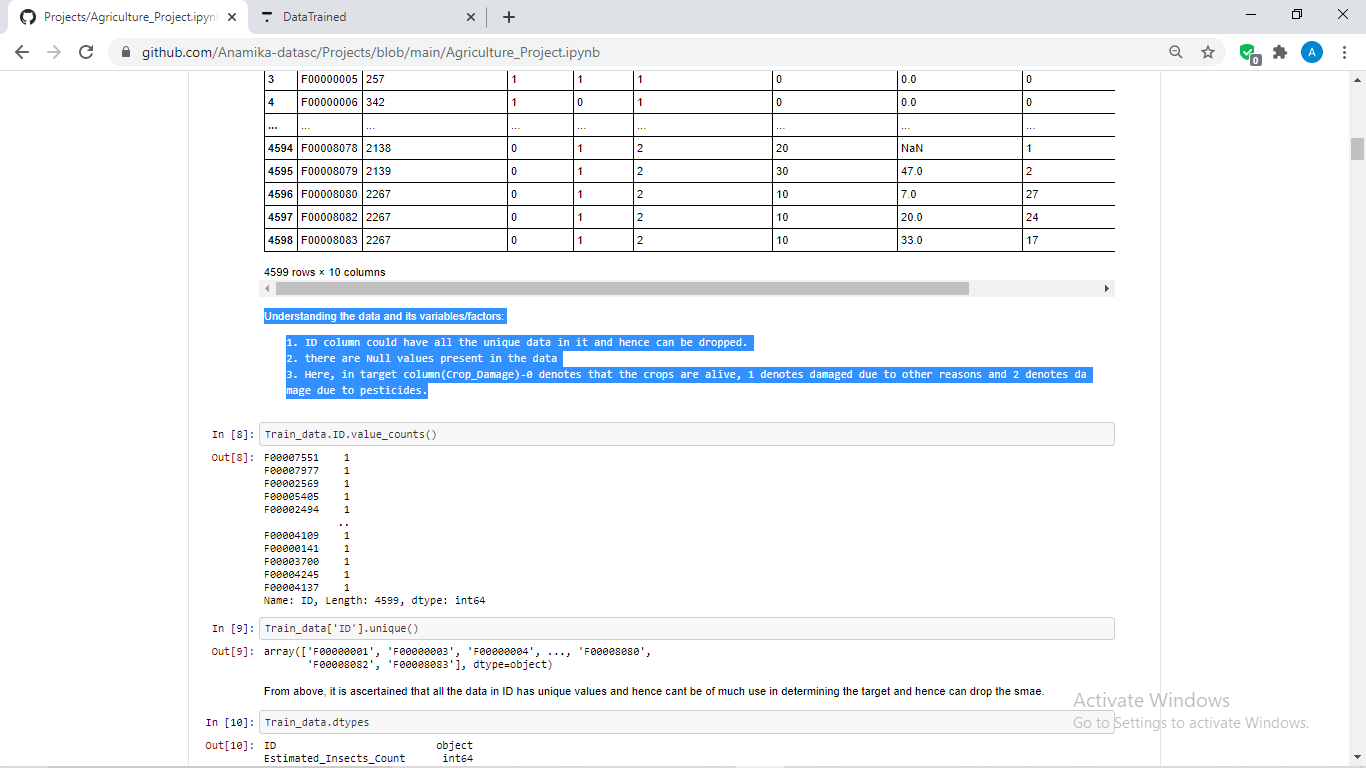
1. ID column could have all the unique data in it and hence can be dropped.

2. There are Null values present in the data

3. Here, in target column (Crop\_Damage)-0 denotes that the crops are alive, 1 denotes damaged due to other reasons and 2 denotes damage due to pesticides.

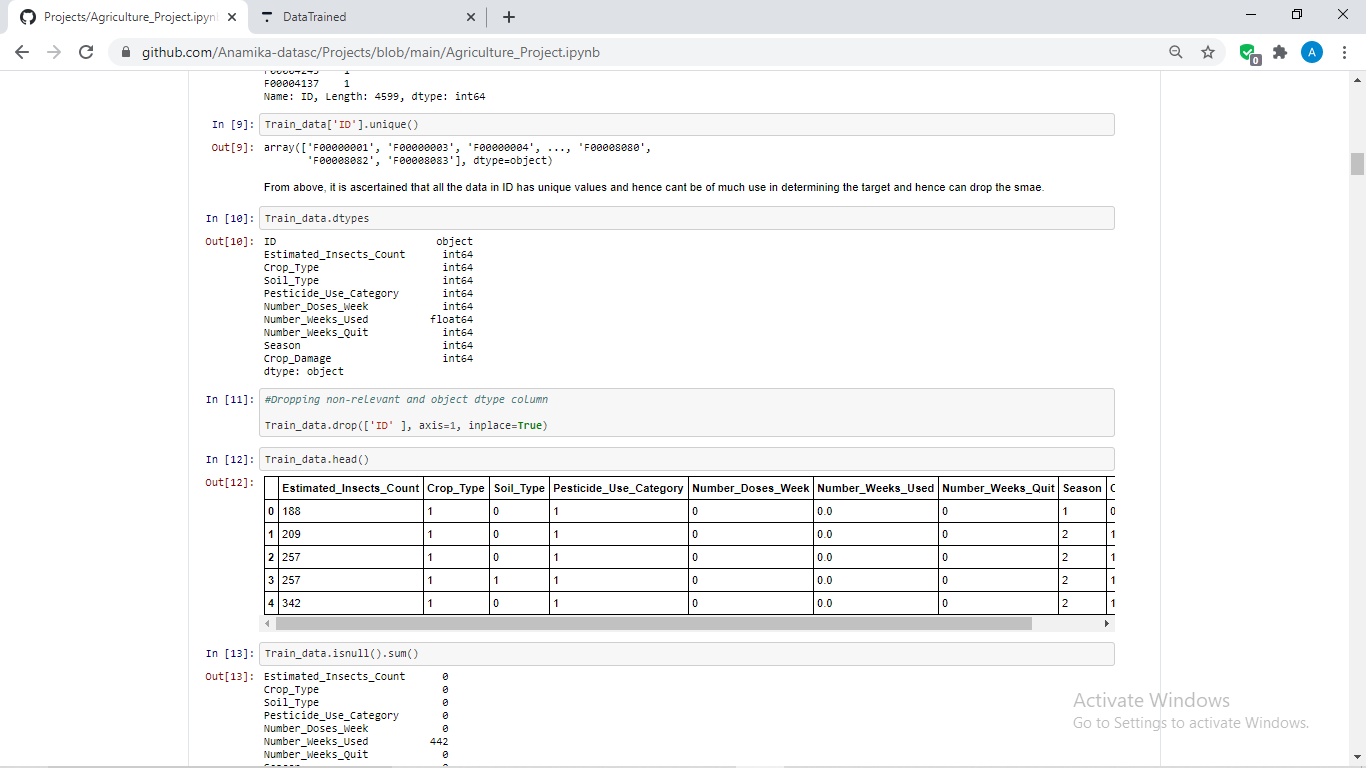
* We have checked the uniqueness of the ID column using below codes and found all the data

present are unique and hence, it will be of no use for analysis further. Therefore, will drop that column.



* Now, checked the data type as below and found ‘ID’ column has only have object data type and we have also checked above that this column has unique data in it and hence, will drop the

same before moving forward:



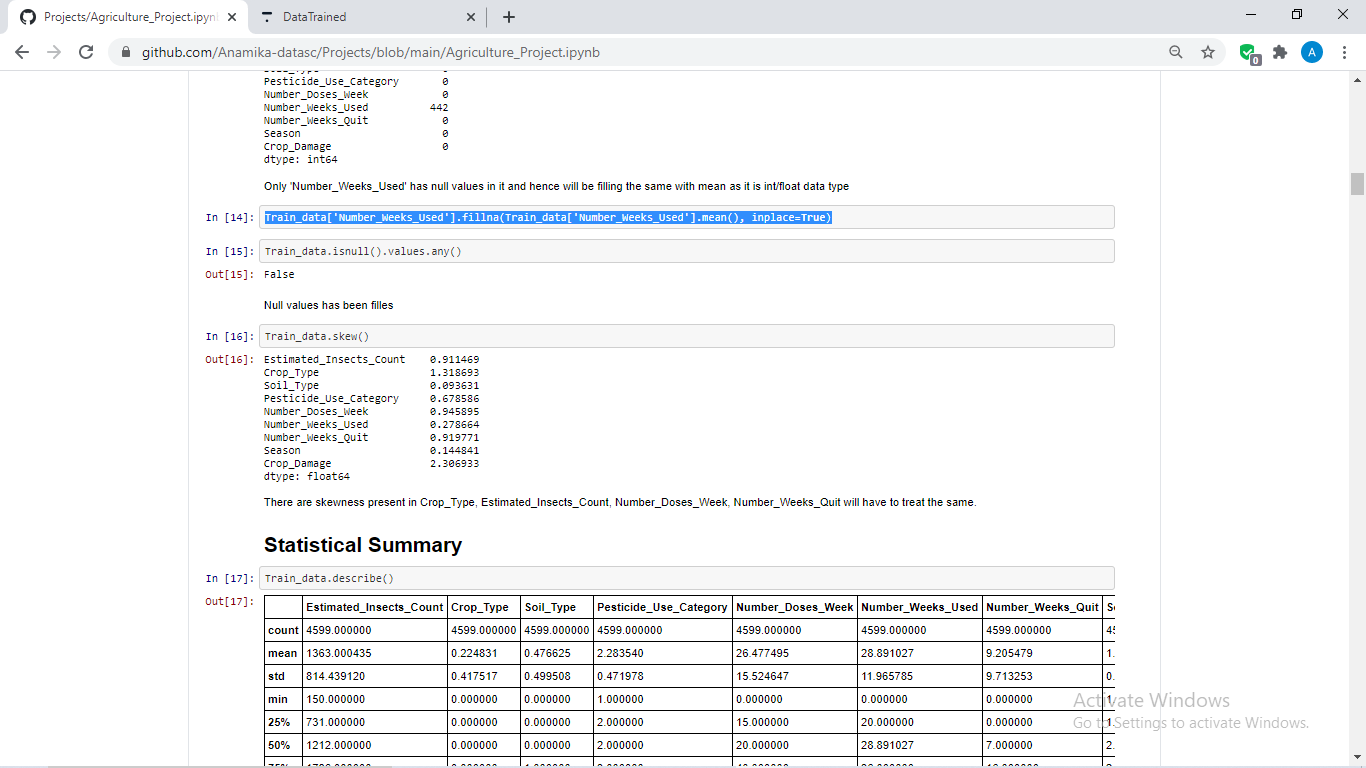
* Further, we have checked the null value presence in the data, after checking the same found ‘Number\_Weeks\_Used’ had null values in it. Hence, we have filled the null value using fillna method as per below code:

Train\_data['Number\_Weeks\_Used'].fillna(Train\_data['Number\_Weeks\_Used'].mean(), inplace=**True**)

Note: We fill the NaN values in this step itself because keeping that could hamper our visual analysis.

* Further, checked the skewness and found presence of the same as follows which we will treat

on later stage:



# Now, it’s time for Statistical Summary which is a very important and useful step which will tell various aspects about the data.

