Machine Learning in Agriculture



Use of Pesticides in Agriculture

Pesticides are substances or mixtures of substances that are mainly used in agriculture to manage numerous pests and sickness carriers, like mosquitoes, ticks, rats and mice. Pesticides are a chemical compound utilized in agriculture to regulate weeds, insect infestation and diseases, With the introduction of pesticides, farmers are ready to manufacture larger crops on less land, Increasing crop productivity by between 20 to 50%. Additionally, pesticides permit farmers to maximise the advantages of different valuable agricultural tools, like prime quality seeds, fertilizers and water resources. Pesticides area unit so an essential tool for the property production of prime quality food and fibres.

Benefits of Pesticides

The primary advantages are the implications of the pesticides' effects – the direct gains expected from their use. As an example the impact of killing caterpillars feeding on the crop brings the first advantage of higher yields and higher quality of cabbage. The secondary advantages are the less immediate or less obvious advantages that result from the first advantages. they'll be refined, less intuitively obvious, or of long term.

• Pesticides help keep food affordable.

• Quality of food

• Protection of crop losses/yield reduction disease control

• Pesticides help keep food affordable.

• Quality of food

• Protection of crop losses/yield reduction

• Disease control

Harmful Effect of Pesticides

A wide variety of chemicals are utilized in international agriculture below the perception that they're basic to achieving most crop yields. There is, however, a growing mental object developing that underpins the distribution and impact of chemicals within the setting and the way they have an effect on the body. Moreover, there's growing demand for restrictive management and management procedures because of the particular challenges that are being recognised.

How to Use Pesticides?

Pesticides are essential to our food supply, to water purification, and for sanitizing surfaces in hospitals, homes and producing. They defend us from sickness carrying insects and plenty of alternative pests. If used consistent with the label, pesticides are safe for human health and therefore the atmosphere. employees in several settings may be exposed to pesticides. varied employee safety programs guarantee their safety. it's most likely pesticide use within the food provider that's most controversial. Safety standards have come back a long way towards reducing chemical residues in foods and prohibiting pesticides that the chance of damage is just too nice. Integrated pest Management (IPM) may be a program of dominant pests that depends on physical, cultural, biological, and chemical ways, rather than simply looking forward to artificial chemicals.

Use of machine learning in Agriculture



Machine learning has emerged with big data technologies and high-performance computing

to create new opportunities for information intensive science within the multi-disciplinary agri-technologies domain. Here we will do analysis dedicated to applications of machine learning in agricultural production systems.

Problem Statement

Though, many of us don't appreciate much, a farmer's job is a 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,

You 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.

DATA ANALYSIS

Data Description

Here we have given a different feature of crops and we have to find out the crop damage level.

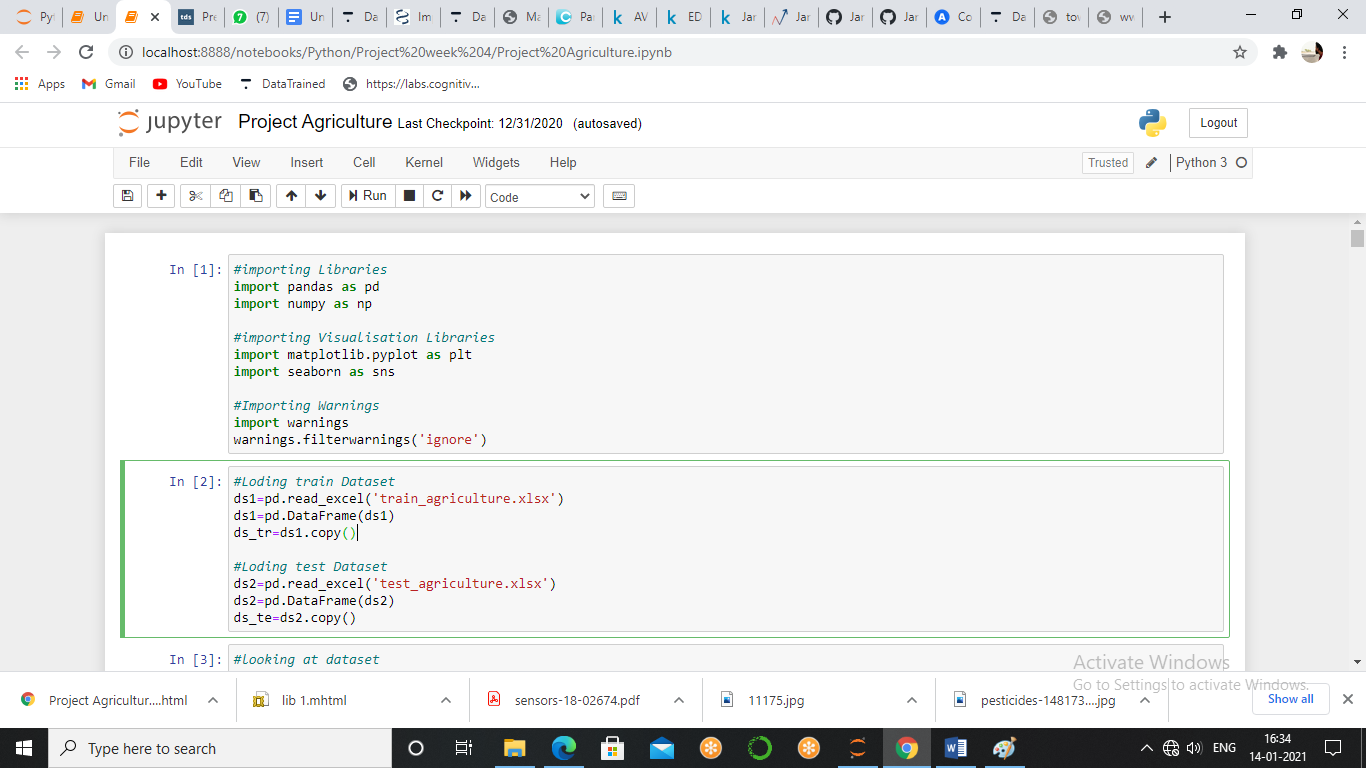
|  |  |
| --- | --- |
| Features | Description |
| Id | Unique ID |
| Estimated\_Insects\_Count | Estimated insects count per square meter |
| Crop\_Type | Category of Crop(0,1) |
| Soil\_Type | Category of Soil (0,1) |
| Pesticide\_Use\_Category | Type of pesticides uses (1- Never, 2-Previously Used, 3-Currently Using) |
| Number\_Doses\_Week | Number of doses per week |
| Number\_Weeks\_Used | Number of weeks used |
| Number\_Weeks\_Quit | Number 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 Pesticides) |

Importing Libraries

Libraries are sets of routines and functions that are written in a given language. A robust set of libraries can make it easier for developers to perform complex tasks without rewriting many lines of code.

Here we are importing 3 types of basic libraries, mathematical i.e Numpy this will help us to perform mathematical operation and second is pandas, to put data into DataFrame which will help us in pre-processing and last one is visualization libraries.

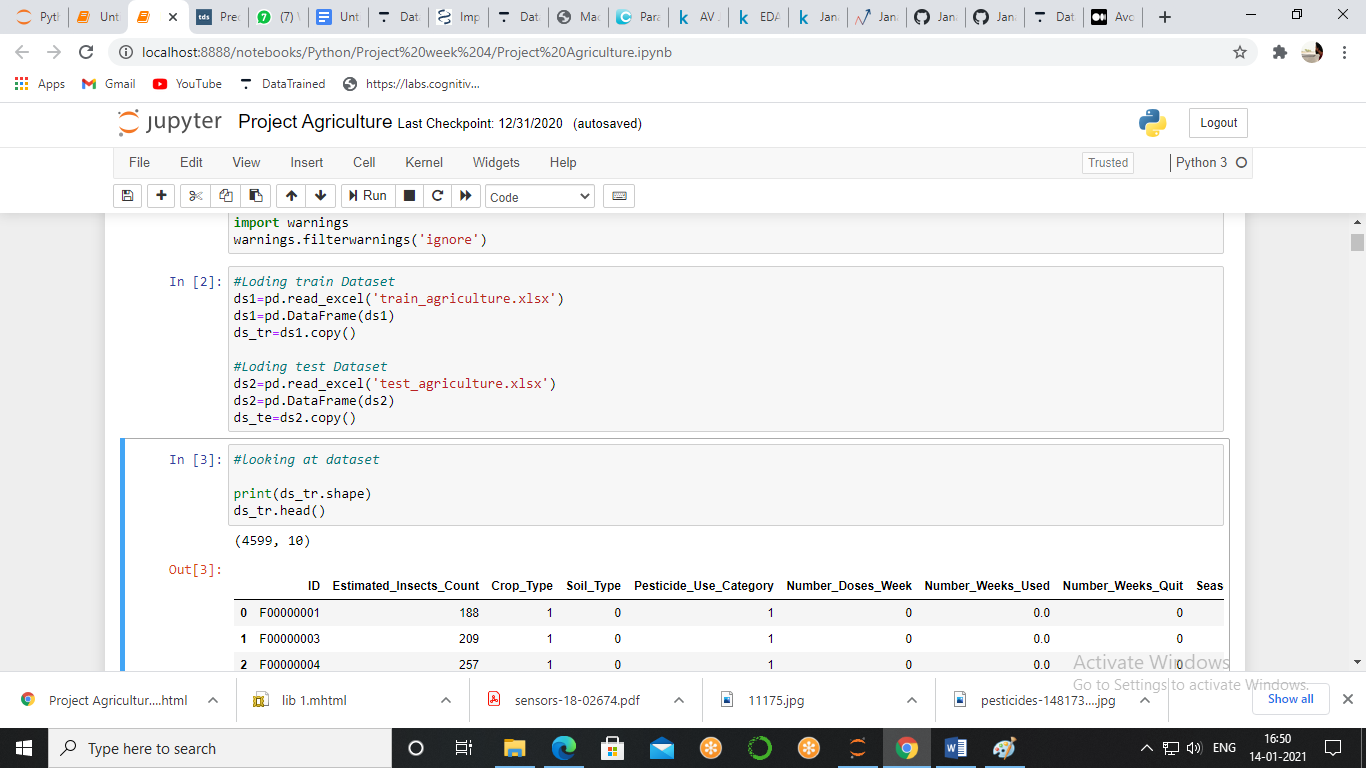
We are also importing warning libraries to ignore unnecessary warnings which will hinder with output of code.



Loading Dataset

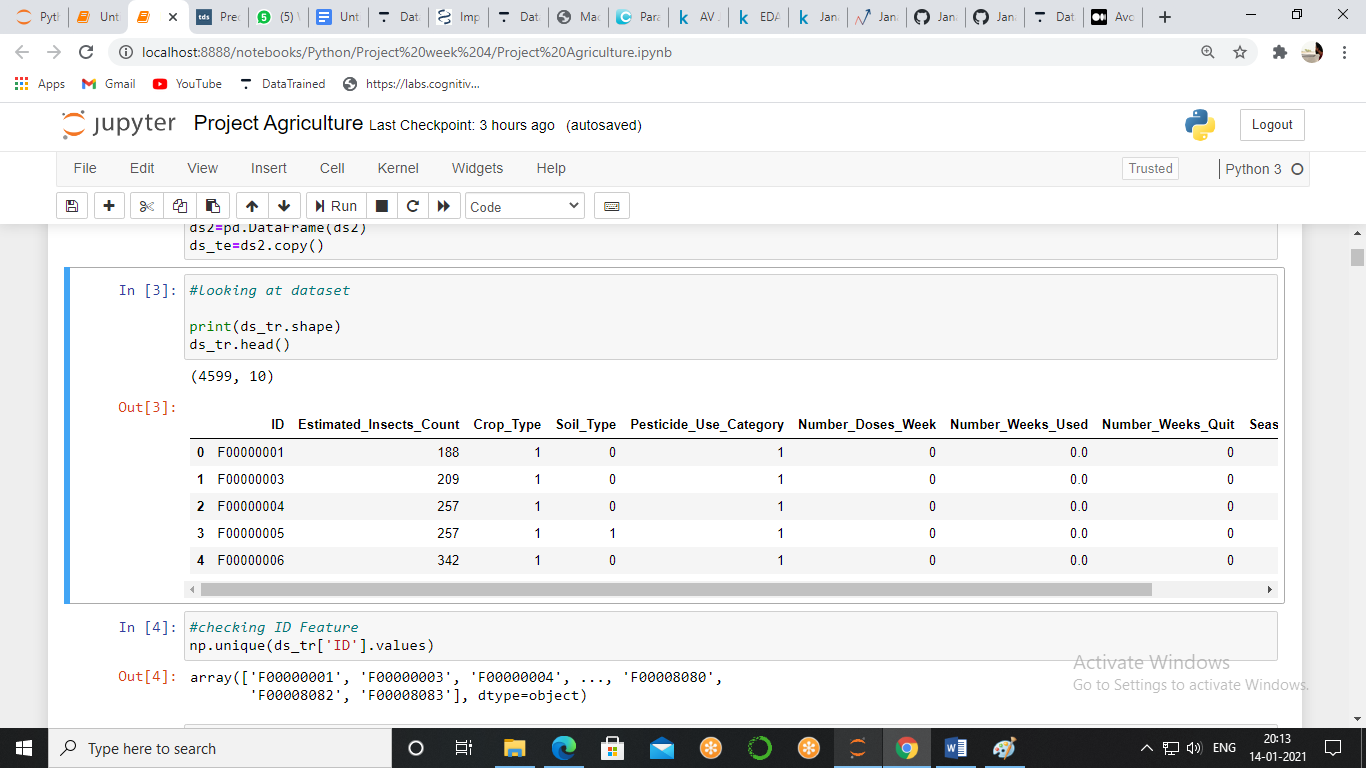
We have two datasets Train and Test, so we need to do all the pre-processing on both the datasets. The difference between both the datasets is that the train dataset has a dependent variable(y) and the test dataset does not have (y).

The raw data is available in excel format so we are importing both the datasets into DataFrame and also making a copy of it at initial stage to be future proof. So that in case if we require the original form of data then there will be no need to import the data again. I put train data in variable ds\_tr and Test data set in variable ds\_te



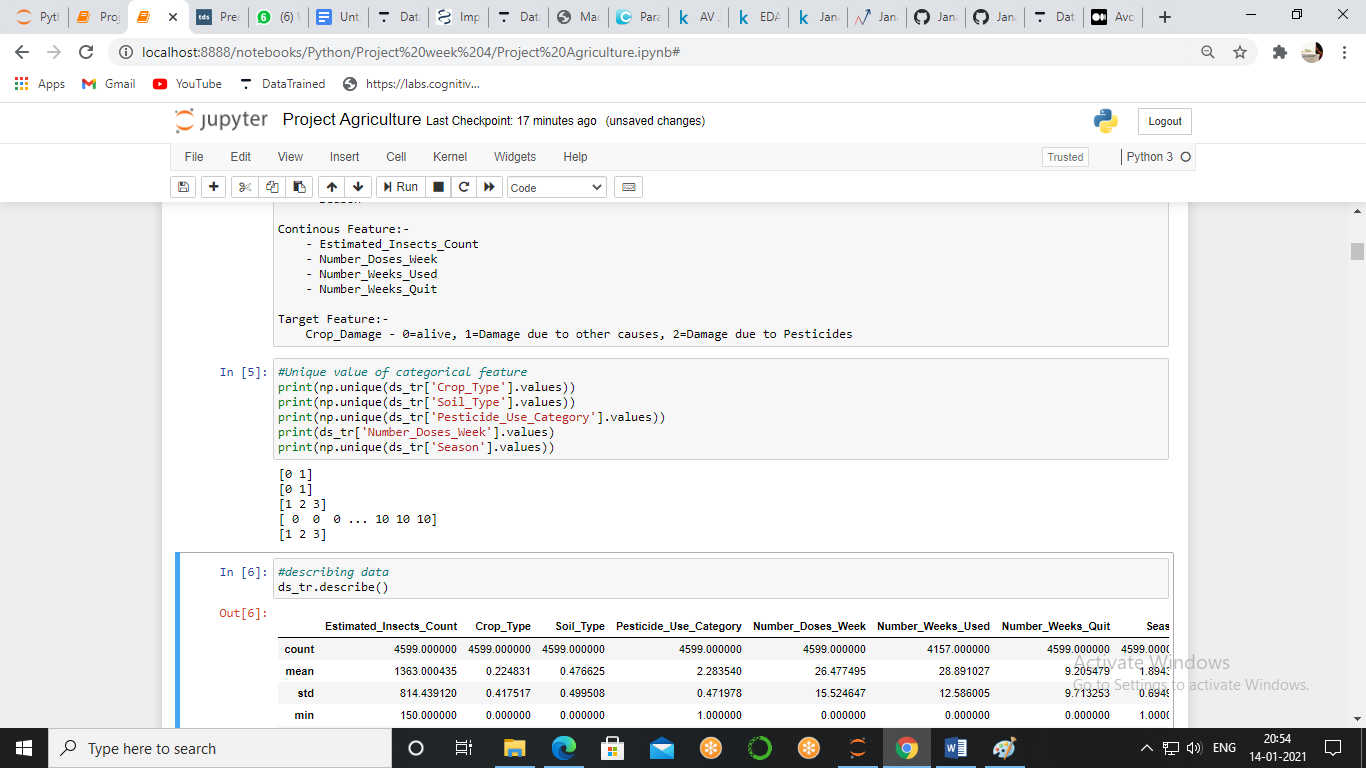
Looking at dataset

Now we will check the size of train data as well as take a peek at the data. We used the head method to load the first five rows because we don’t the data and info we will be needing will be visible through the first five rows only.

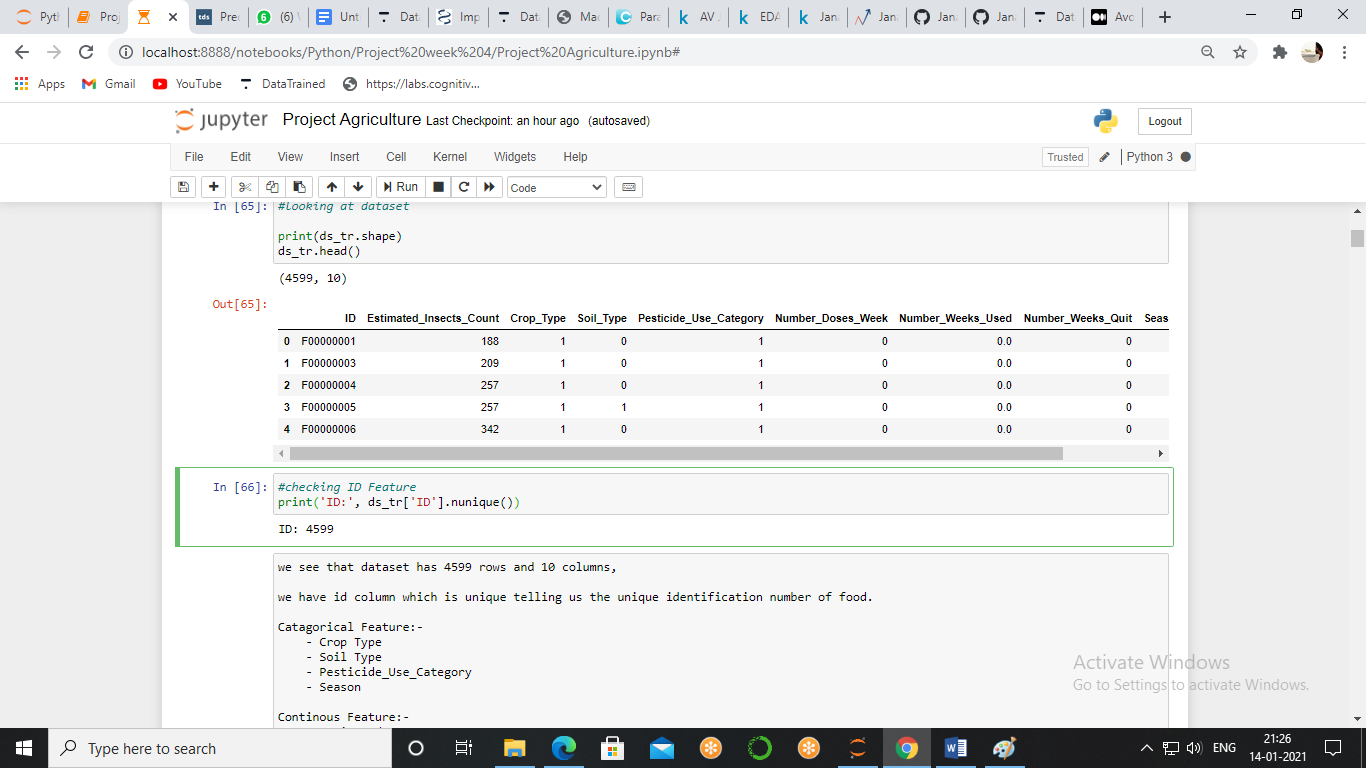


We see that our dataset has 4599 rows and 10 columns plus we have id column which is unique telling us the unique identification number of food and has categorical feature *Crop Type, Soil Type, Pesticide Use Category, Number Doses Week, Season* and numerical feature *Estimated Insects Count, Number Doses Week, Number Weeks Used, Number Weeks\_Quit* while Crop Damage as our Target Variable.

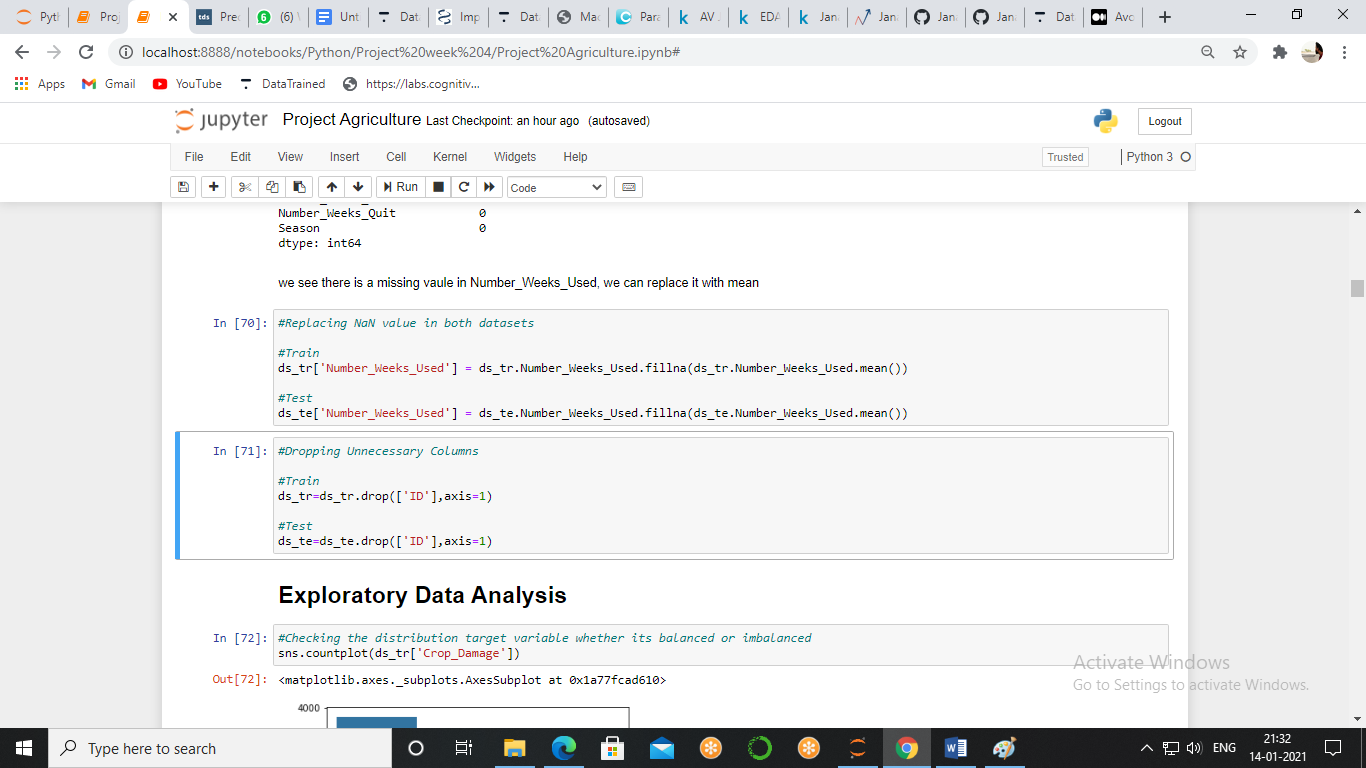
Now, we will look at the unique value of categorical features because in numerical features all the values will be different as they are continuous, so, will look at the unique features.



We have 2 types of crop, 2 types of soil, 3 type of pesticides use category and 3 seasons but now we have to look at the ID feature to find its uniqueness, as we assume that it has unique feature for every row but still we want to be confirmed about that because one line of assurance hurts no one.

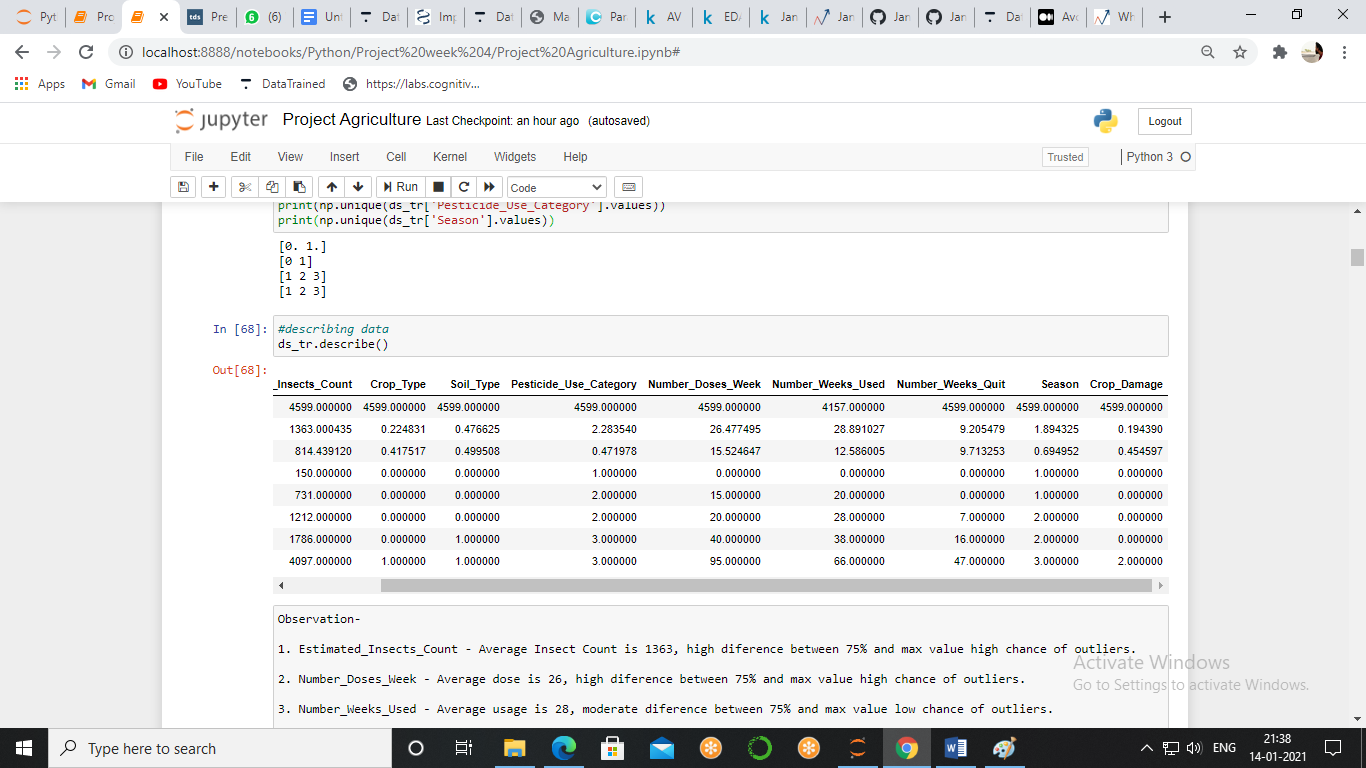
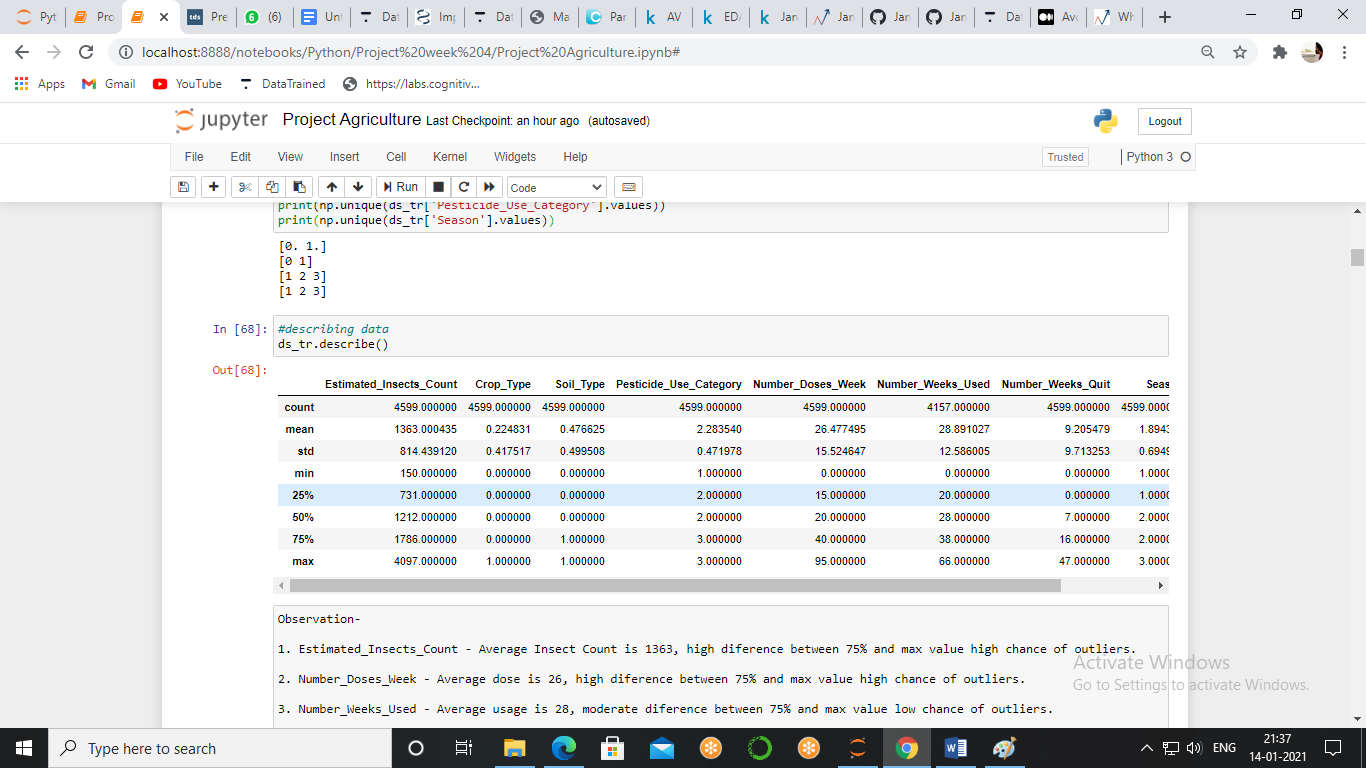


ID has 4599 unique elements and our data has 4599 rows which establish the fact that ID has all unique elements plus it is not going to add any value in the analysis part, so we will not proceed with it.