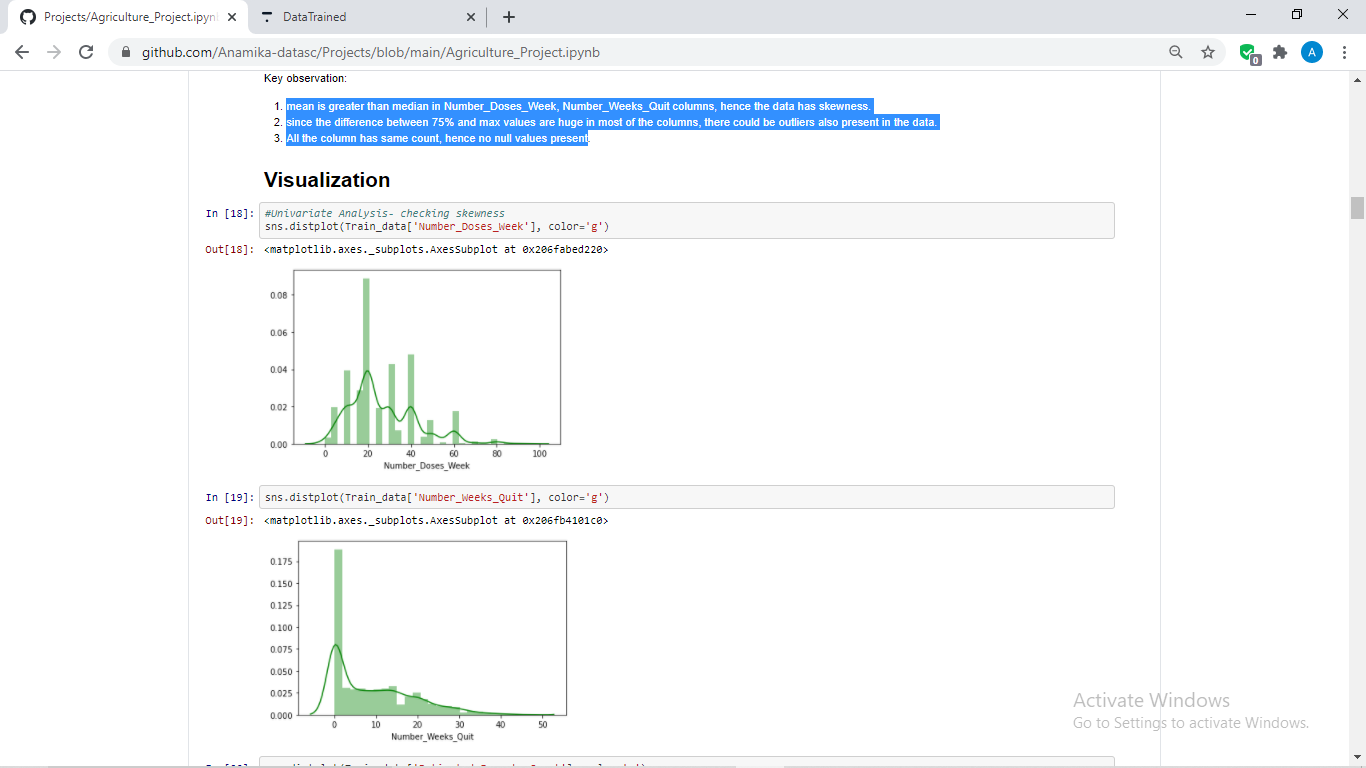
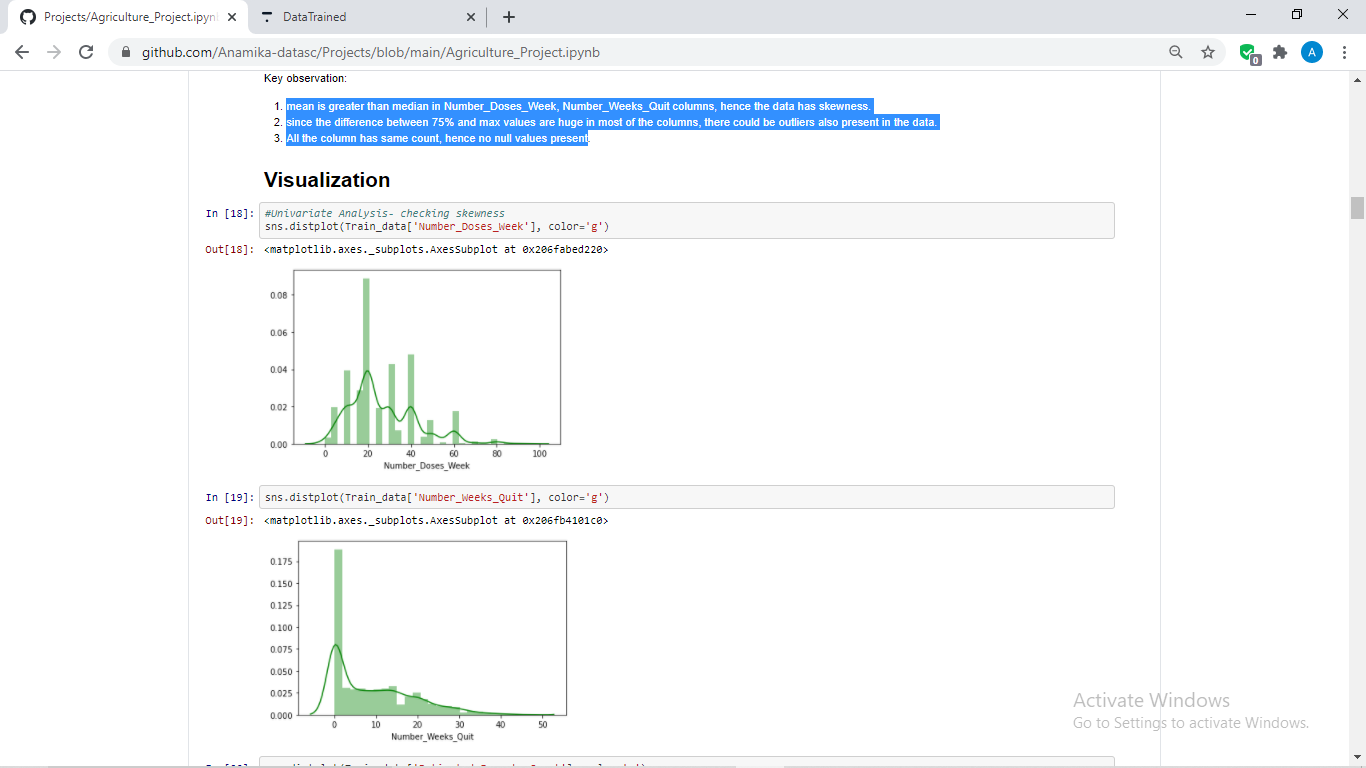
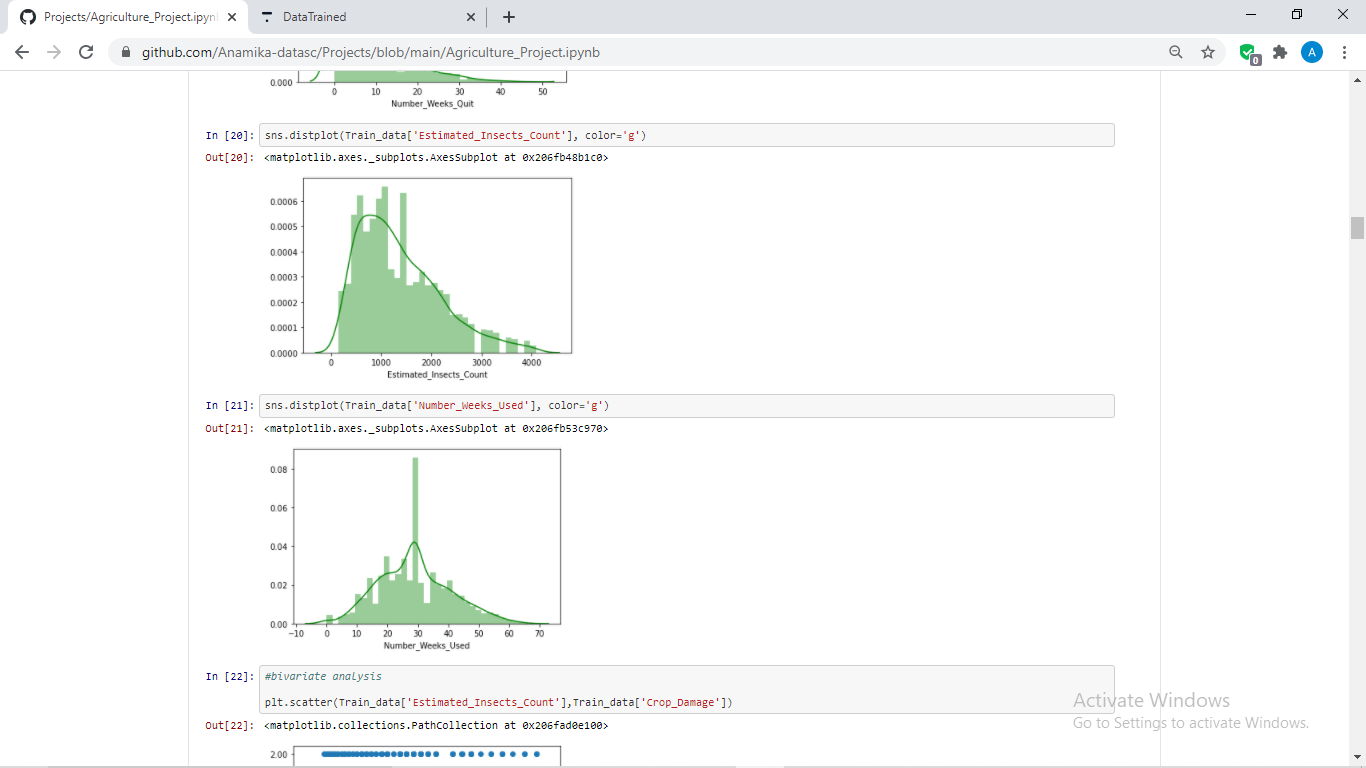
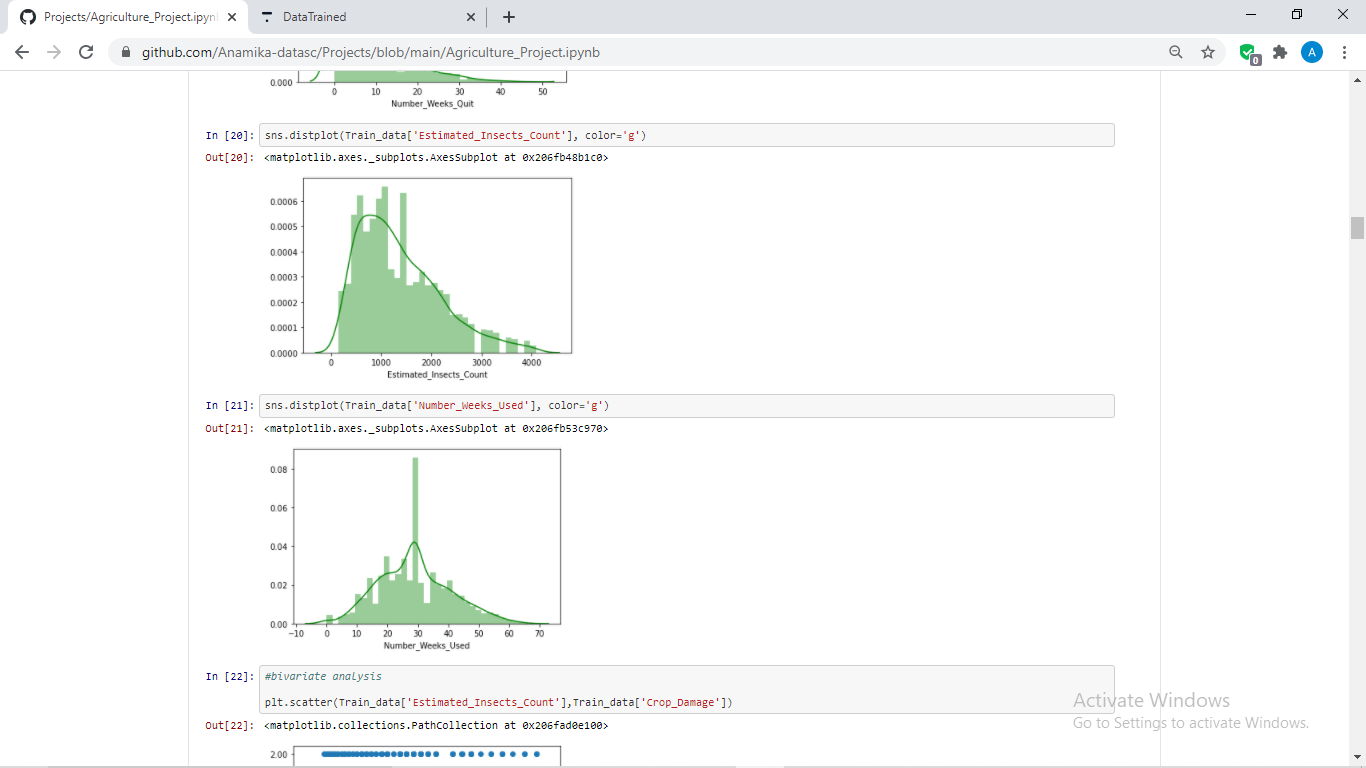
# 

# From Statistical summary, we have observed the following:

1. Mean is greater than median in “Number\_Doses\_Week”, “Number\_Weeks\_Quit” columns, hence the data has skewness.
2. Since the difference between 75% and max values are huge in most of the columns, there could be outliers also present in the data.
3. All the column has same count, hence no null values present now.

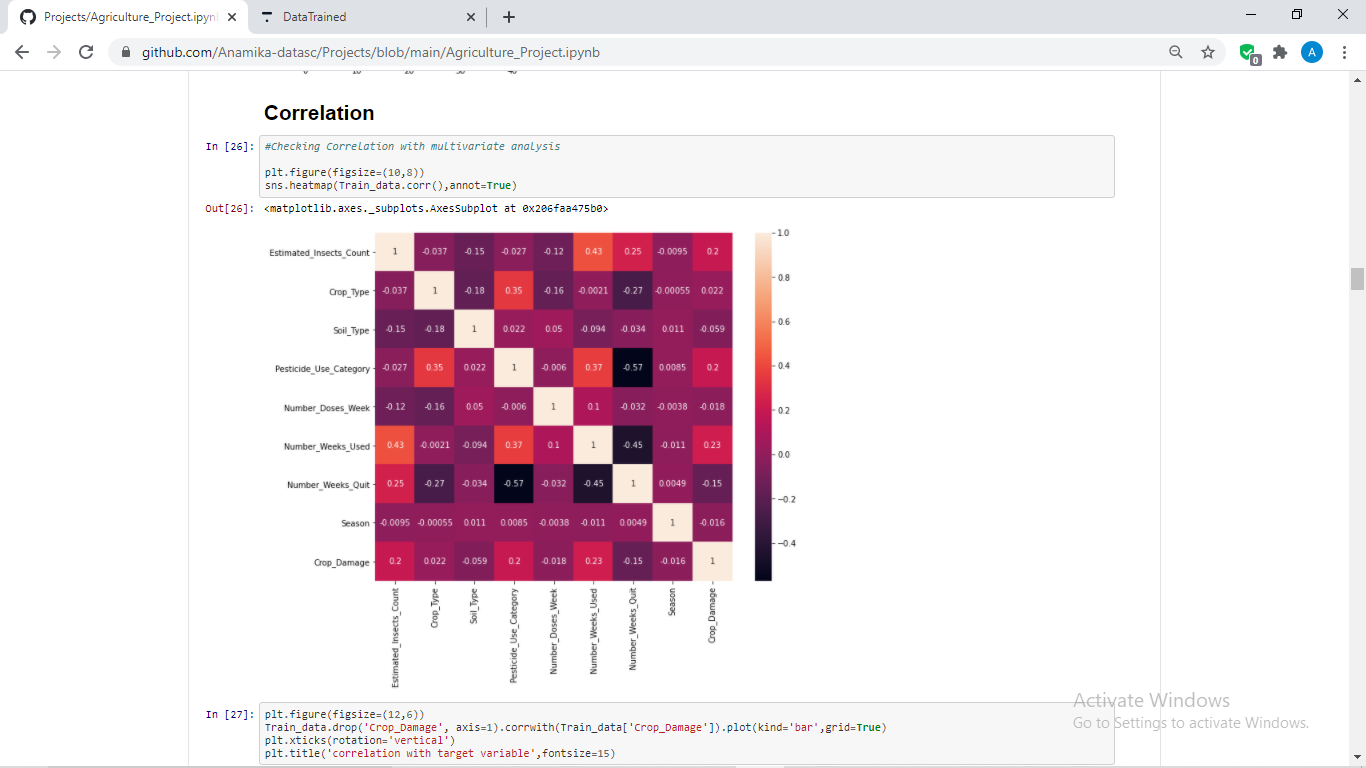
**EDA Through Visualization**

* Now, will check the above observations which we have extracted with the help of data analysis through **visualization** methods:

From above, it can be seen that the data does not have normal curve and has skewness in them which needs to be treated.

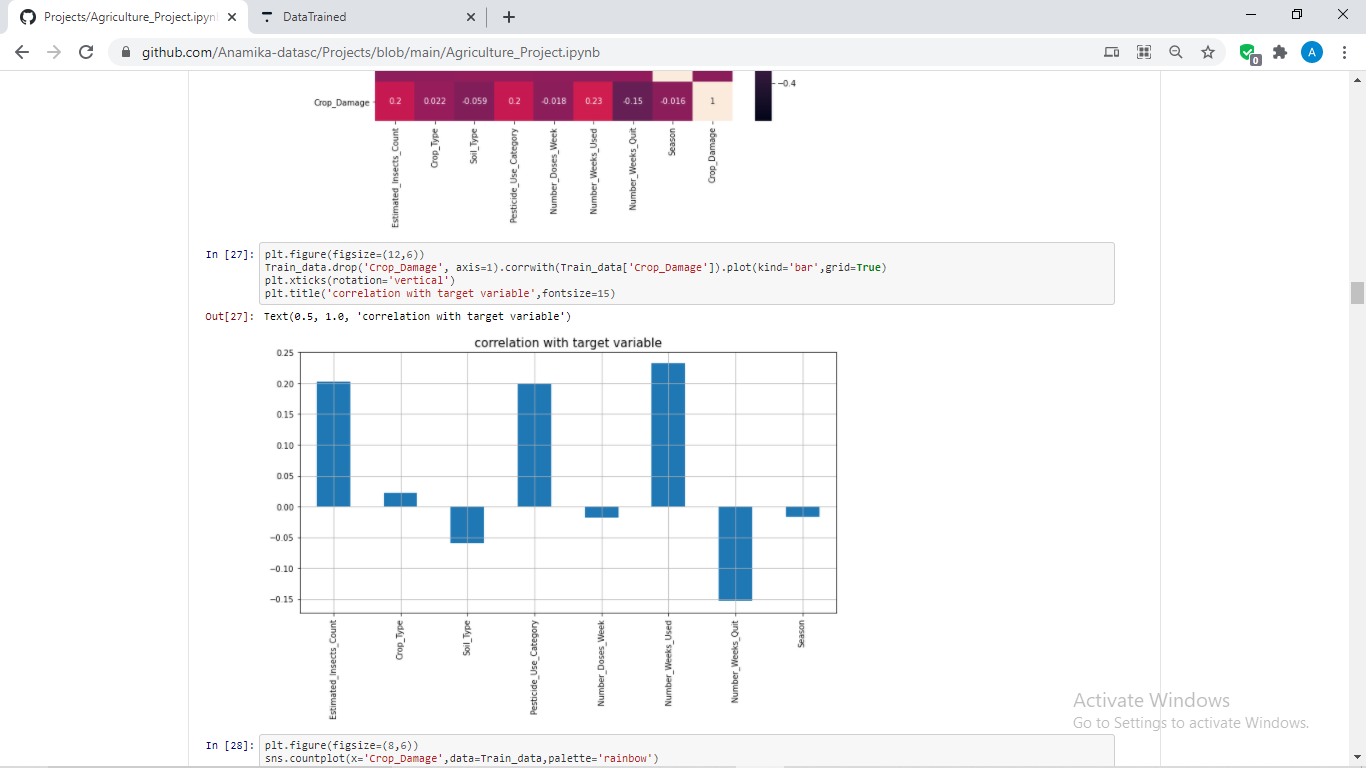
* Checked **Correlation** using heatmap



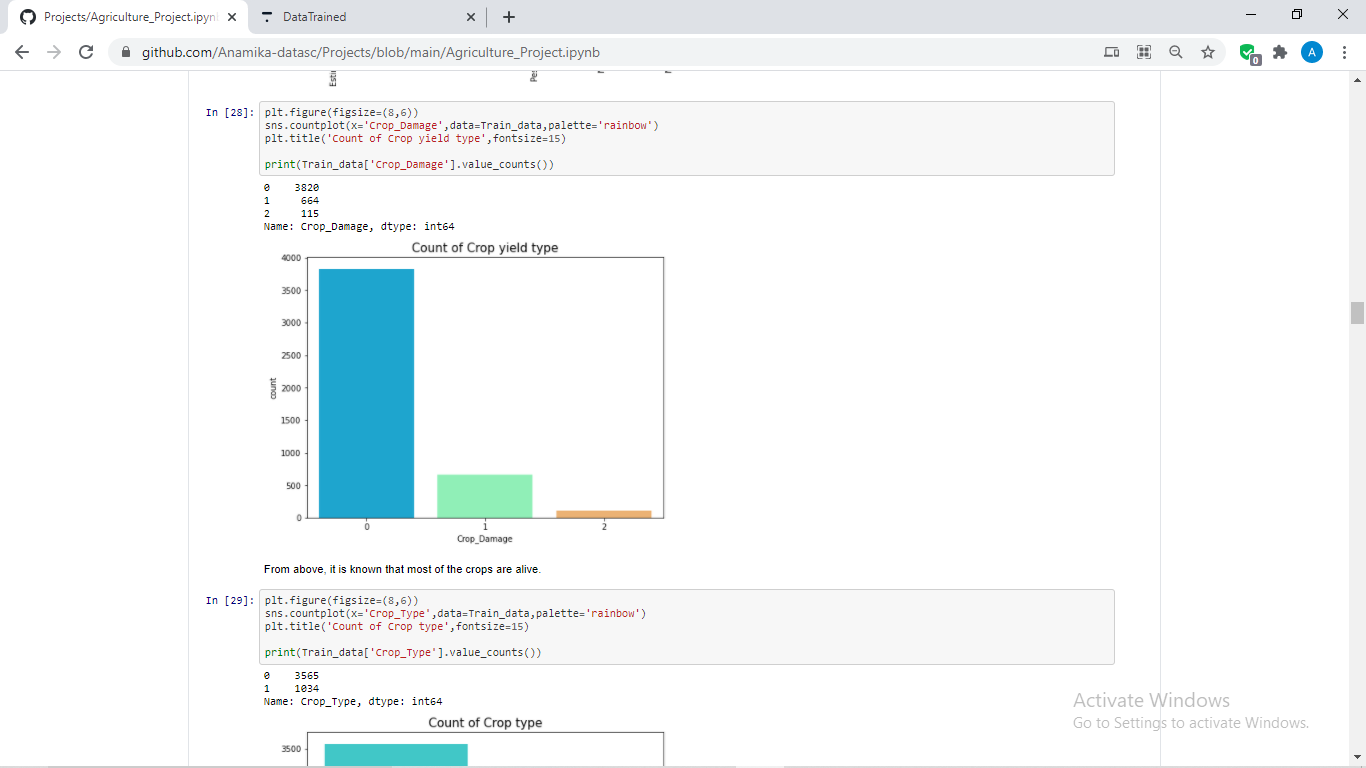
From the above heatmap, we have noted the following correlations:

* 1. Target has highest correlation with **Number\_Weeks\_Used** column followed by **Estimated\_Insects\_Count** and **Pesticide\_Use\_Category**.
  2. Target has negative correlation with **Number\_Weeks\_Quit**
  3. **Crop\_Type**, **Soil\_Type**, **Number\_Doses\_Week** and **Season** has very less correlation with the target variable.

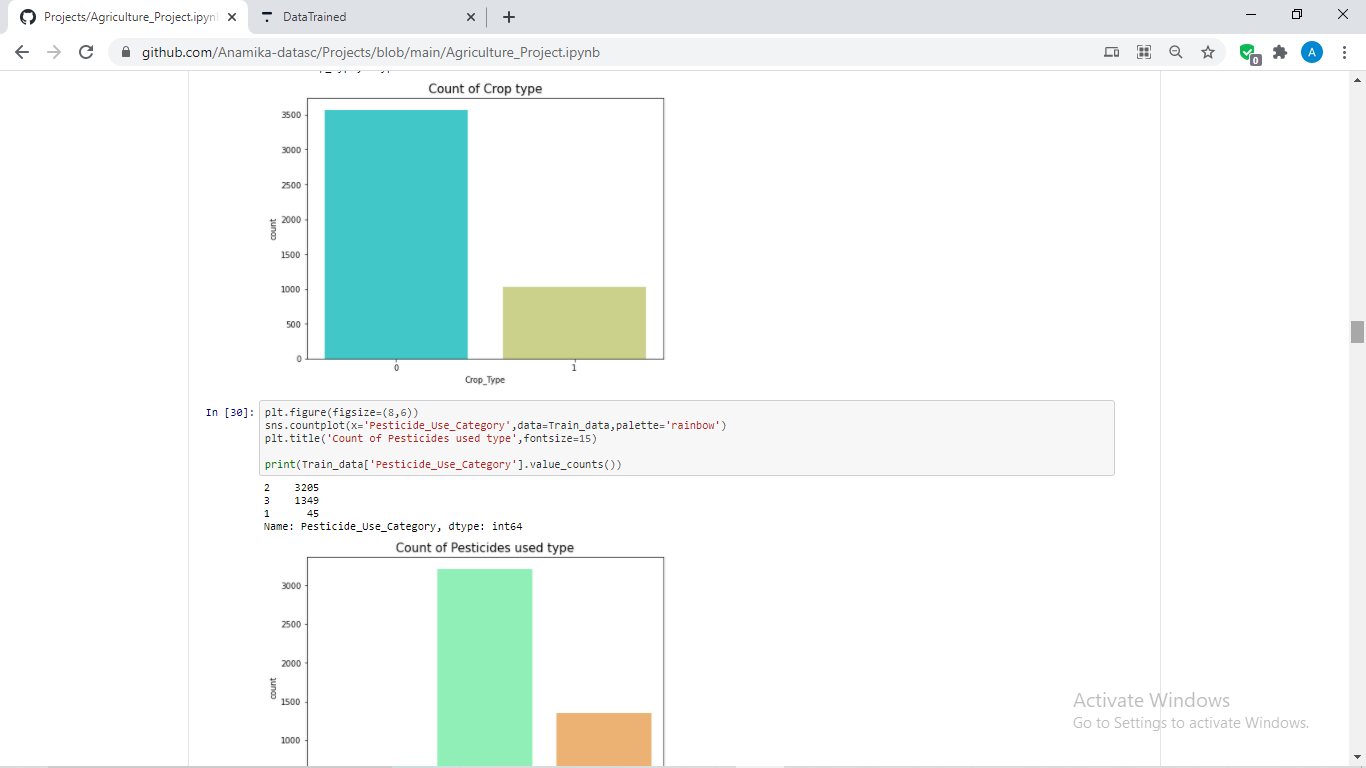
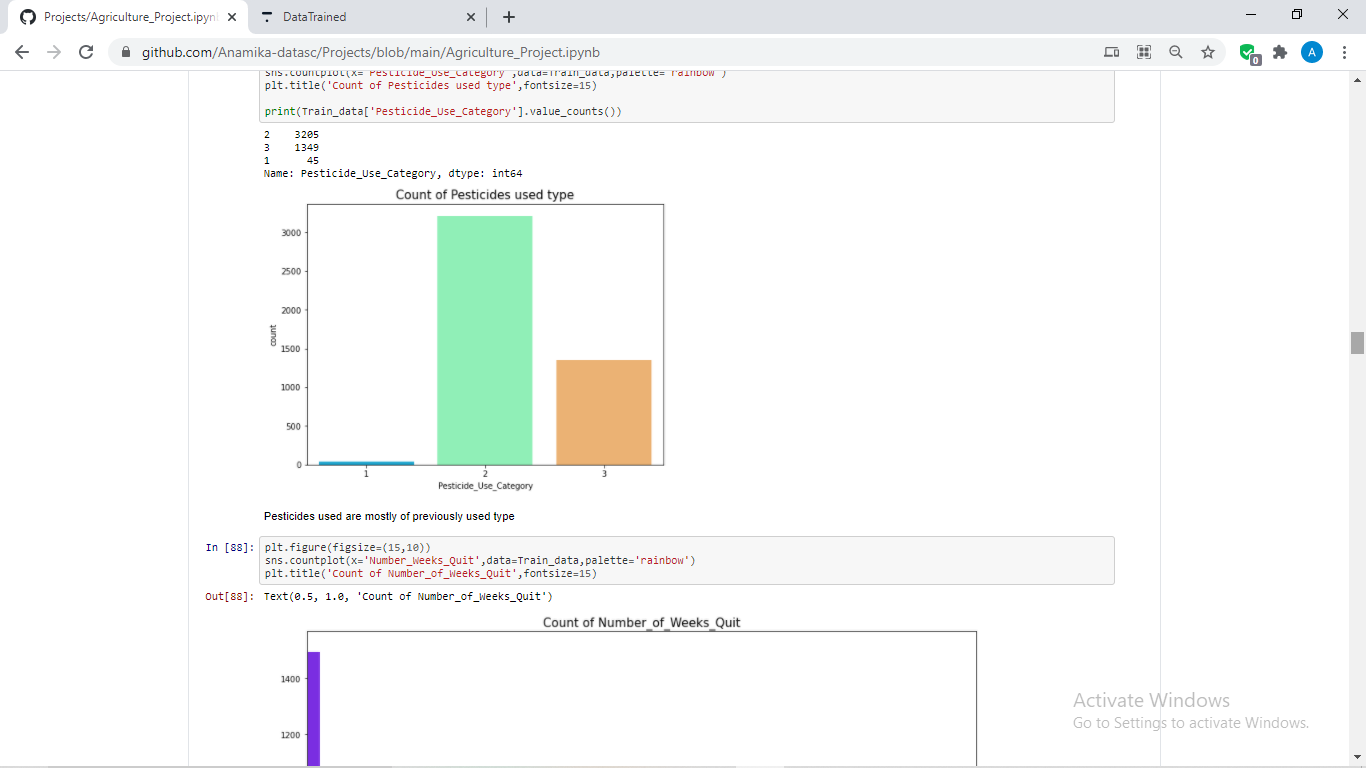
We have noted the above correlation using barplot as well.