Describing Data

Statically data shows us so much information about the dataset, like its skewness, potential outliers, min, max, standard deviation, quartiles.

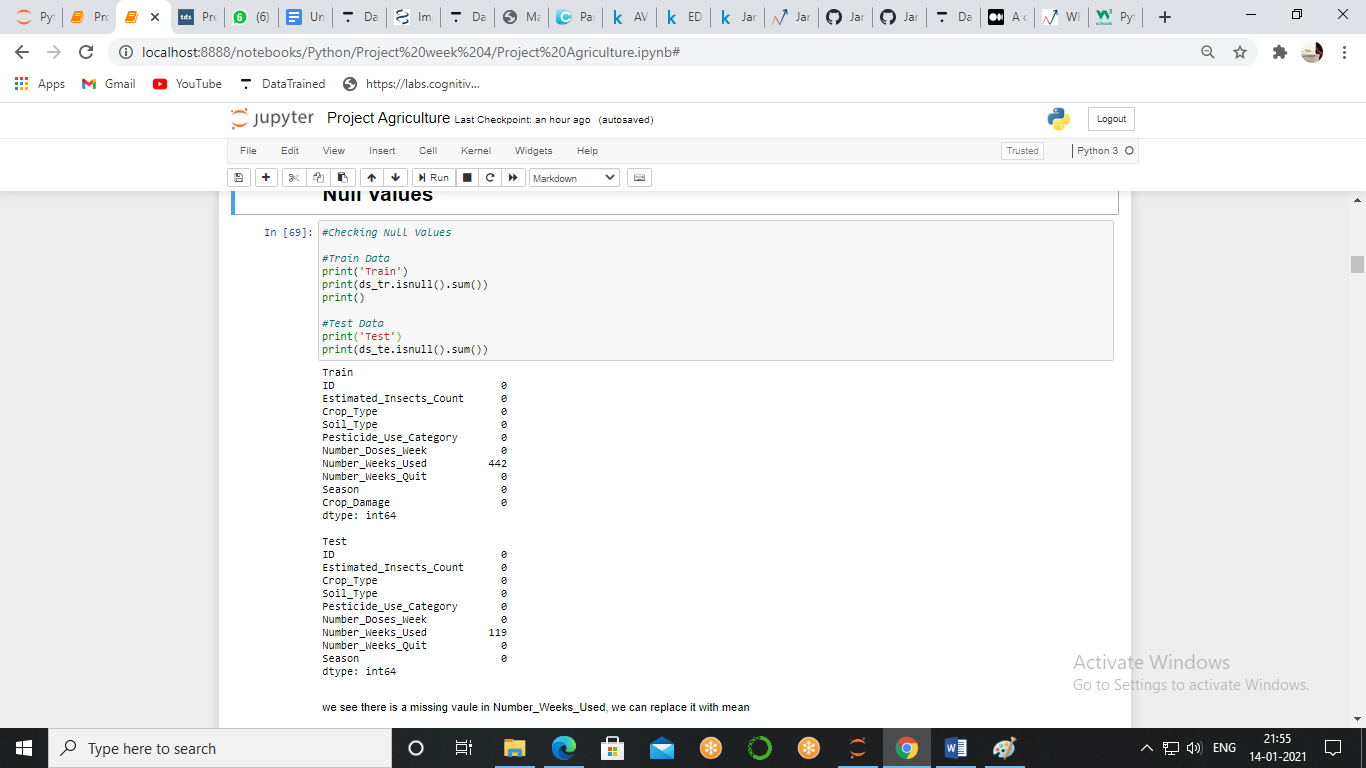


The high difference between 75% and max shows us outliers and when we see Estimated\_Insects\_Count, Number\_Weeks\_Used , Number\_Weeks\_Quit we see high outliers which we will remove later. Now, we are not going to analyse skewness here because we can do it with one line of code and that thing will be much more time efficient. So, we will see average value here

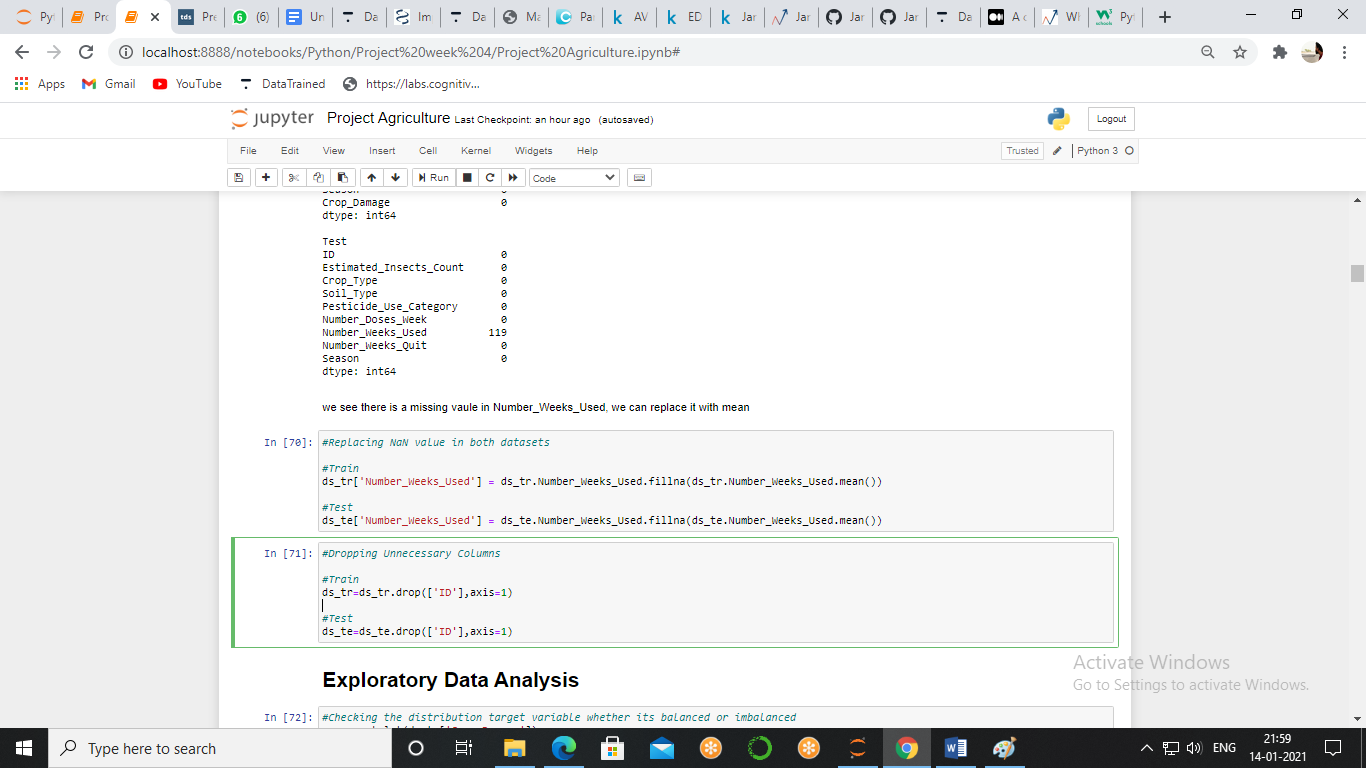
* Estimated\_Insects\_Count- Average Insect Count is 1363
* Number\_Doses\_Week - Average dose is 26
* Number\_Weeks\_Used - Average usage is 28
* Number\_Weeks\_Quit - Average number is 9

Null Values

The None keyword is used to define a null value, or no value at all, None is not the same as 0, False, or an empty string. None is a datatype of its own (None Type) and only None can be None.



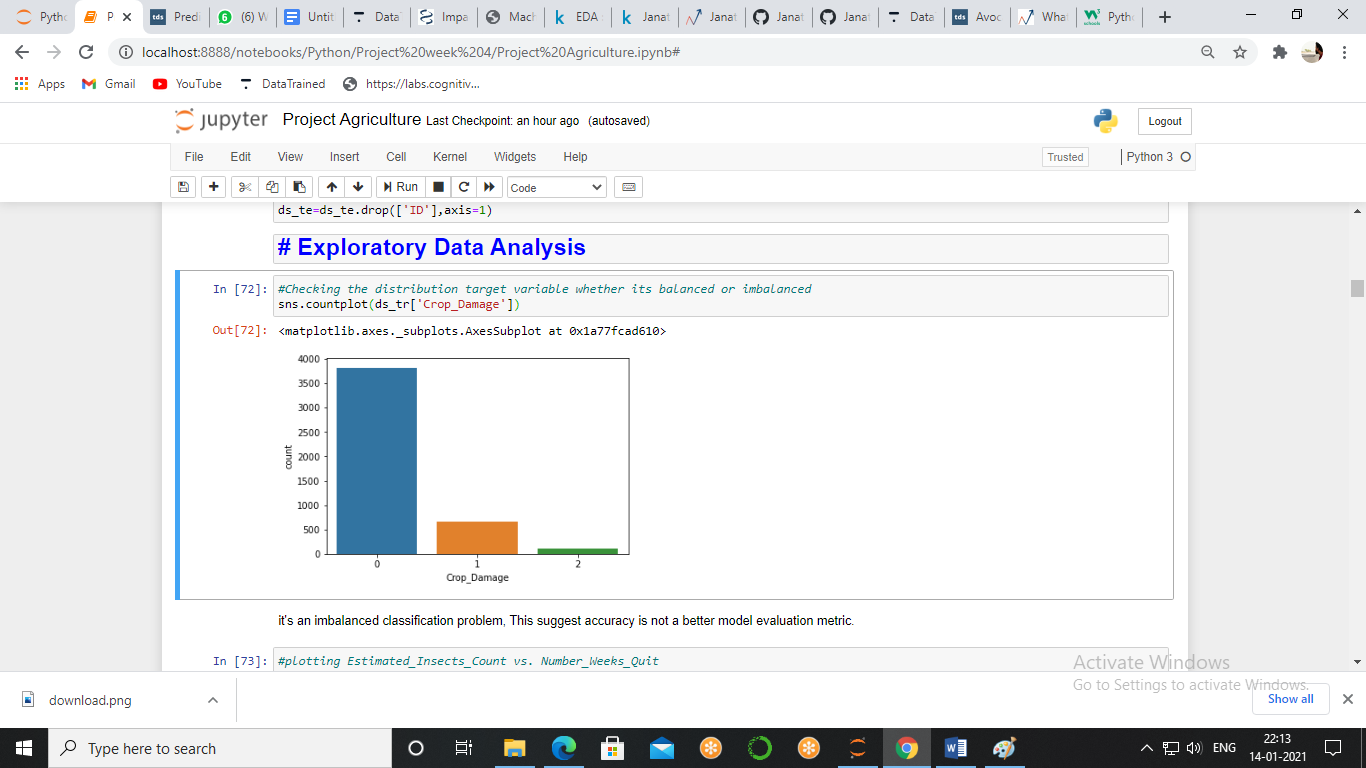
There is only one column which has approx. 1% of null values, now we see that the Number Week Used feature is a numerical feature which we can replace with the mean of the entire column. We will be using the fillna method.



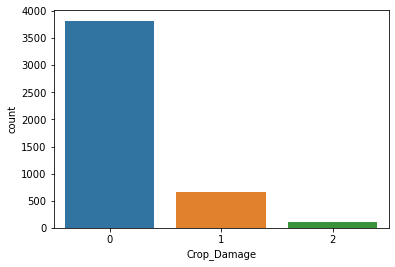
Exploratory Data Analysis

Exploratory data analysis is an approach to analyse the data. It's where a data enthusiast would be able to get an idea of the overall structure of a dataset by bird's eye view. Data science often consist of advanced statistical and machine learning techniques

First we will look at the Target Variable i.e. Crop Damage



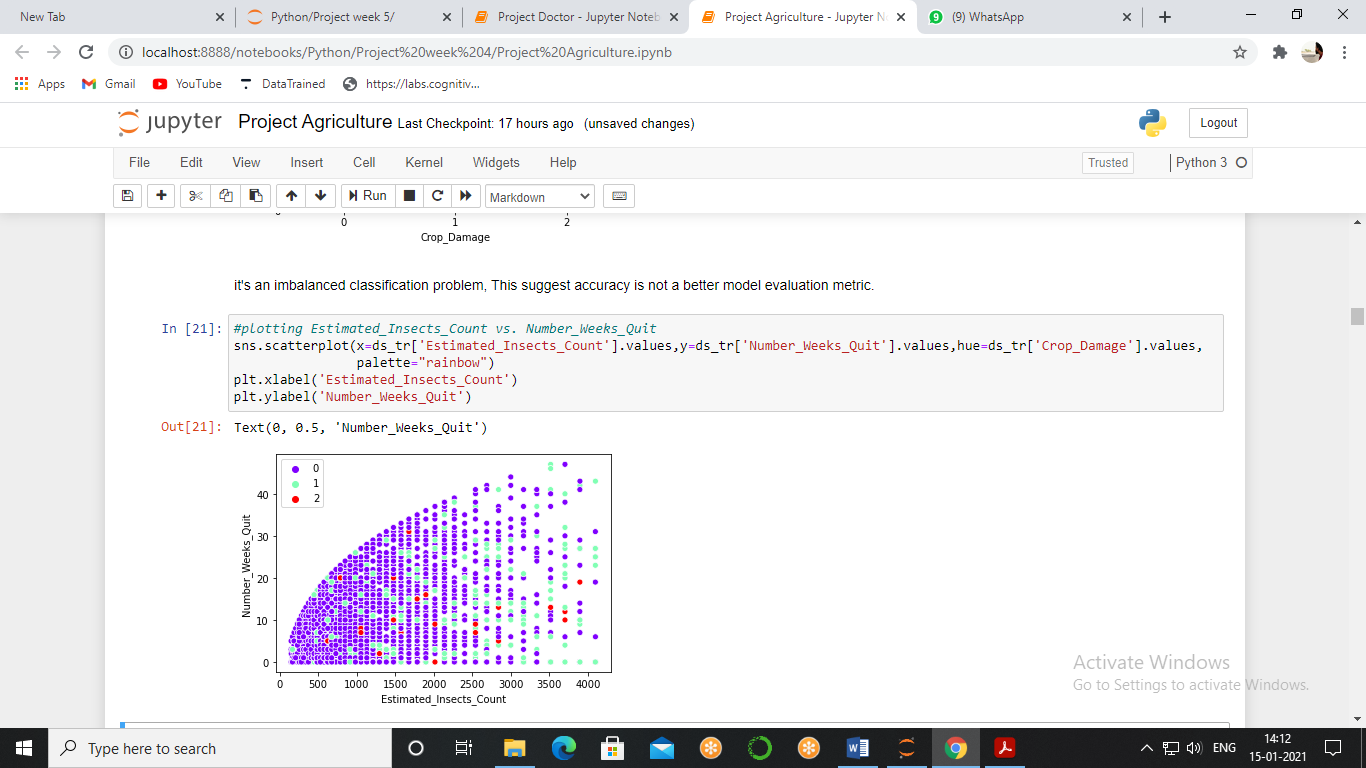
We are using count plot function to check the number of values of Crop Damage

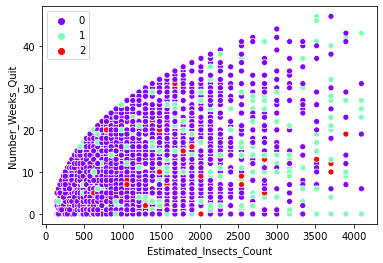


It's a highly imbalanced classification problem, this suggests accuracy score is not a better model evaluation metric we have to go for auc-roc and sampling a data as well.