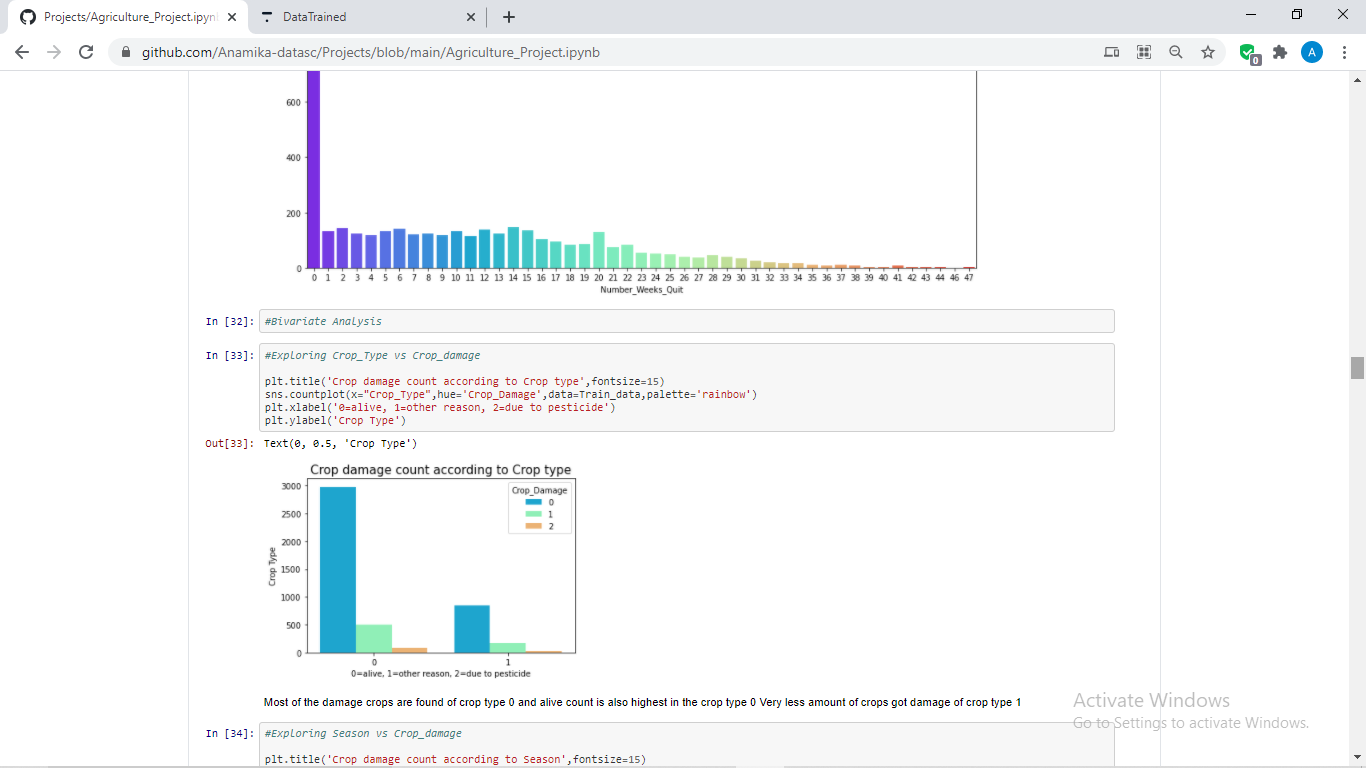
* Checked few **counts** to know the data distribution.



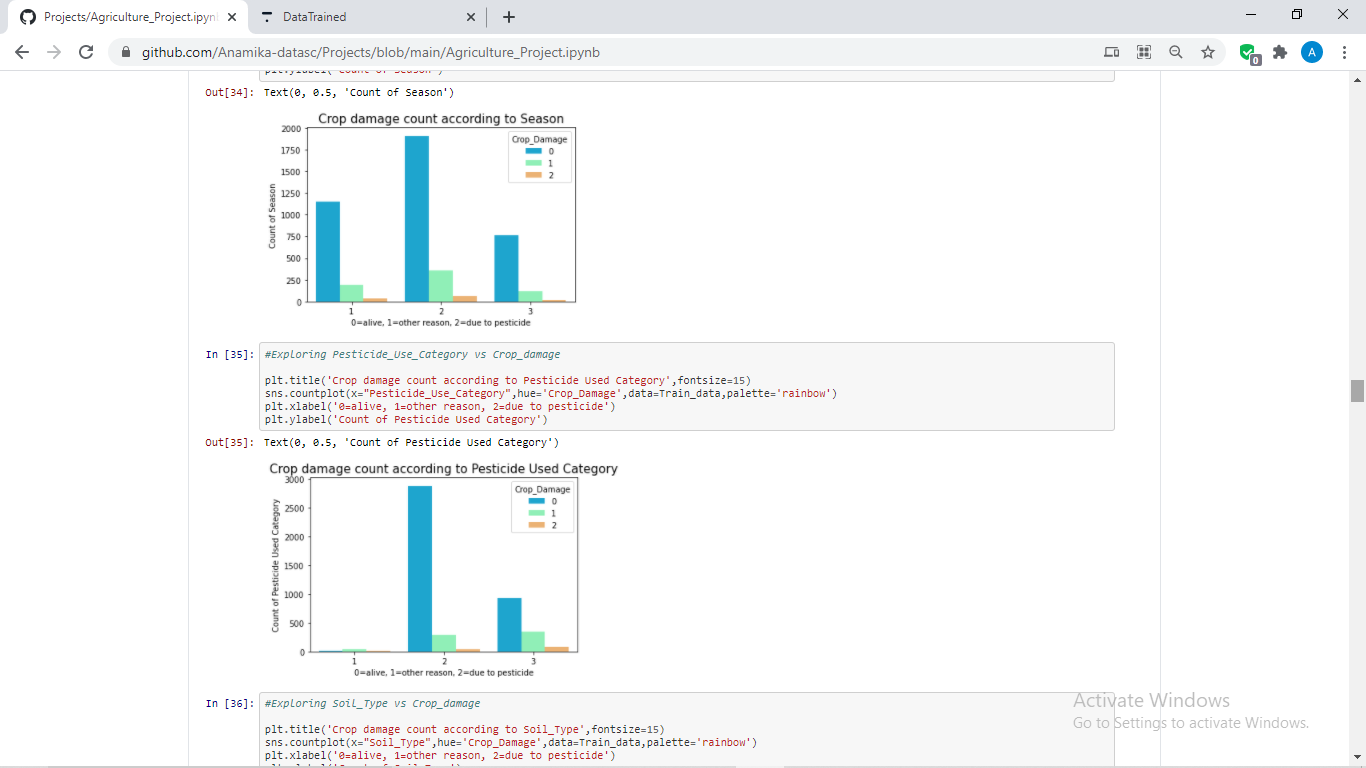
From above, it is understood that most of the crops are alive.

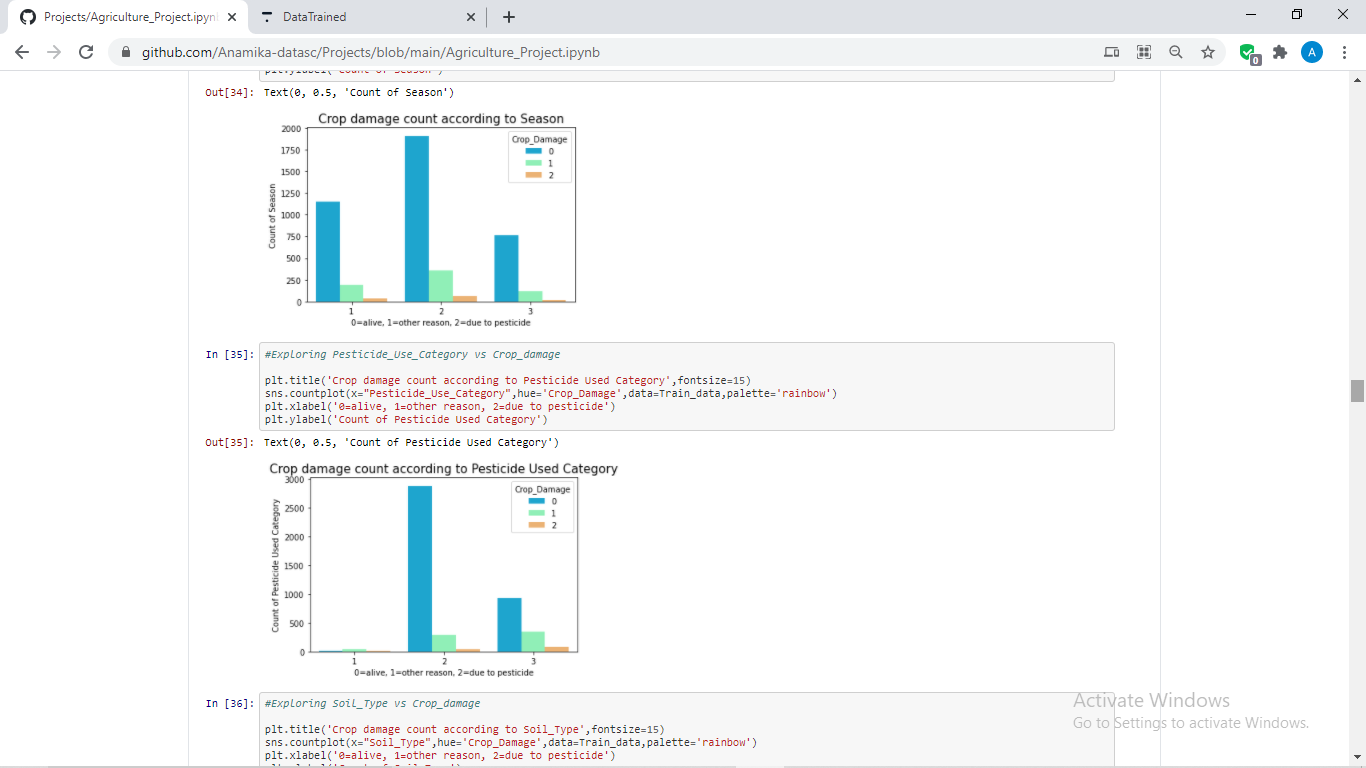
* Further, used **bivariate analysis** to do some comparisons like below:



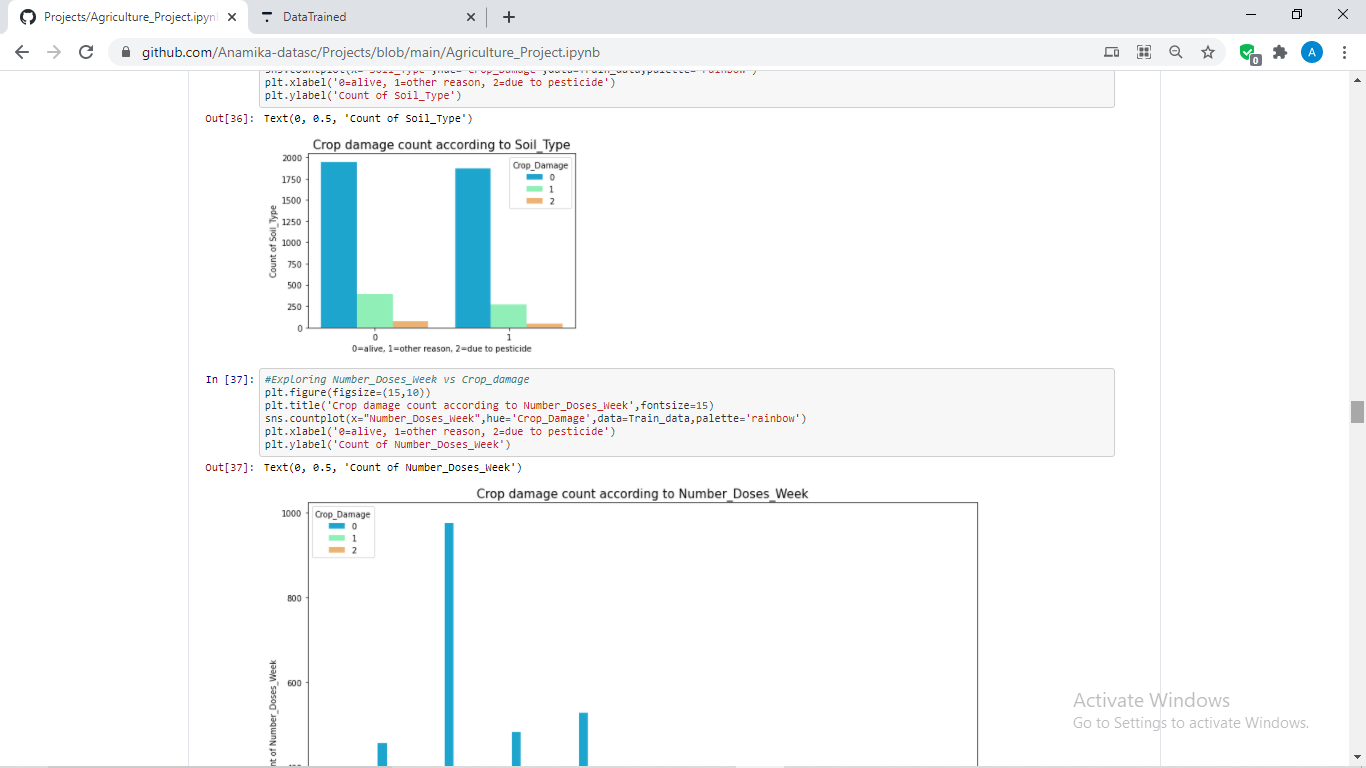
Most of the damage crops are found of crop type 0 and alive count is also highest in the crop type 0 Very less amount of crops got damage of crop type 1.



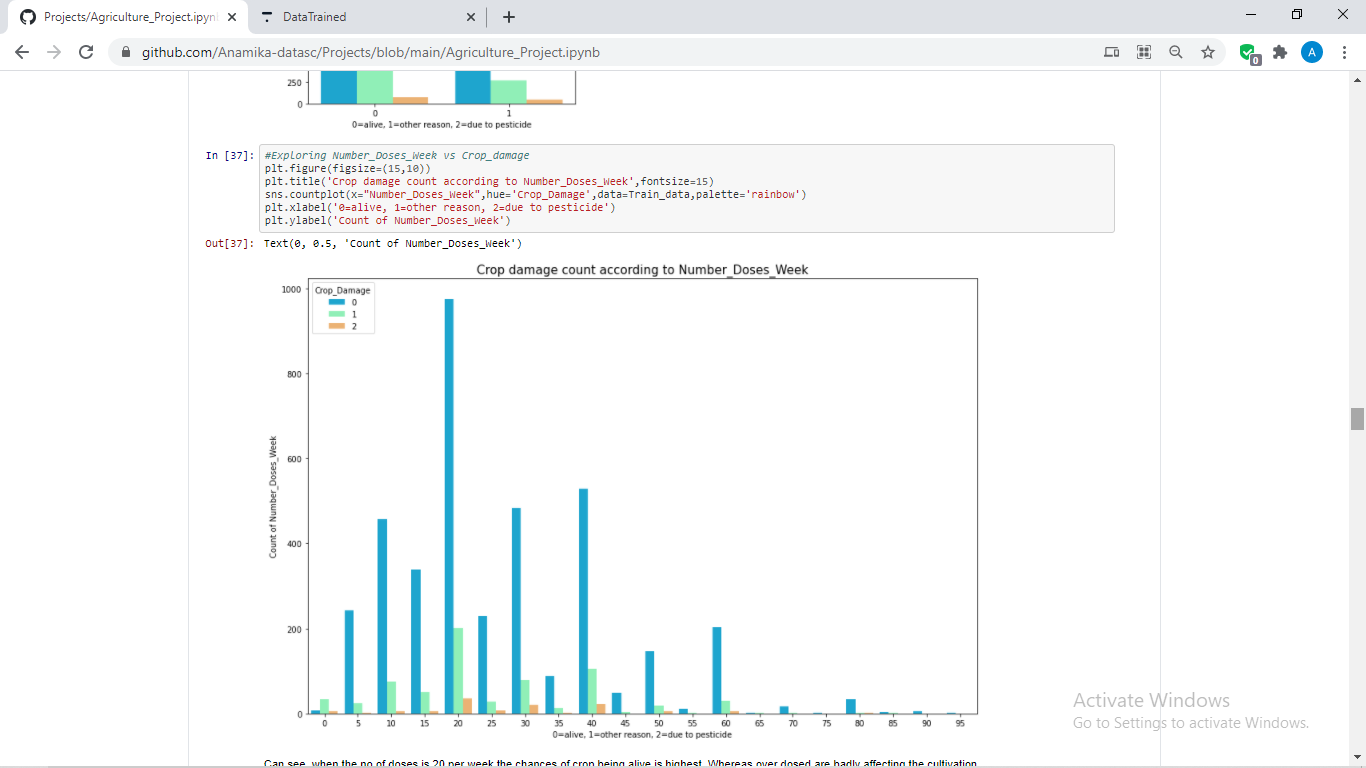
In the second season, chances of crops to live are highest.



Pesticide category 2 has highest alive crop count whereas Pesticide 0 has no impact on the crop cultivation and Pesticide 3 has average impact on crop cultivation.



Soil type does not affect much to crop cultivation, has similar results in both the type.

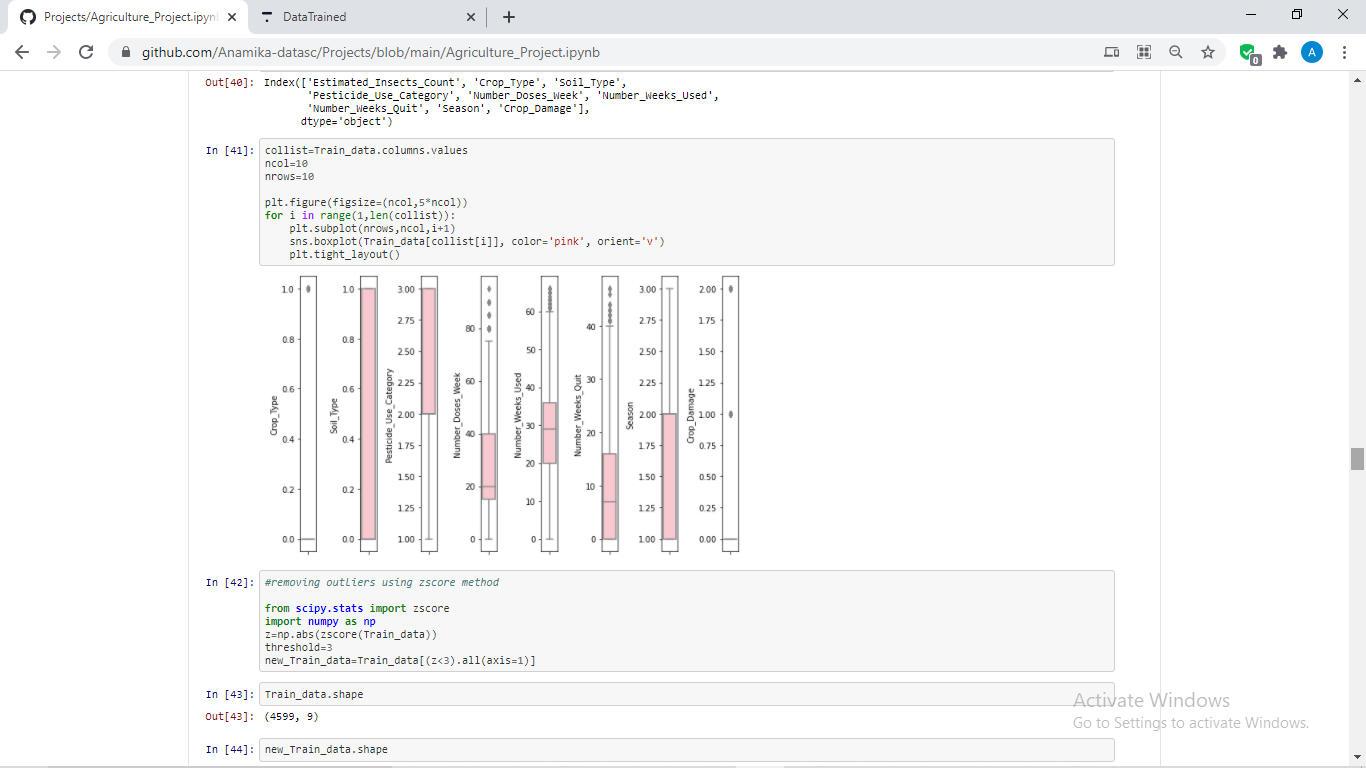


Can see, when the no of doses is 20 per week the chances of crop being alive is highest. Whereas over dosed and under dosed are badly affecting the cultivation.

* Now, will check the outliers using boxplots.

Used for loops to check outliers in all the columns in a one go as we have less columns here. However, all the data may not have less columns like this one. There we wil have to adapt different approach to find out the presence of outliers.

That’s why it is said data science is always **data driven**.



It can be observed that all the columns except Soil\_Type and Season have outliers presence in them.

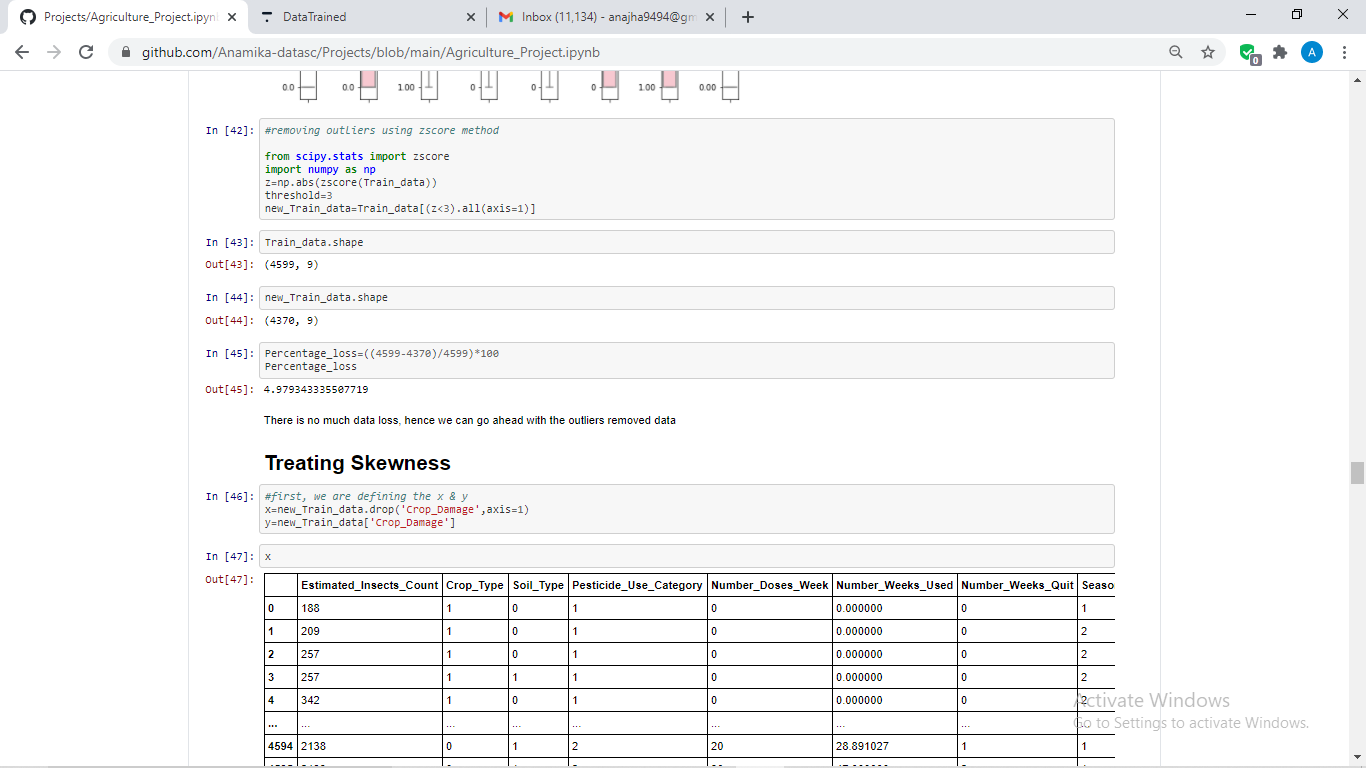
1. **EDA CONCLUDING REMARKS**

From the above EDA and our data analysis done we have extracted the followings insights about the data which will be useful in our data preparation:

1. ID columns has unique data in it and has object data type and hence dropped the same already.
2. ‘Number\_Weeks\_Used’ only had null values in it which has been filled by mean().
3. Data hs some skewness presence which has been computed using **describe() and visualizations** method.
4. There is also a presence of outliers which has been observed using describe() and visualizations methods.
5. Target has highest correlation with **Number\_Weeks\_Used** column followed by **Estimated\_Insects\_Count** and **Pesticide\_Use\_Category**. The correlation is distributed among the columns.
6. Pesticide category 2 has highest alive crop count whereas Pesticide 0 has no impact on the crop cultivation and Pesticide 3 has average impact on crop cultivation.
7. When the no. of doses is 20 per week the chances of crop being alive is highest. Whereas over dosed and under dosed are badly affecting the cultivation.
8. In the second season, chances of crops to live are highest.
9. **PRE-PROCESSING PIPELINE**