Pesticides, Insects and Damage

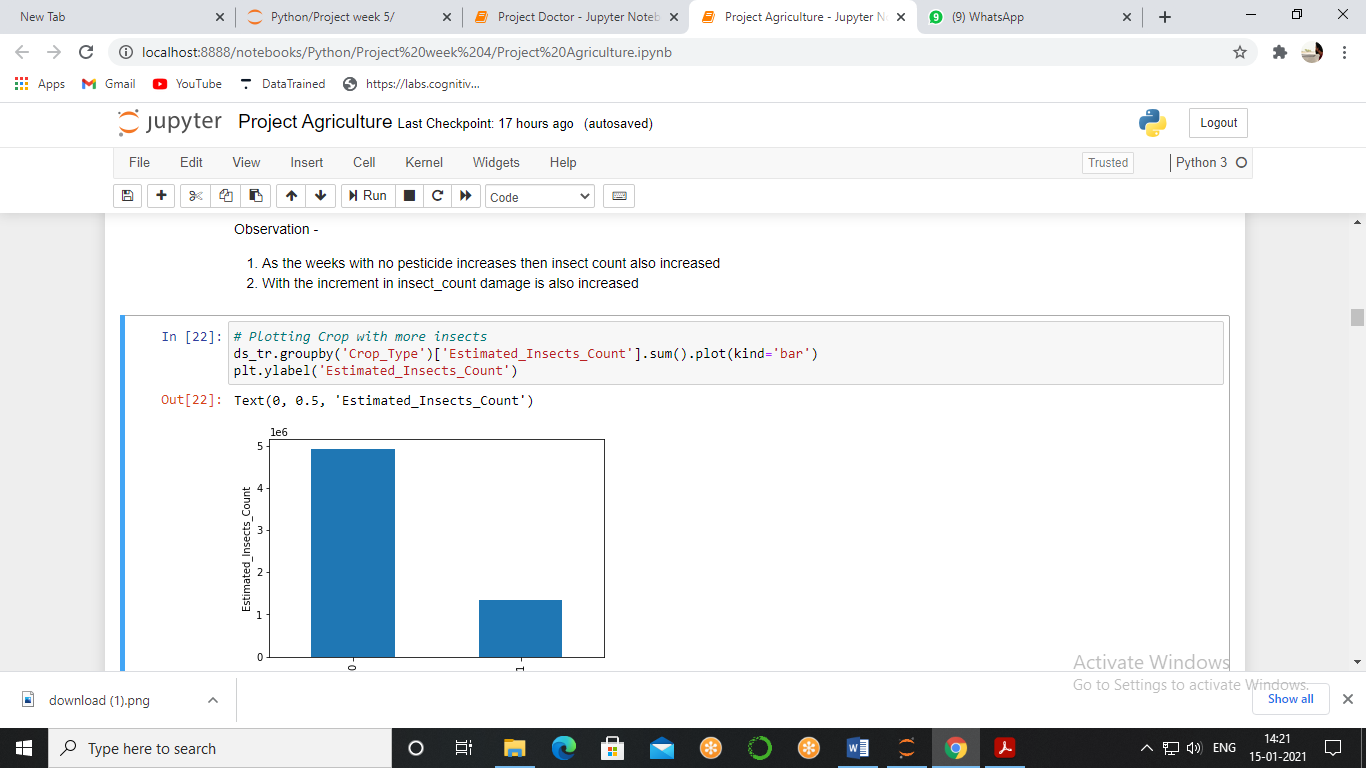
We see above that pesticides are used to repel insect so we will see scatter plot to see the relationship between use of pesticides and instant number of insect we will also see the crop damage done by various factor in this analysis

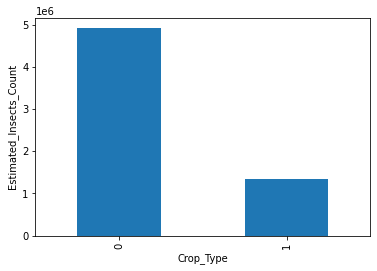




As the use of pesticides stops the number of insect increased and we can clearly see that above but there is also interesting fact that with the increment in insect the crop damage Type 1 is also increased which is the damage done by other factor and here we can presume that other factor is, increment in number of insect. Here our final take is if we completely stop using the pesticides then the insect will damage the crop.

We saw above that the insect plays a vital role in damaging a crop but there we have to type of crops and before the want to analyse which type of crop is most affected by the insect so we will plot a bar graph by making a group of Crop type and estimated insect count and adding insect count On particular crop

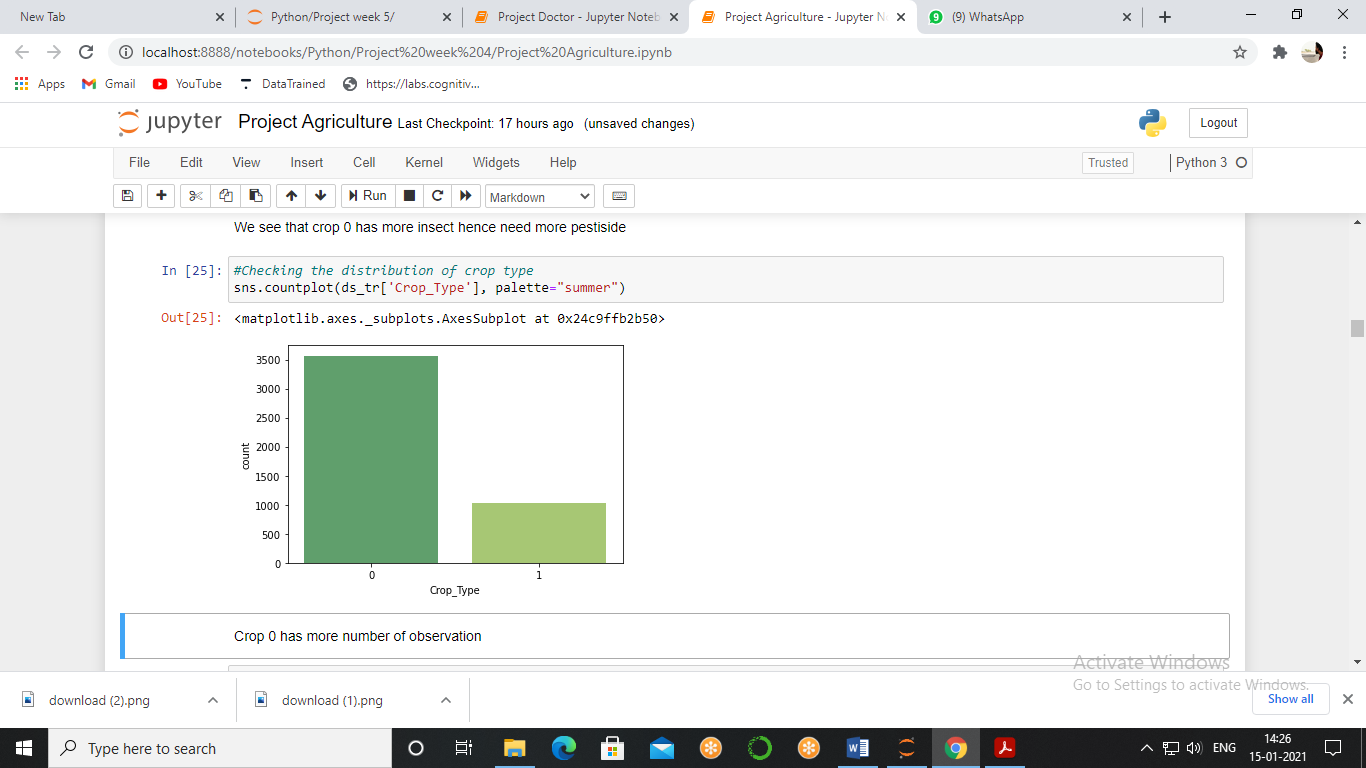


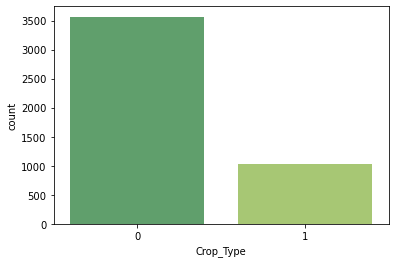


The crops zero has more insects so it will require more attention by that we mean it requires more pesticides than the crop one because pesticide has the repellent effect to insects which we proved above.

Crop Yield

Now we want to check the yield of a particular crop which we can find by checking the distribution of Crop type for which we are using seaborn countplot

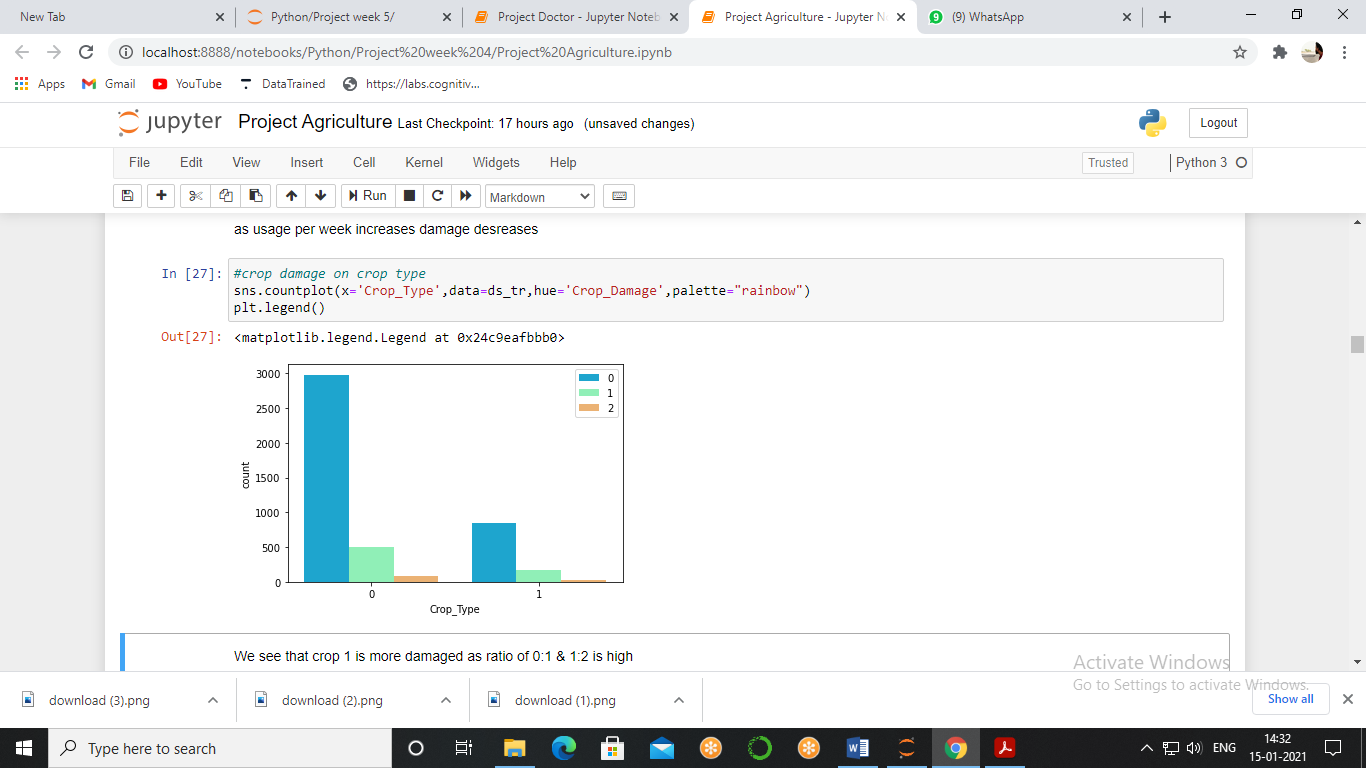


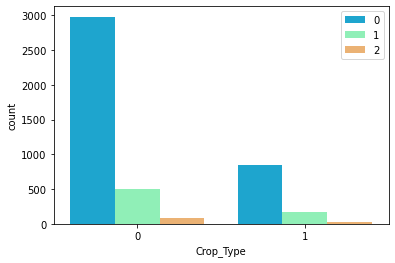


Looking at the graph we can clearly see that crop 0 has more production than the crop one but still the production of crops 0 can be further increased, and also we got no information about which type of crop should have been increased its production so we can't comment that *we should increase the production of Crop Type 1* our basic motivation is to increase the production by minimising the damage of any crop whatsoever.

Factors Affecting Damage/Production

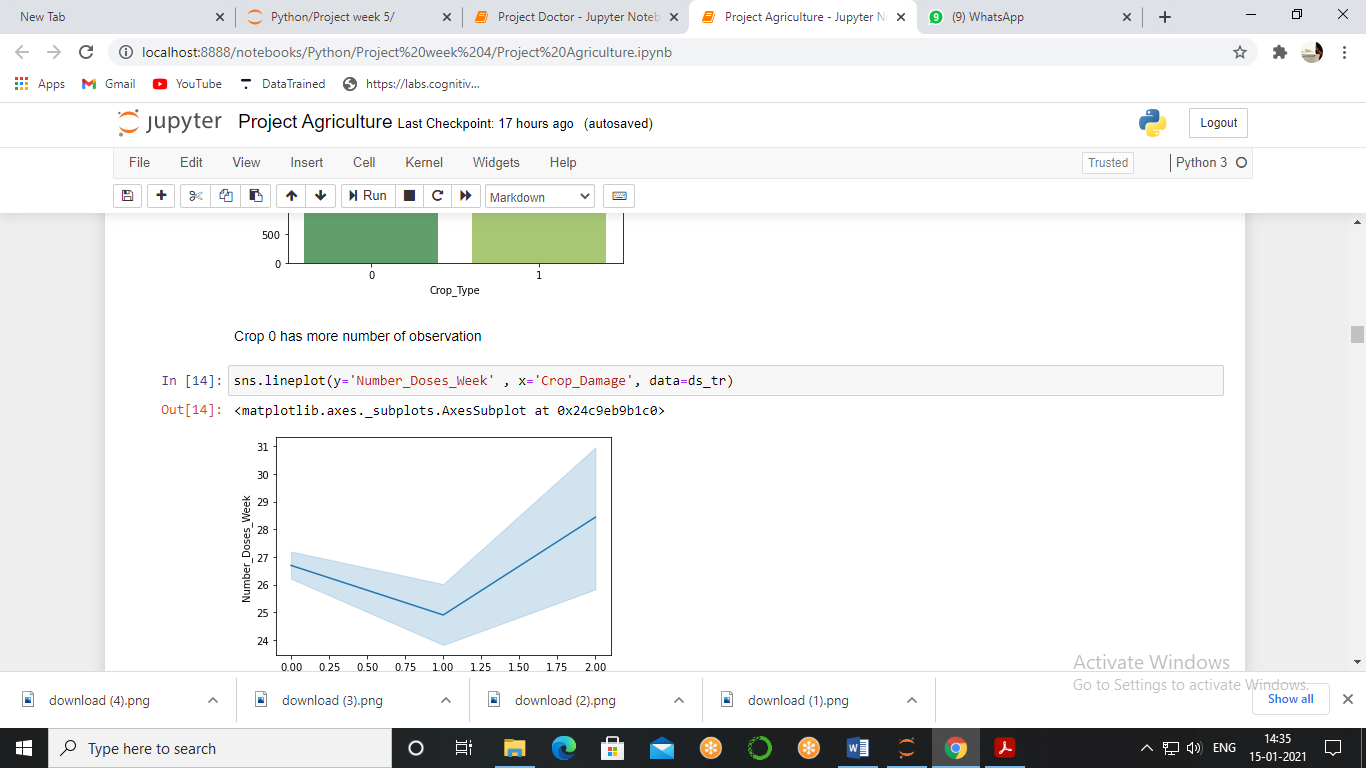
We will now check which type of crop is more profitable or we can say which type of Crop has less damage

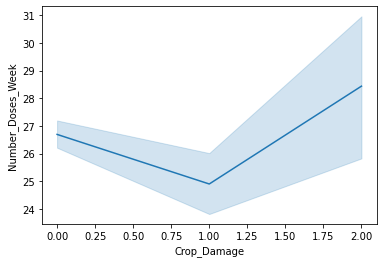




Before we can analyse further we should understand this thing that here we will not see the number of observations rather we will see the ratio between the damage of crops. We see that crop 1 is more damaged as the ratio of 0:1 & 1:2 is high.

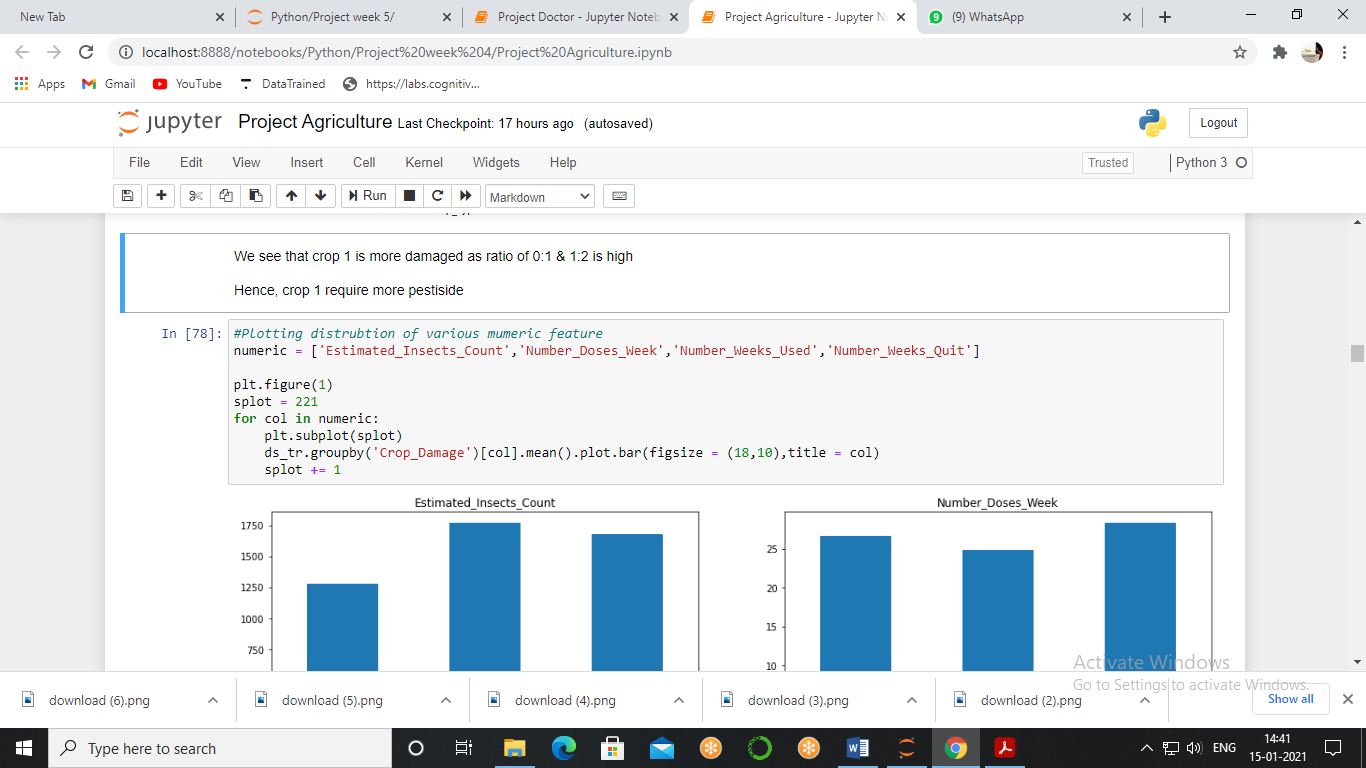
But now should we increase or decrease the use of pesticides per week? to answer these questions we will plot a graph to see the relationship between use of pesticides per week to damage





We see two properties of graph first decreasing then increasing but if we look at the initial and end point we see that the overall increment in the damage as the number of doses decreases the ideal amount is around 25 and both decreasing and increasing is damaging the crop.

Damage of crop is directly related to many factors so instead of plotting graph with each factor to ease it out we used to for loop and plotted all the graph at once

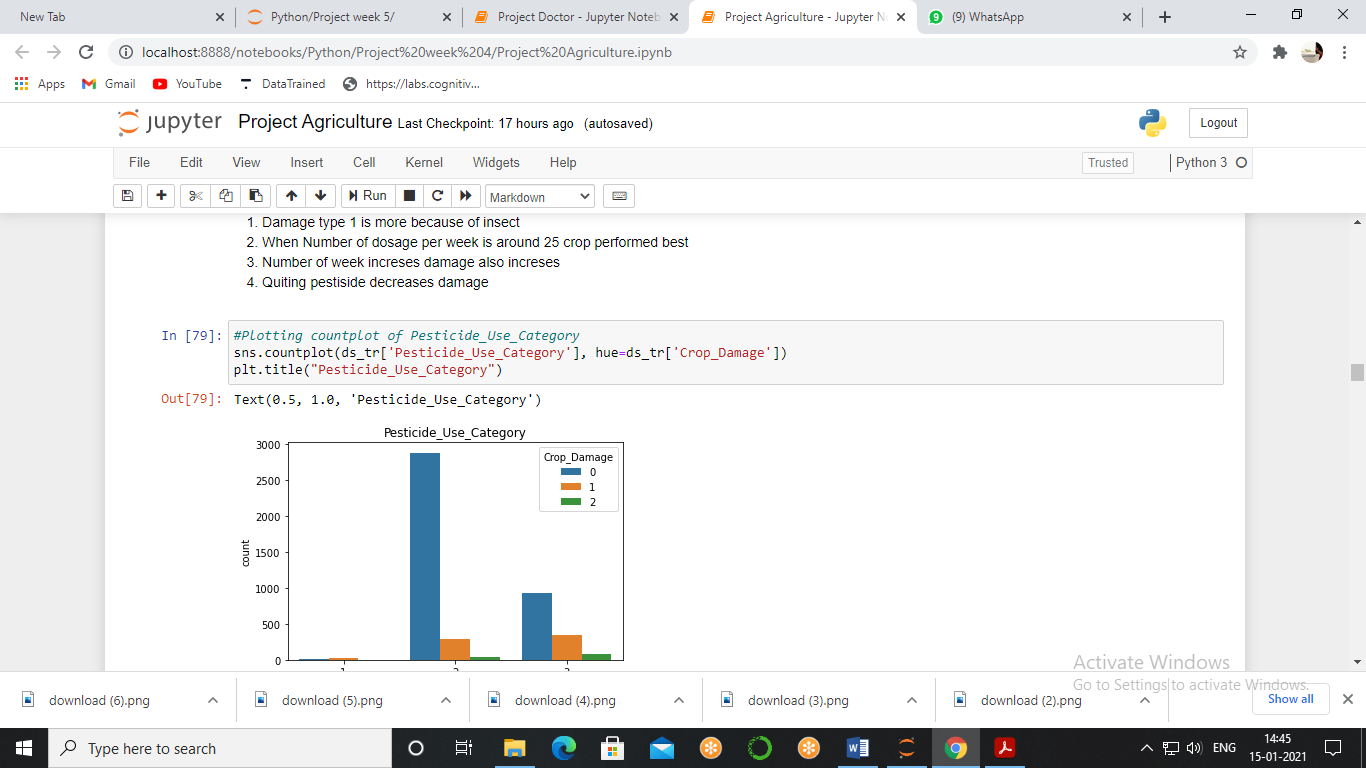


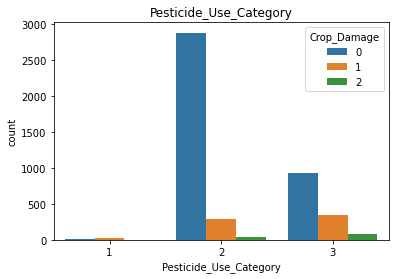


Here we see that with the increment in dose per week the type 2 damage is also increased which is basically a damaged because of pesticides second have the use of pesticides continued for number of weeks type 2 damage increased and third after quitting the pesticides type 2 damage decreased so the whole observation states that the excessive use of pesticides destroy the crop.

Pesticide Usage

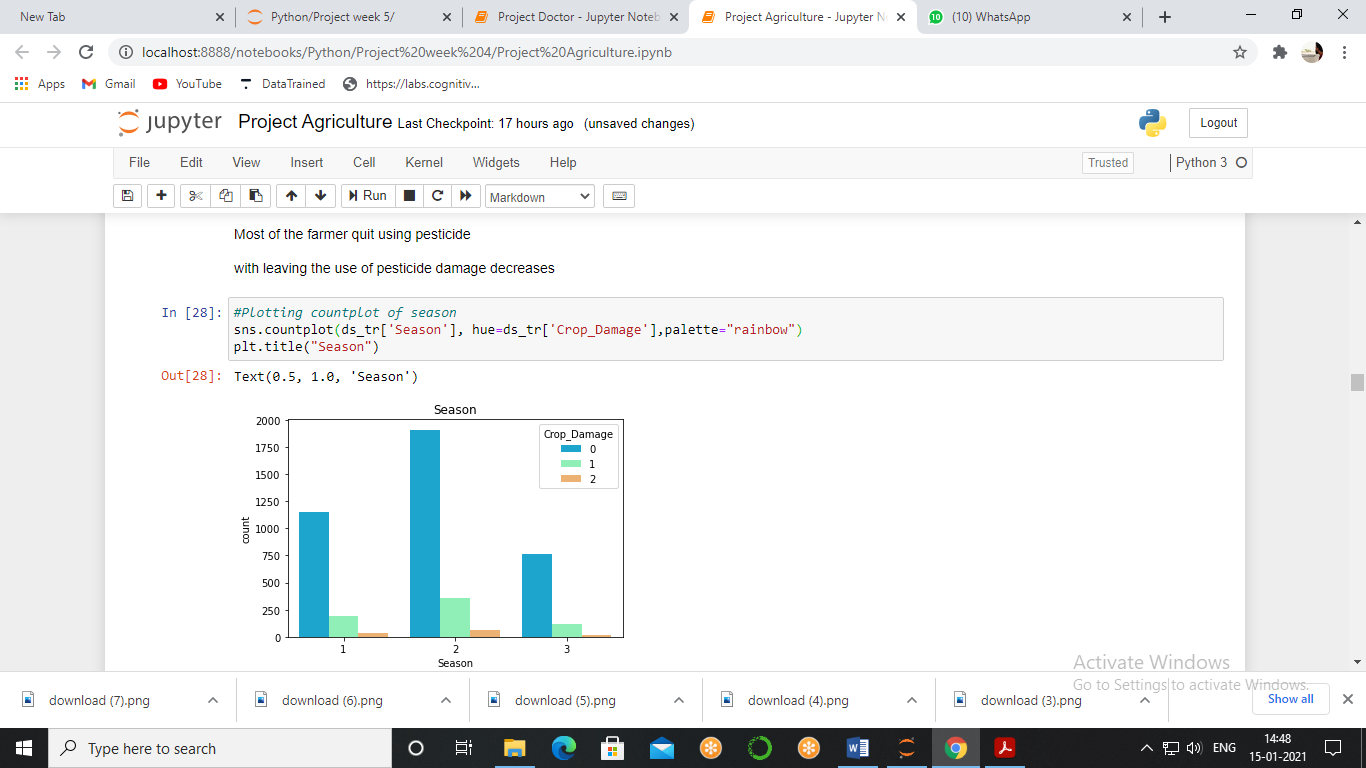
We are done with pesticides and its damage, now we want to see the overall usage of pesticides and also its effect on crops, which can be seen by Pesticide\_Use\_Category where 1- Never, 2-Previously Used, 3-Currently Using.

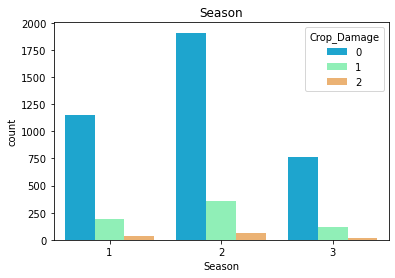




We see that most of the farmer quit using pesticides, the farmers who never used pesticides indulge with more damage type 1, when the use of pesticides begins damage type 3 i.e. because of pesticides increased but damage type 2 i.e. because of other factors decreased and when farmed quit using pesticides suddenly the damage decreased so much. Therefore, the periodical use of regulated amounts can boost productivity.

Now, all the factor we took into consideration above are human controlled factors but there is one factor which is not controlled by human i.e. season, so, we will look the role of season on damage.

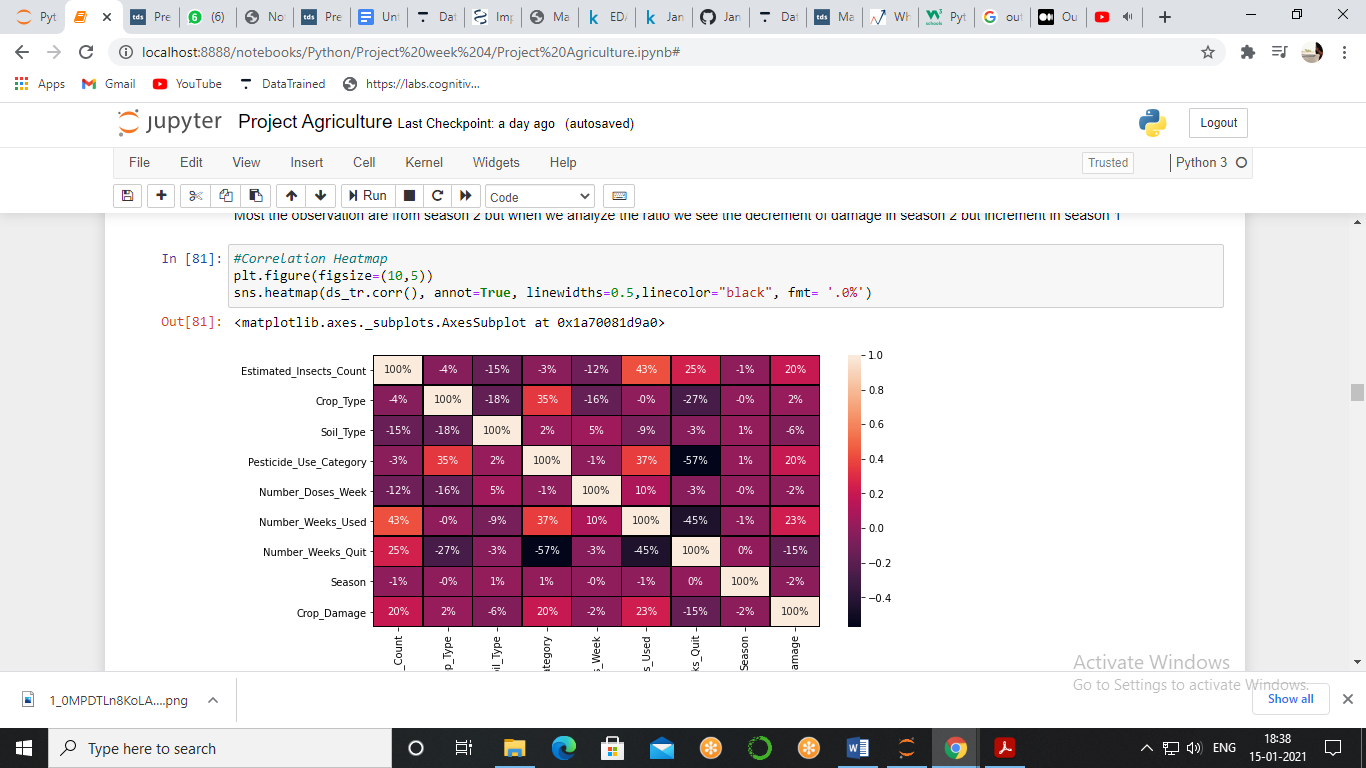


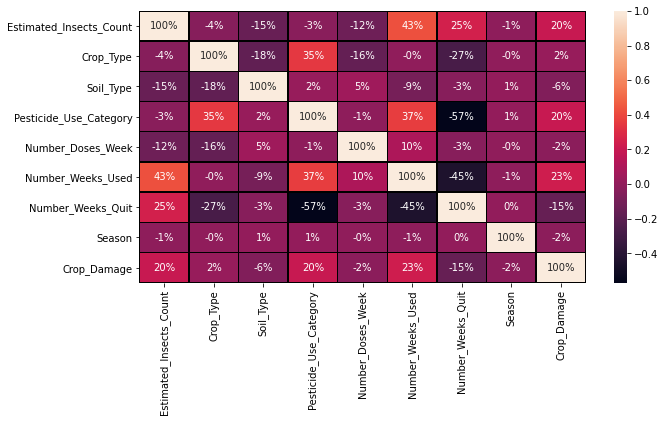


When we analyse the ratio we see the decrement of damage in season 2 but increment in season 1 and 3.