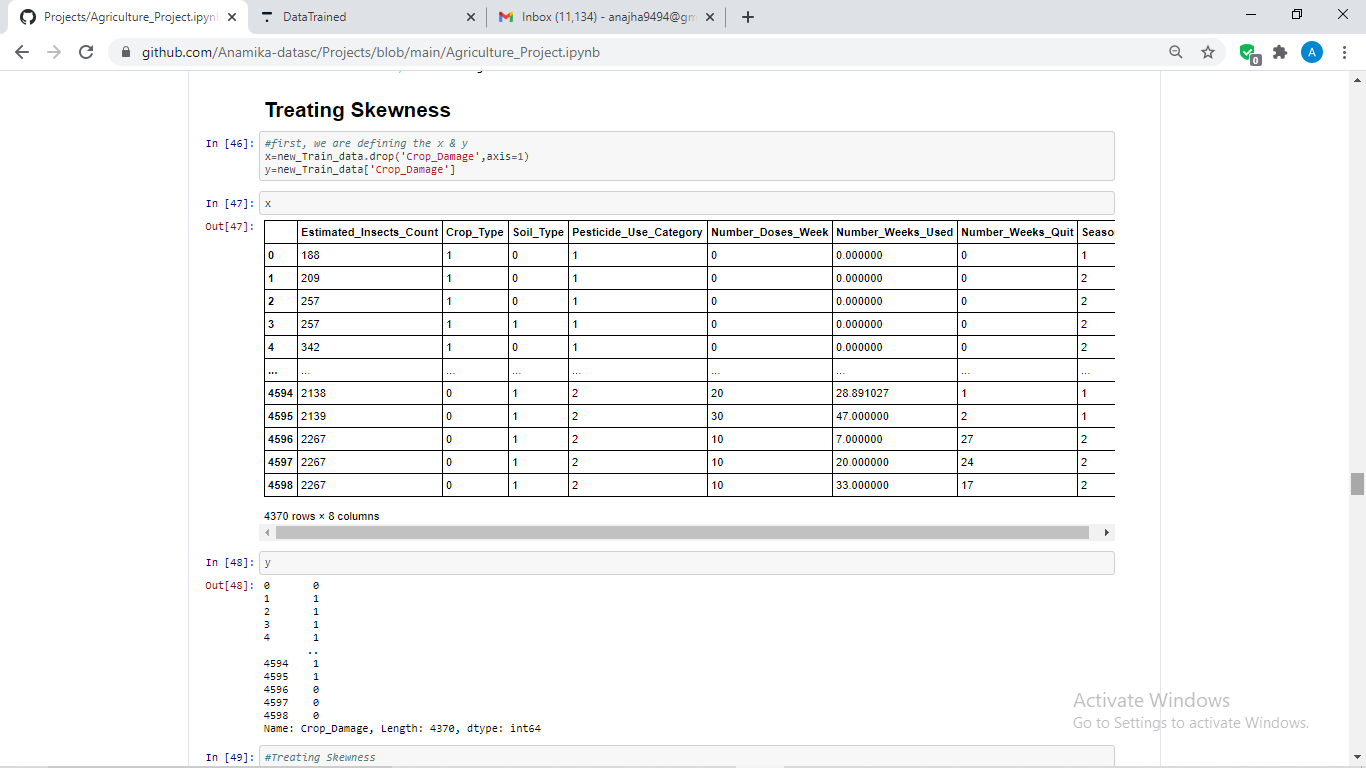
For better prediction and interpretation, we will do some data cleaning which is also called as data pre-processing, pipeline of the same are listed below:

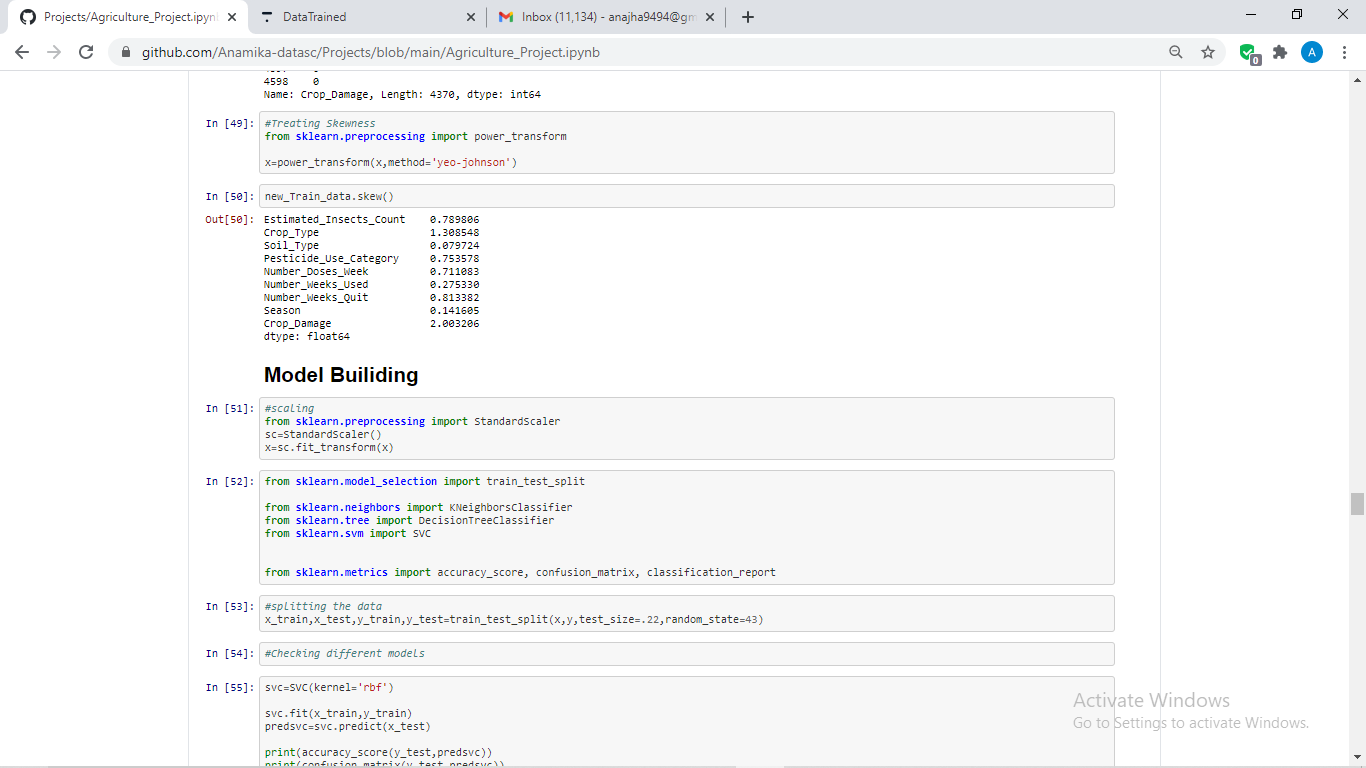
1. Some of the data cleaning has already been performed at the time of data analysis as it was the requirements for further processing like filling NaN values, dropping non-useful and categorical column like “ID”.
2. Since, only ID column had object type data in it we don’t have to do encoding here. All other columns has int/ float data type which is readable by the ML algorithms.
3. From the correlation output, it came to our notice that the correlation is distributed among the columns and since we have less data, we cannot drop weakly correlated columns as well.
4. Some columns have outliers in them and hence will treat the outliers using **z score method** as below:



1. Since the data has skewness, will have to treat the same but we will remove skewness from the x variables only, as removing data from y variable could hamper our predictions.

For the same, will have to define the x and y variable first:





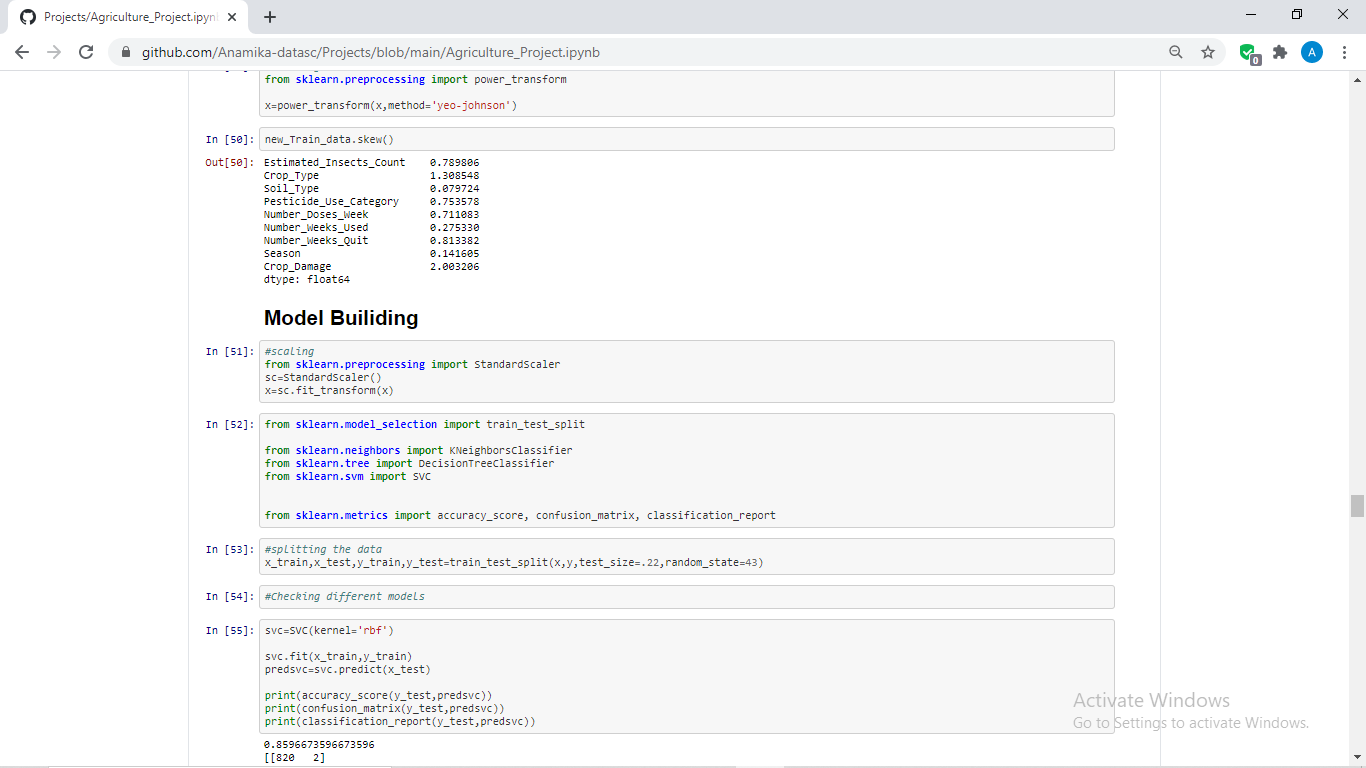
Treated the skewness using **yeo-johnson** method as this is the most affective for skewness removal.

1. **BUILDING MACHINE LEARNING MODELS**

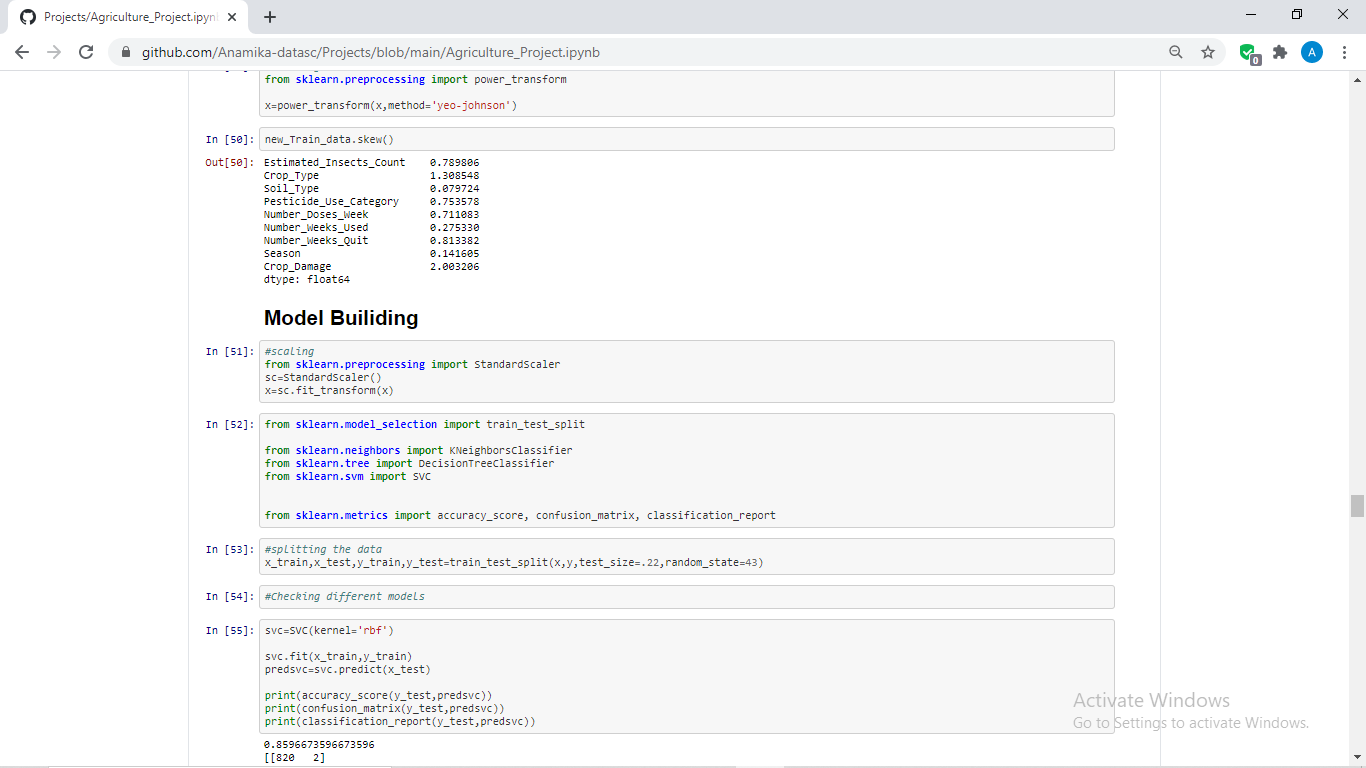
Building ML model involves scaling the data, finding best parameters for algortithms, training the model and finally testing the model.