Relationship of Factors

The correlation heatmap. Correlation matrices are an essential tool of exploratory data analysis. Correlation heatmaps show in a glance which variables are correlated, to what degree, in which direction, and alerts us to potential multicollinearity problems.

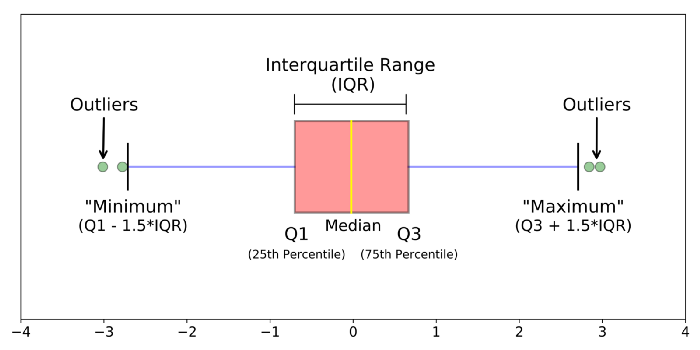




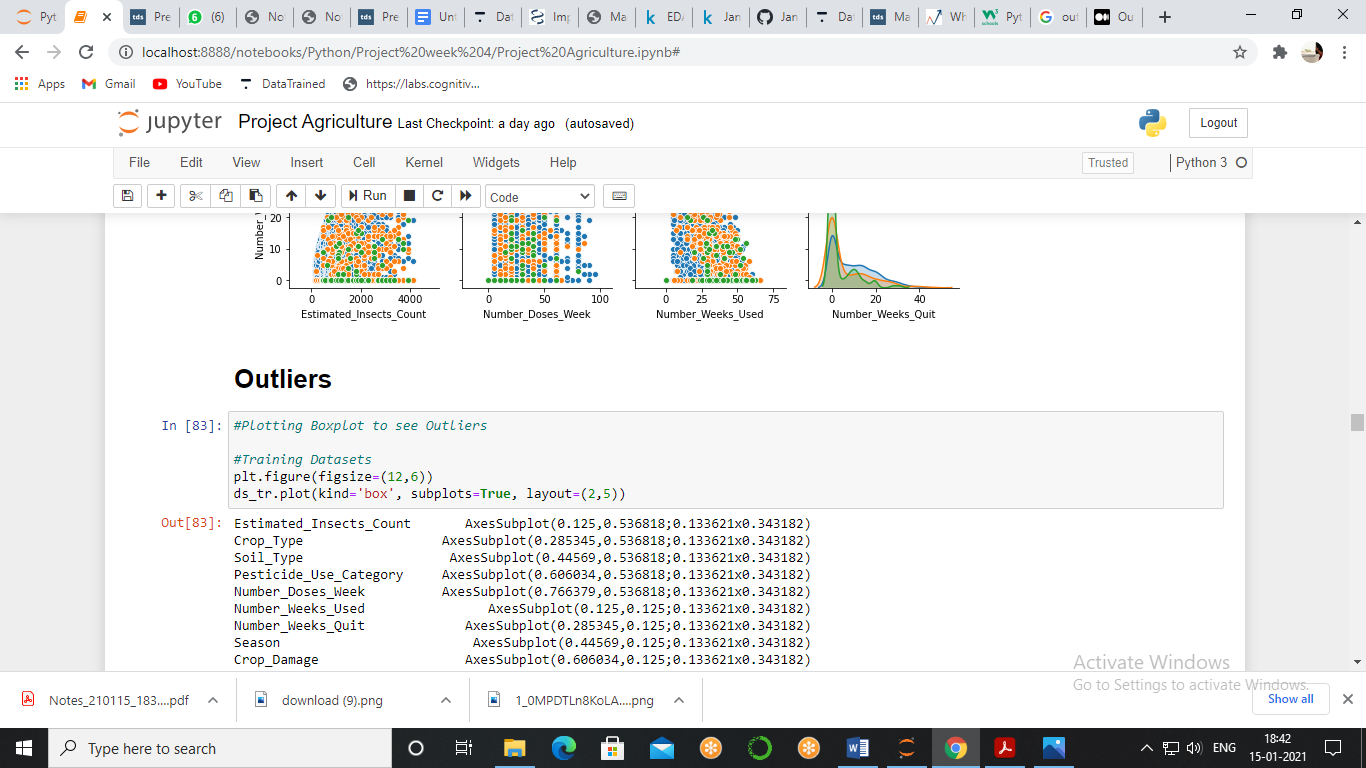
* High Correlation= Insect count, Use Category, Number of week used
* Mild Correlation= Crop type
* Low Correlation = Soil Type, Number of Dosage per week, Number of Week Quit, Season

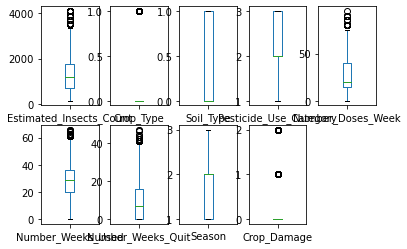
Outliers

Boxplot *—*displays a five-number summary of a set of data, this type of plot is used to easily detect outliers. It can also tell us if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.



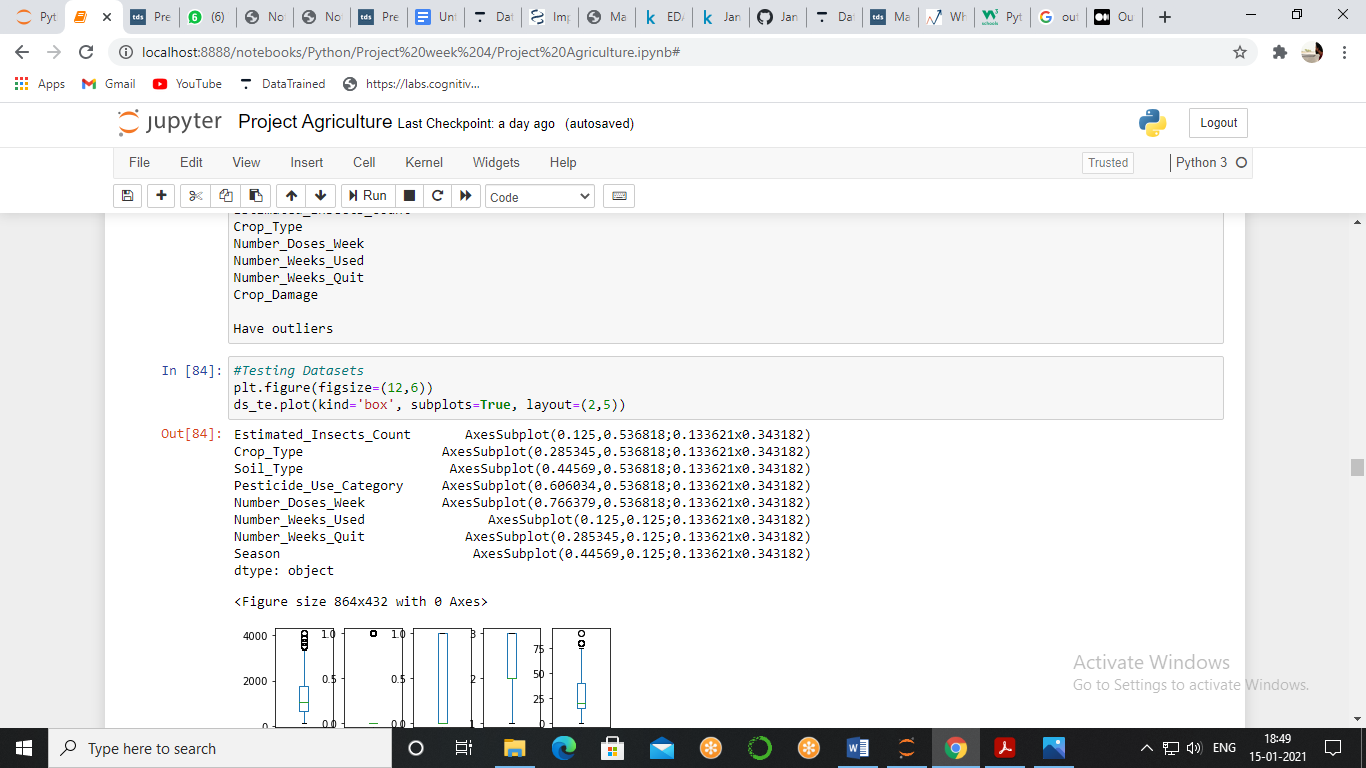
To check the outliers, we use box plot and with the use of subplot we will plot all the boxplot in One Frame.

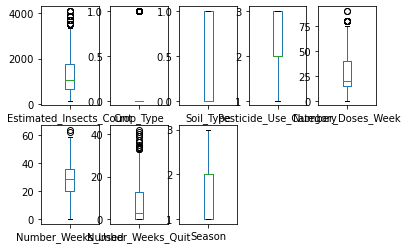




After looking at the boxplot of train dataset it is clearly shown that all the data above fourth quartile is outlier and Crop\_Type, Number\_Doses\_Week, Number\_Weeks\_Used, Number\_Weeks\_Quit, Crop\_Damage Have outliers but we will not remove the outliers of Crop Damage as it is our target variable.

Now, we are checking the outliers of Test Dataset





After looking at the boxplot of Test dataset it is clearly shown that all the data above the fourth quartile is outlier and Crop\_Type, Number\_Doses\_Week, Number\_Weeks\_Used, Number\_Weeks\_Quit Have outliers.

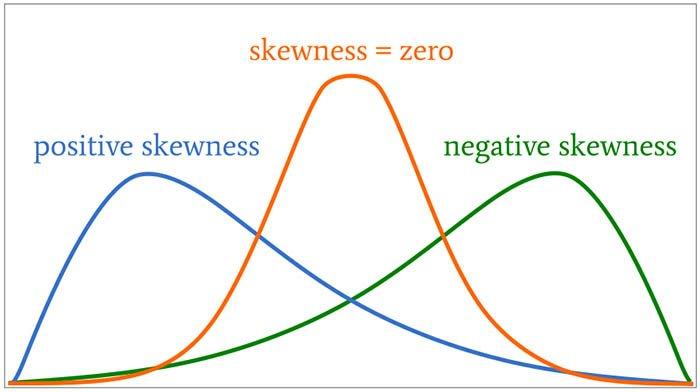
To remove the outliers, we use a statistical method known as Z score. We will find the value of Z score, we consider threshold value of Z score is +- 3, above and below it all the values are considered as outliers so we will check all the rows which Z score is above absolute 3.



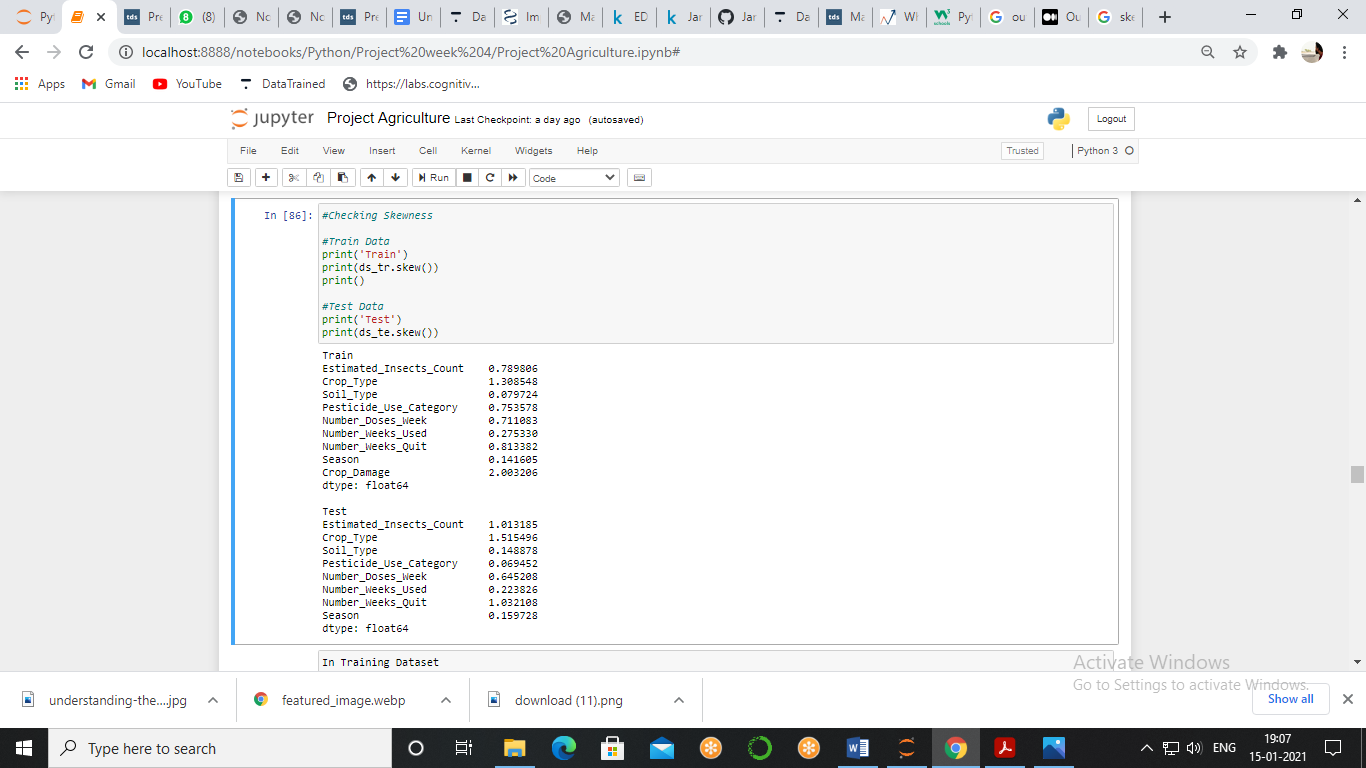
Now there are many ways to remove an outlier either replace them with mean, median or mode or we can simply remove the rows which have outliers, here we consider the removing row option because outliers are less and it will not affect the data.

Skewness

Skewness is the measure of how much the probability distribution of a random variable deviates from the [normal distribution](https://www.analyticsvidhya.com/blog/2020/04/statistics-data-science-normal-distribution/?utm_source=blog&utm_medium=what-is-skewness-statistics)

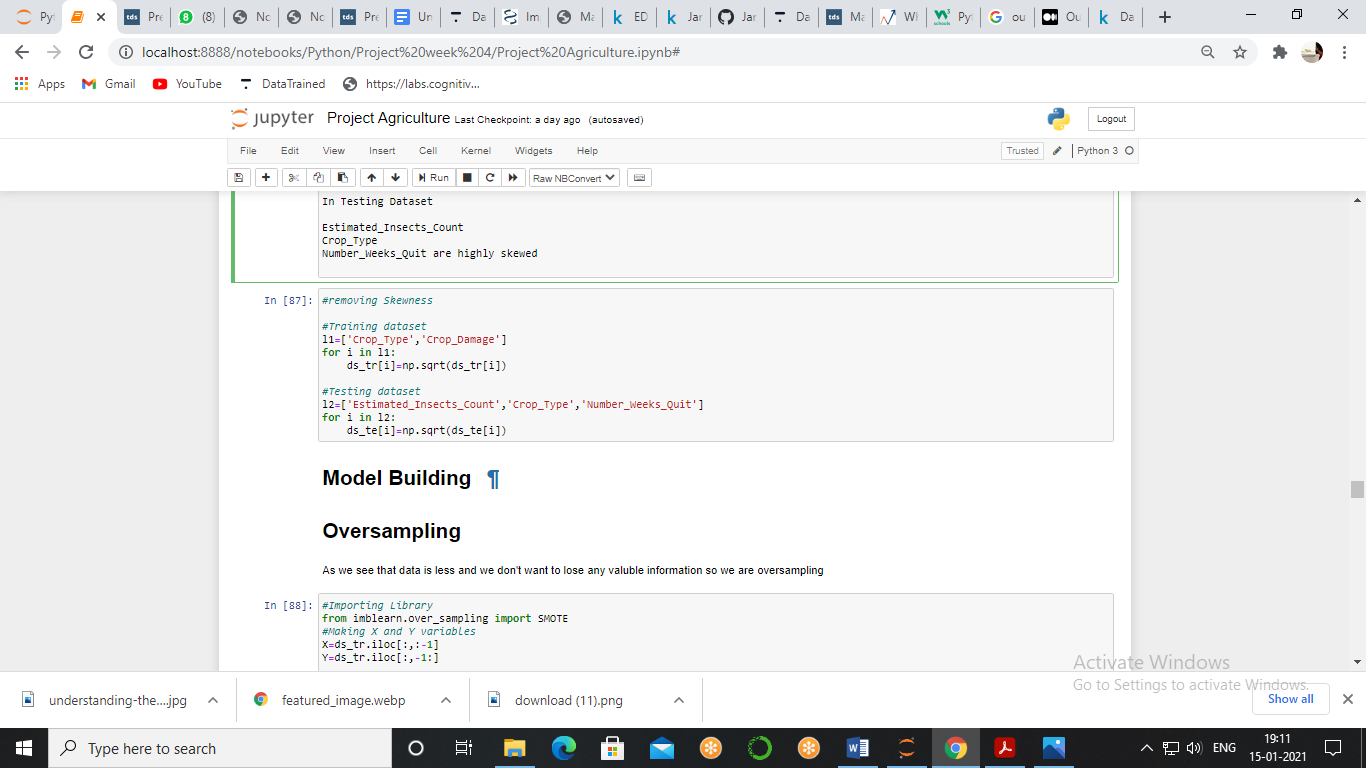


Skewness can be seen by using a simple one-line code Dataset. Skew(), If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.



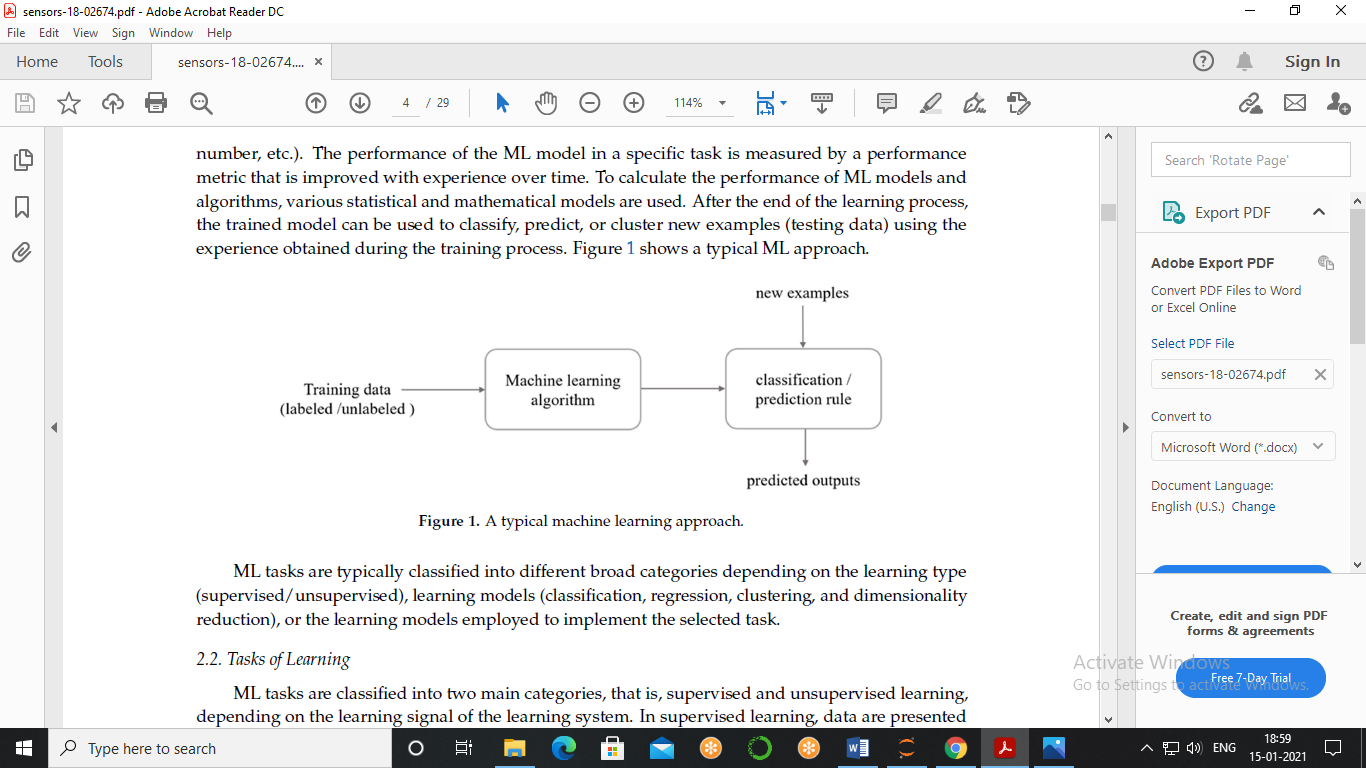
In Training Dataset Crop Type and Crop Damage are highly skewed while In Testing Dataset Estimated\_Insects\_Count, Crop\_Type, Number\_Weeks\_Quit are highly skewed.

As our data value is small plus it has 0 values too it is safe to use the square root method, in this method we will take square root of all the variables with skewness.



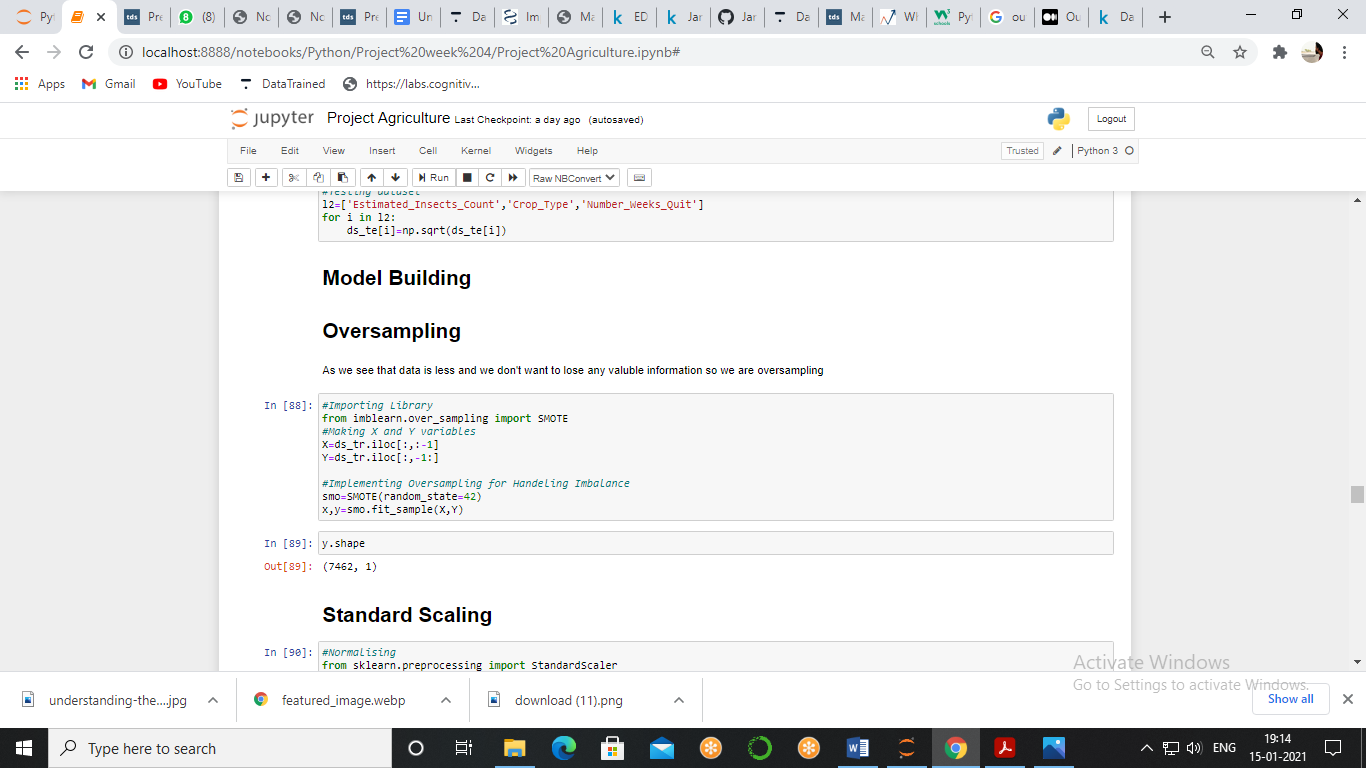
Machine Learning Model

Agriculture plays a critical role in the global economy. Pressure on the agricultural system will increase with the continuing expansion of the human population. Agri-technology and precision farming, now also termed digital agriculture, have arisen as new scientific fields that use data intense approaches to drive agricultural productivity while minimizing its environmental impact.



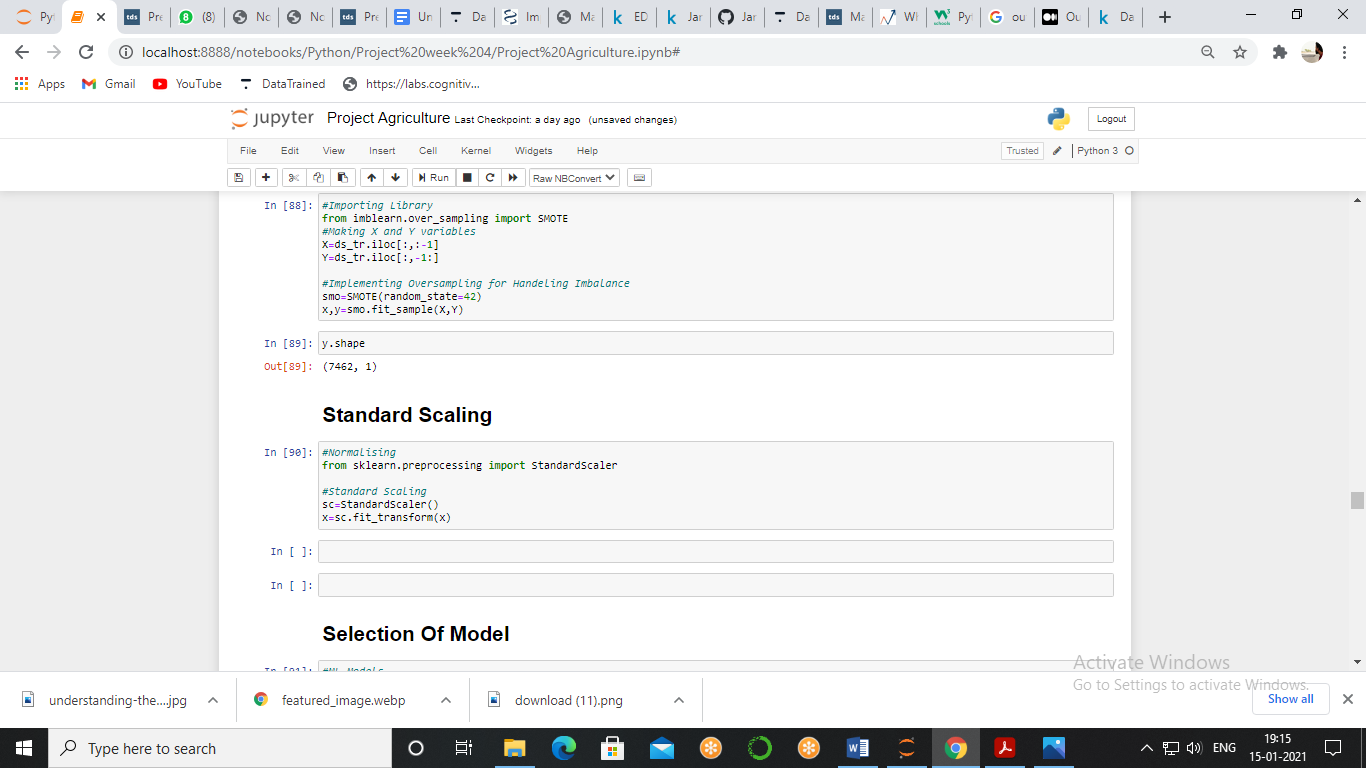
First we will have separated data into X and Y where our X will be the input data and Y will be the target variable, as our target variable is categorical type.

When we see the distribution on the first graph we analysed the distribution was highly imbalance so to improve the result, we will use sampling moreover we will choose over sampling because imbalance ratio can be seen around 70:30, so, we choose over sampling rather than under sampling.

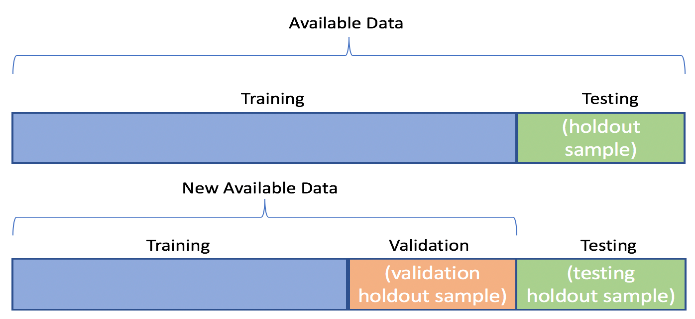


As the rows of our data are increased hands over sampling is done.