Since, data science is always data driven in this data we will opt for following steps to find the best algorithms having maximum accuracy by checking various algorithms of classification models.

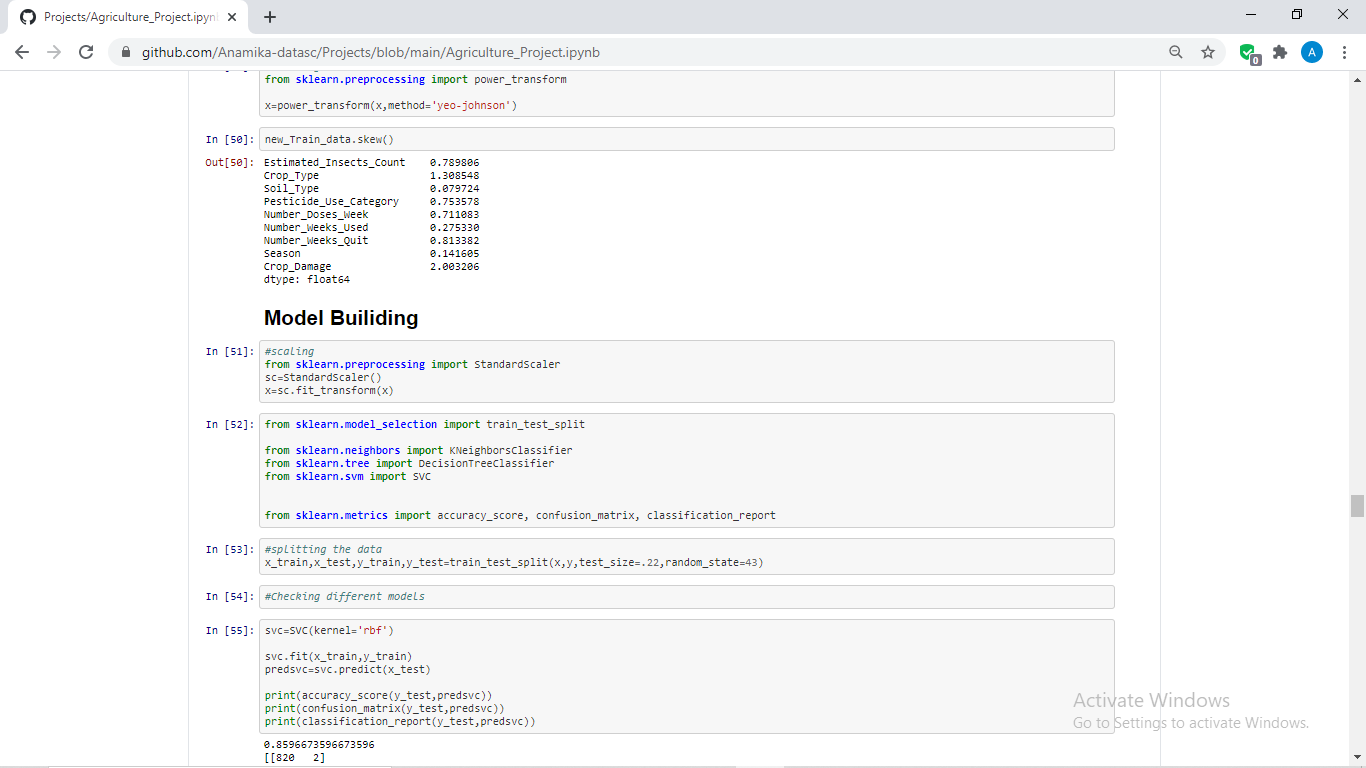
* First, we will scale the data to sum up the data on the same page with same scale of values. Scaling is done as below:



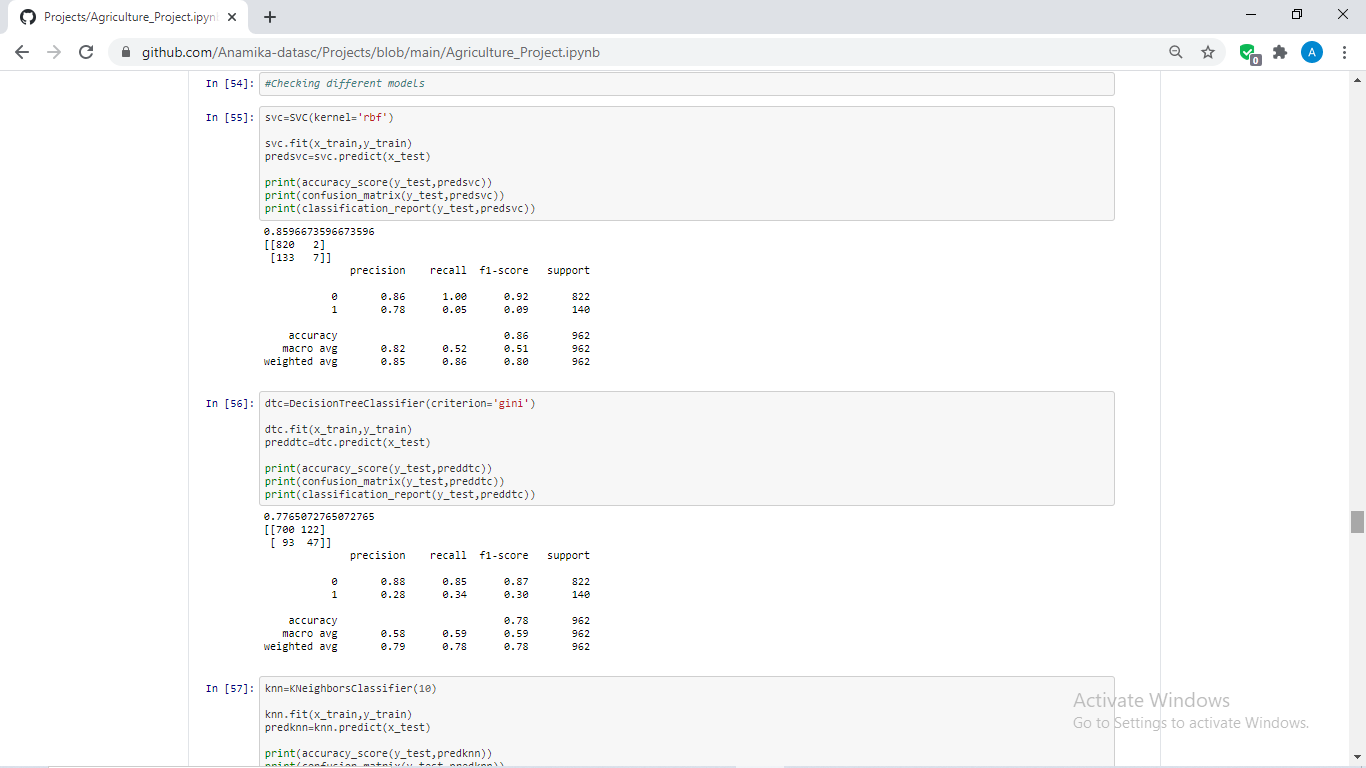
* Next, we will import some necessary libraries for model building as below:

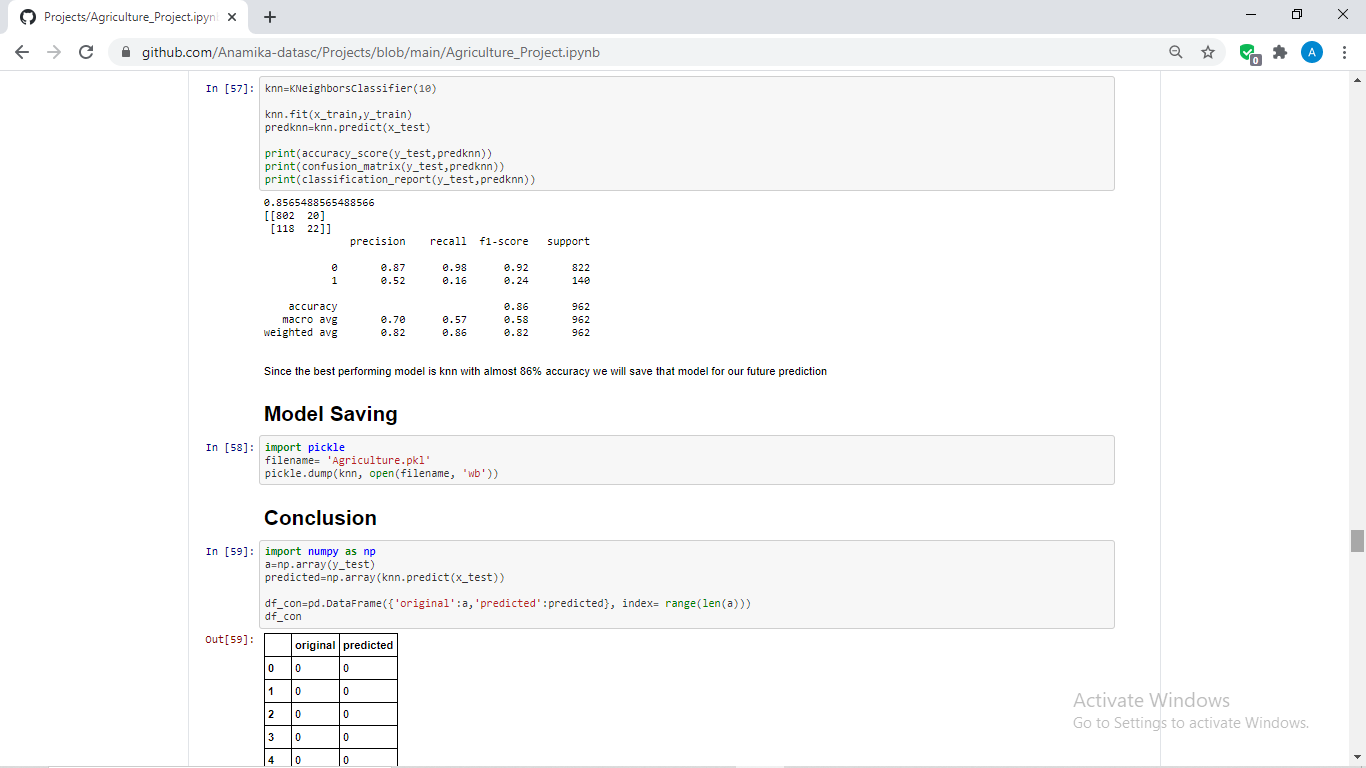


* Next, we are splitting the train and the test data for our model training and testing:



* Now, we are passing our train and test data to different models for the training and the testing and we will check the scores of them with the help of matrices like accuracy\_score, confusion\_matrix and classification\_report. You can check the codes and the model performances as below:

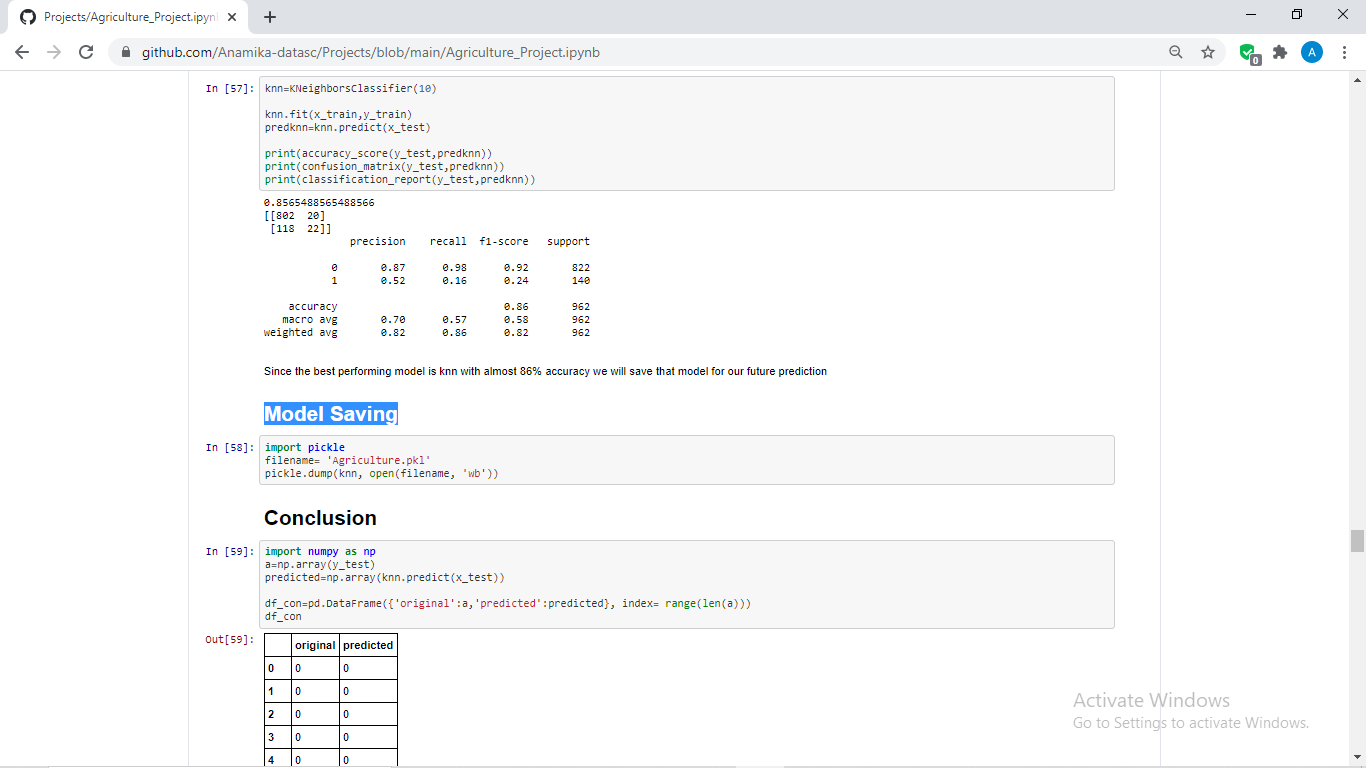




As we can clearly observe, **KNN** is the best performing model with the accuracy of 85%, we will go ahead with saving this model for our future predictions.

* Model Saving

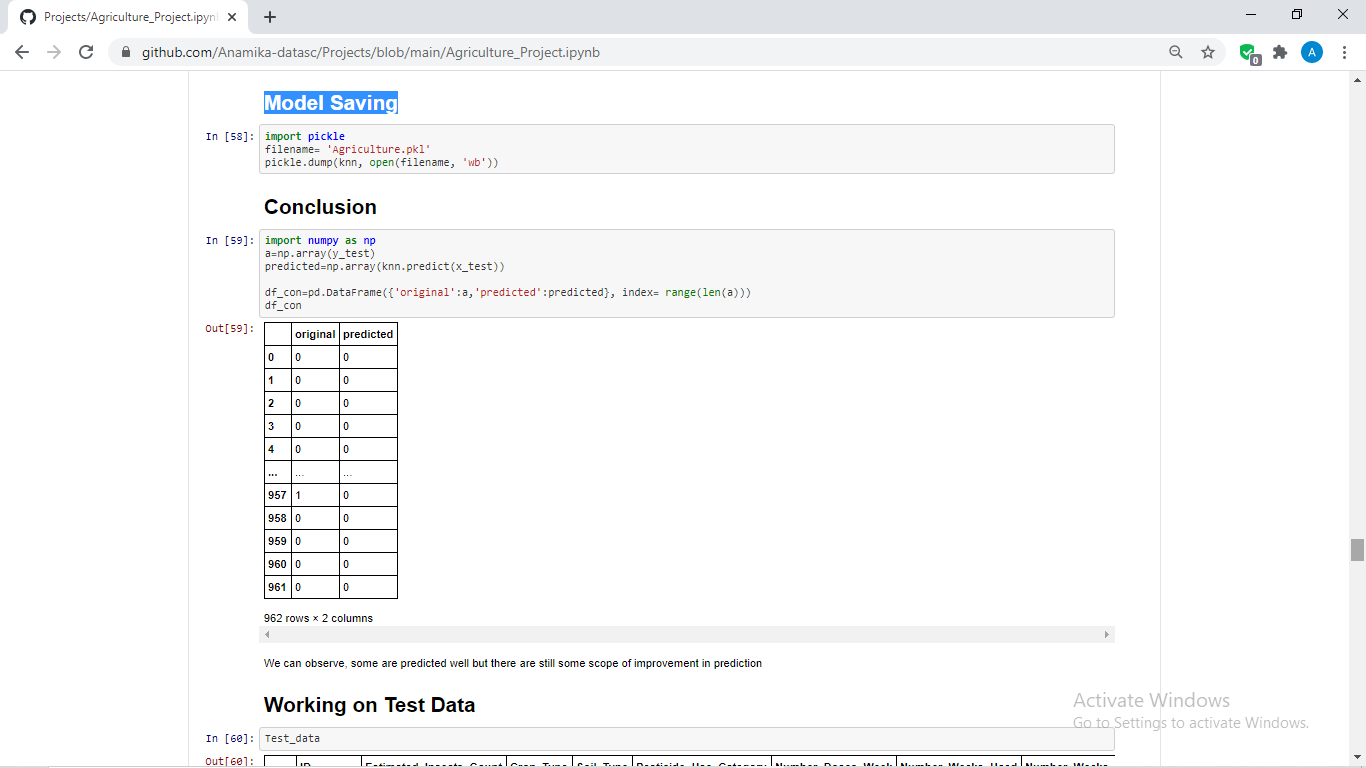
Now, we are saving the model in pickle as per below codes:



Here, we have saved our model with specifying file name in the write binary mode.

1. **CONCLUDING REMARKS**

After model saving will check the accuracy of the model by comparing original which is y\_test and the predicted data, which is done as below:



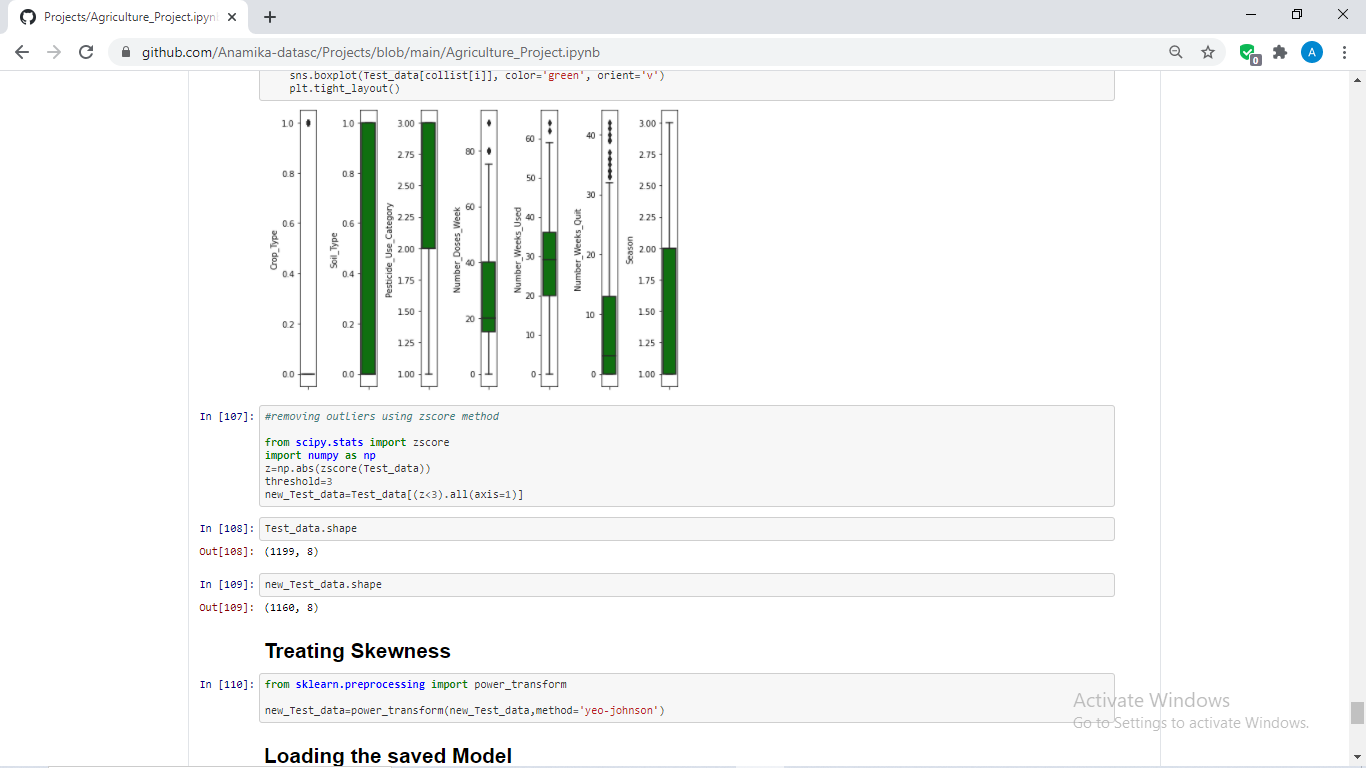
We can observe, the model is performing quite well with almost 86% accuracy but there are still some scope of improvement in prediction which can be achieved by applying certain hypertuning methods, checking the best random state, with some boosting and ensemble methods and testing more models.

1. **WORKING ON TEST DATA AND TESTING THE SAVED MODEL ON OUR TEST DATA**

Here, there is 1 more step to be done which is to work on our Test data which we have highlighted while loading the datasets that this dataset also have Test data in it which we can use for testing our saved model.

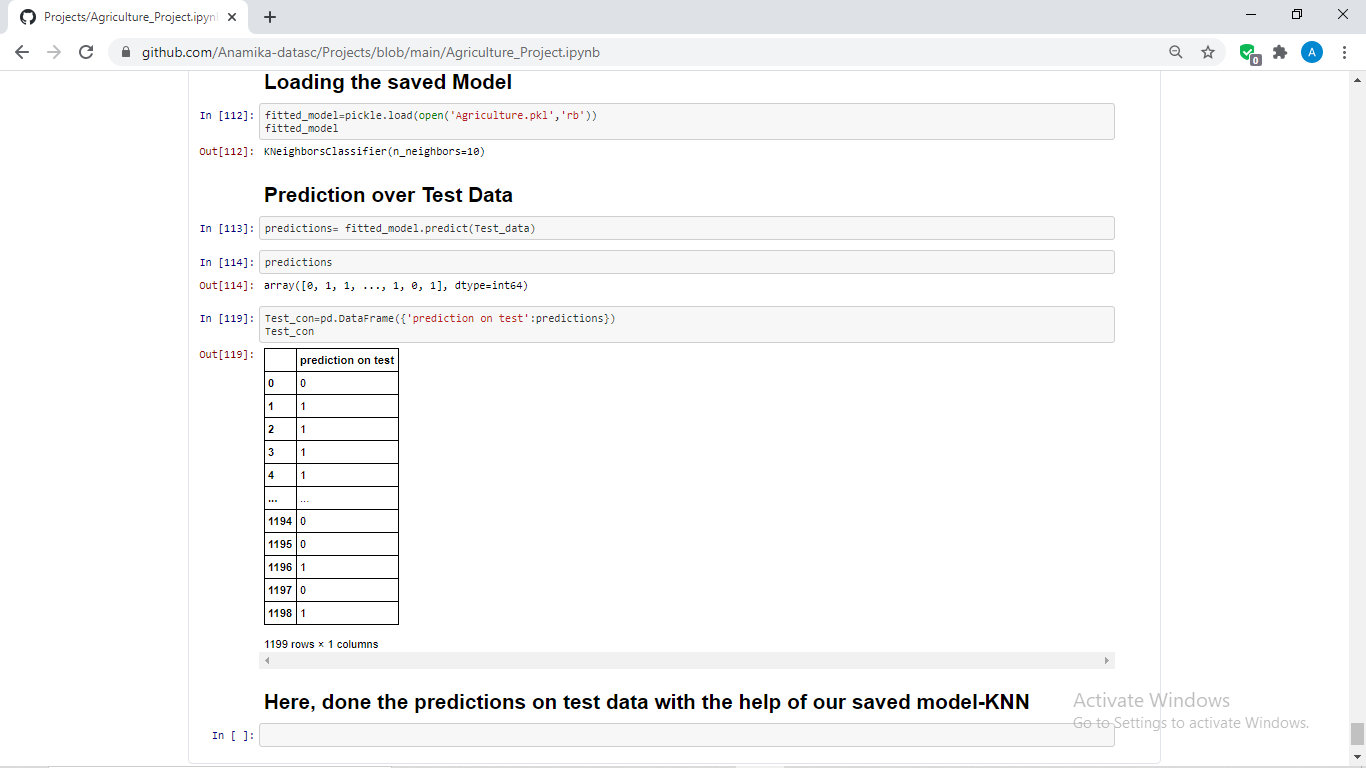
Since, all the EDA, Data Analysis and data pre-processing are almost same as we have done on our Train data, showing that again makes no sense.

There is only 1 difference which is while removing skewness. Since there is no y variable in this dataset we won’t be splitting the data before skewness removal. For better understanding, please refer below codes.



Shown the steps just before skewness removal for better understanding.

Now, we will use our saved model for predicting the outcome on test data where the Crop\_damage i.e. our target variable is not available. Please refer the below codes for the same:



Above, we have loaded the saved model with the **pickle.load** function. Further done the predictions on Test data with our saved model KNN which has shown with the variable **Test\_con**.

**Summary**: From our above ML project we have seen how we can predict whether the crop will survive or not on the basis of various factors like usage of pesticide, soil type, season etc. Similarly, we can apply ML on any kind of data set and in any field with the intention to make better use of the data available everywhere.

Open for the suggestions and more insights on this ML project!!

Thank you for reading my article.

Author

(Anamika. Jha)