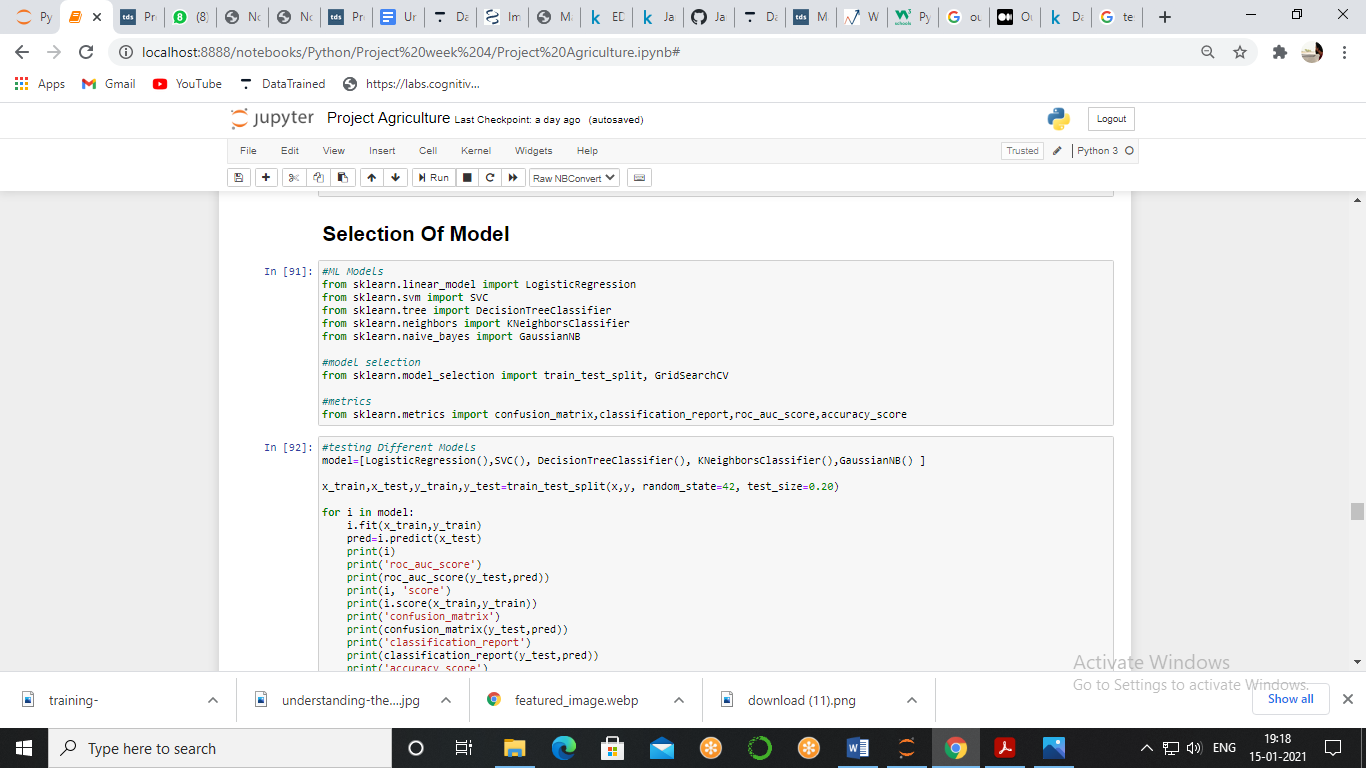
**Scaling** - Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation, here we use standard scaler for standardization.



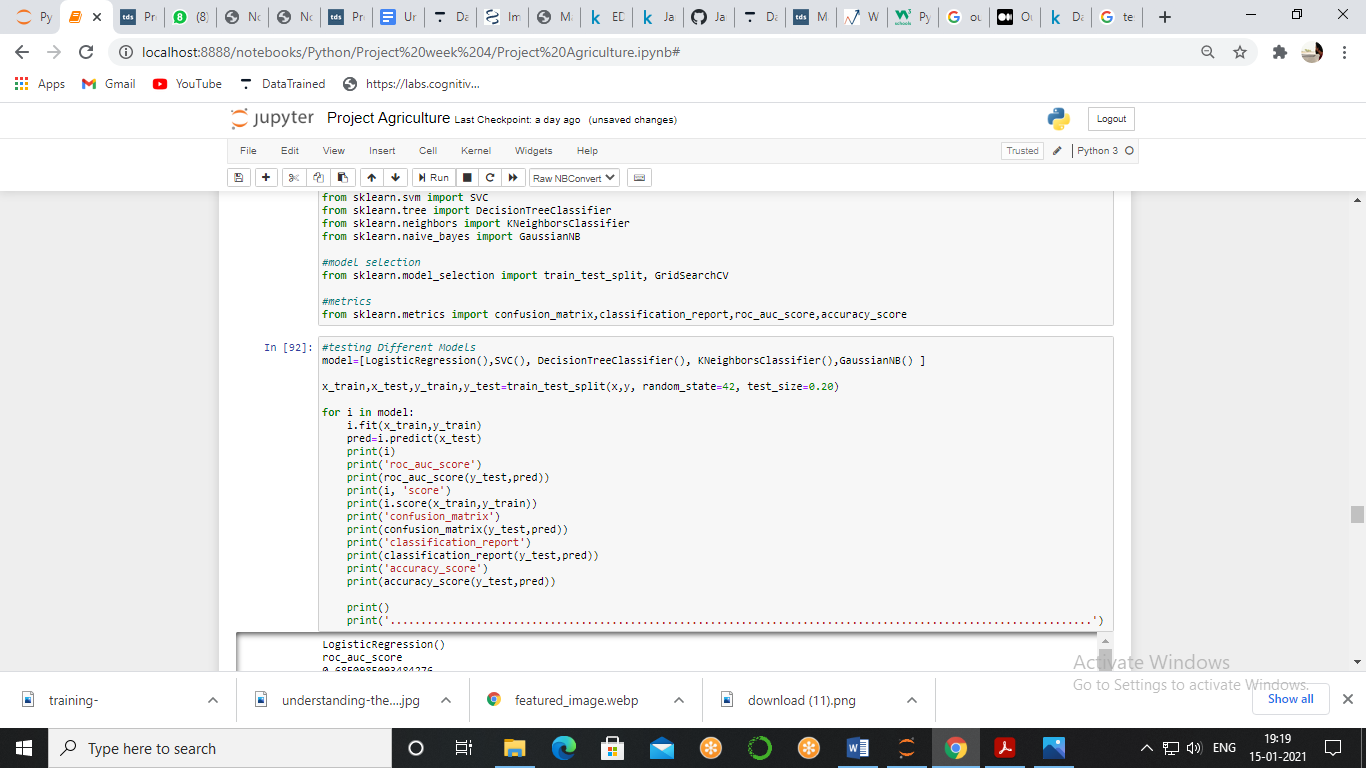
Now we will be importing libraries of machine learning to make our model, first we will import classification models then we will import a library to select a model and that will be train test split this library will split a model into four categories.



Then we will import mattresses which will tell the performance of our model initially our model was imbalanced type but after over sampling our model became balanced, so, here we should take mainly 2 matrices into consideration one is accuracy score and second is auc roc curve.



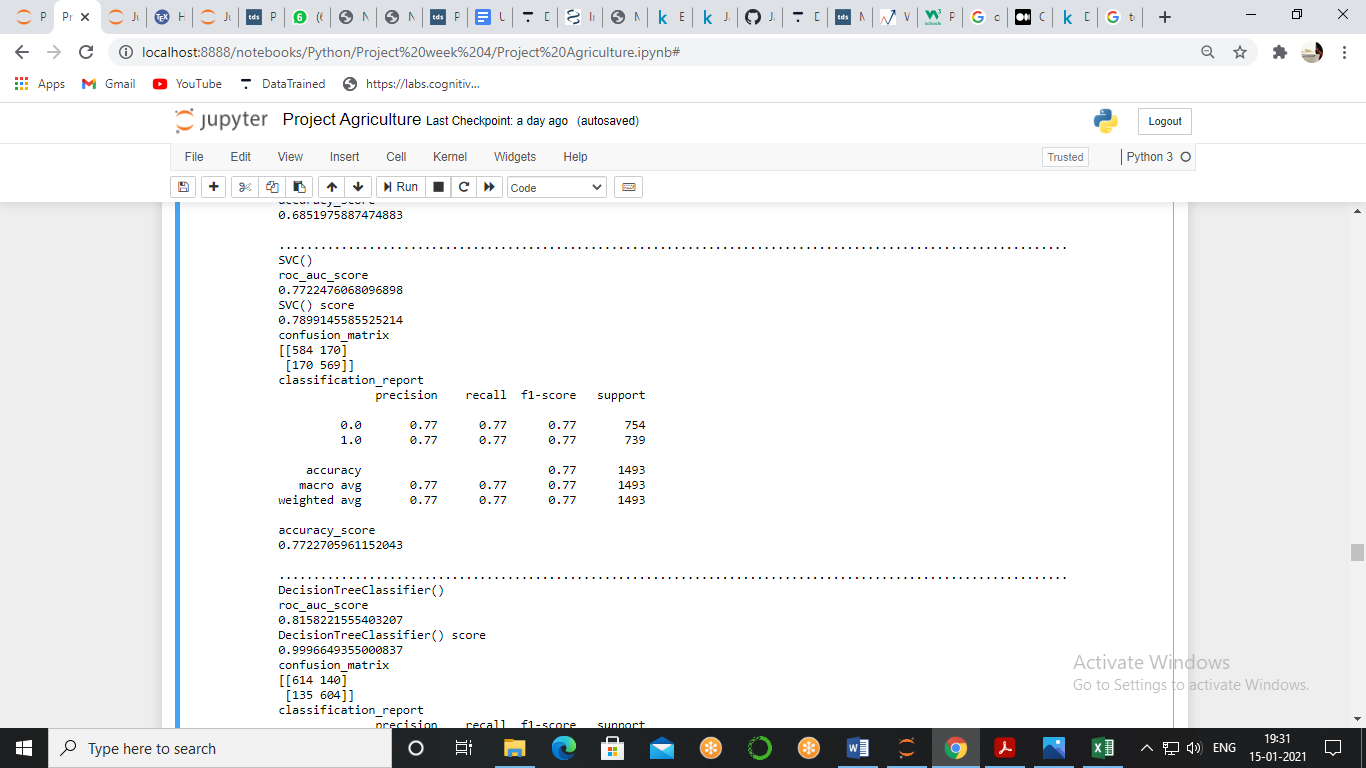
To check all models at once we made a list of callable name of the models and pass it to the for loop result of which we will get the best of all the models which will make us easy to choose the best model.



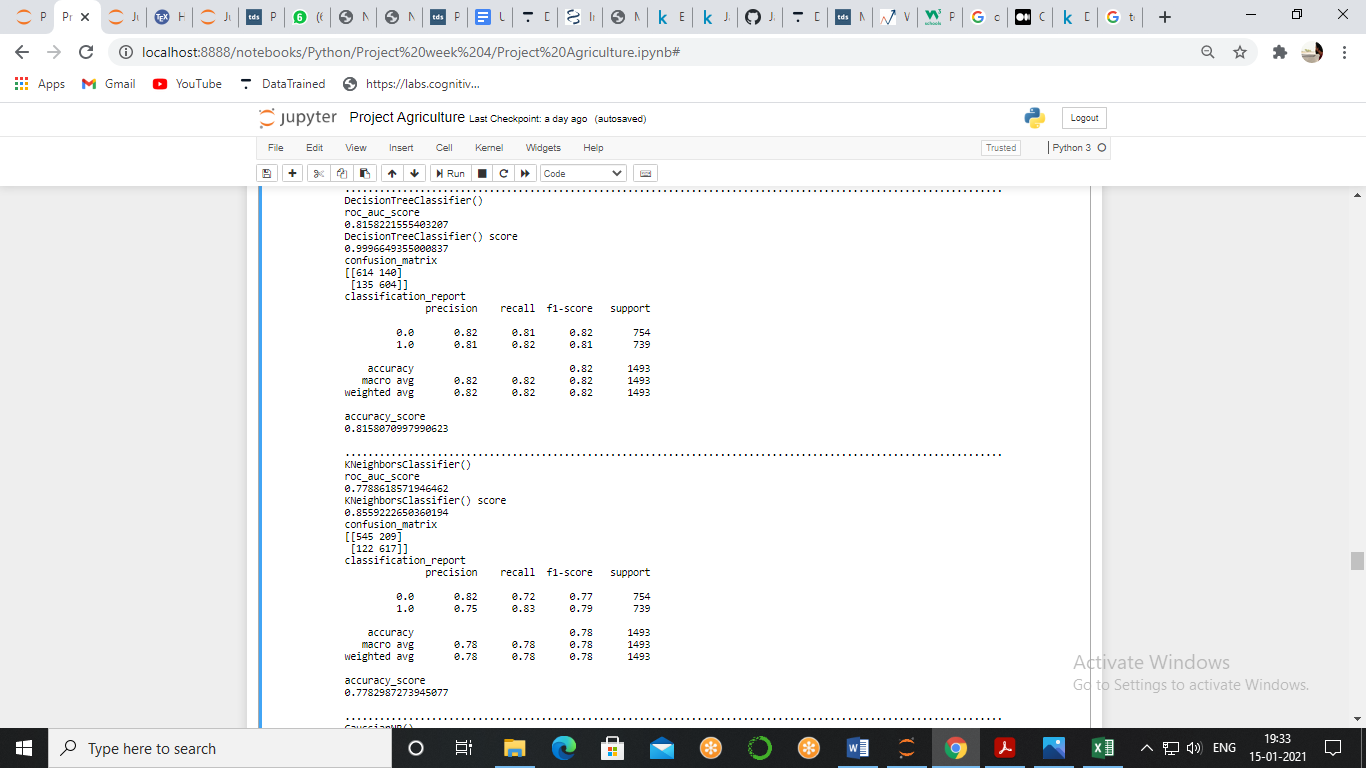
**Logistic Regression**



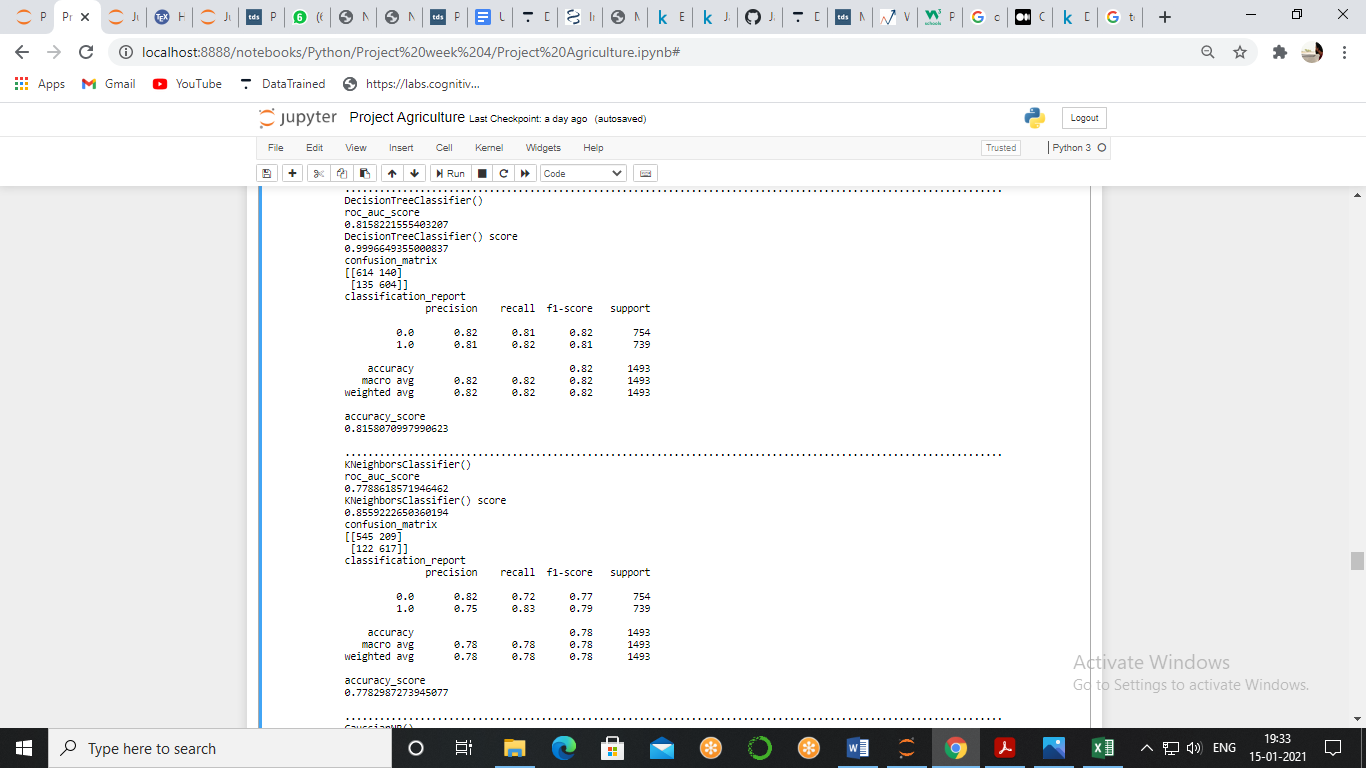
**Support Vector Classifier**



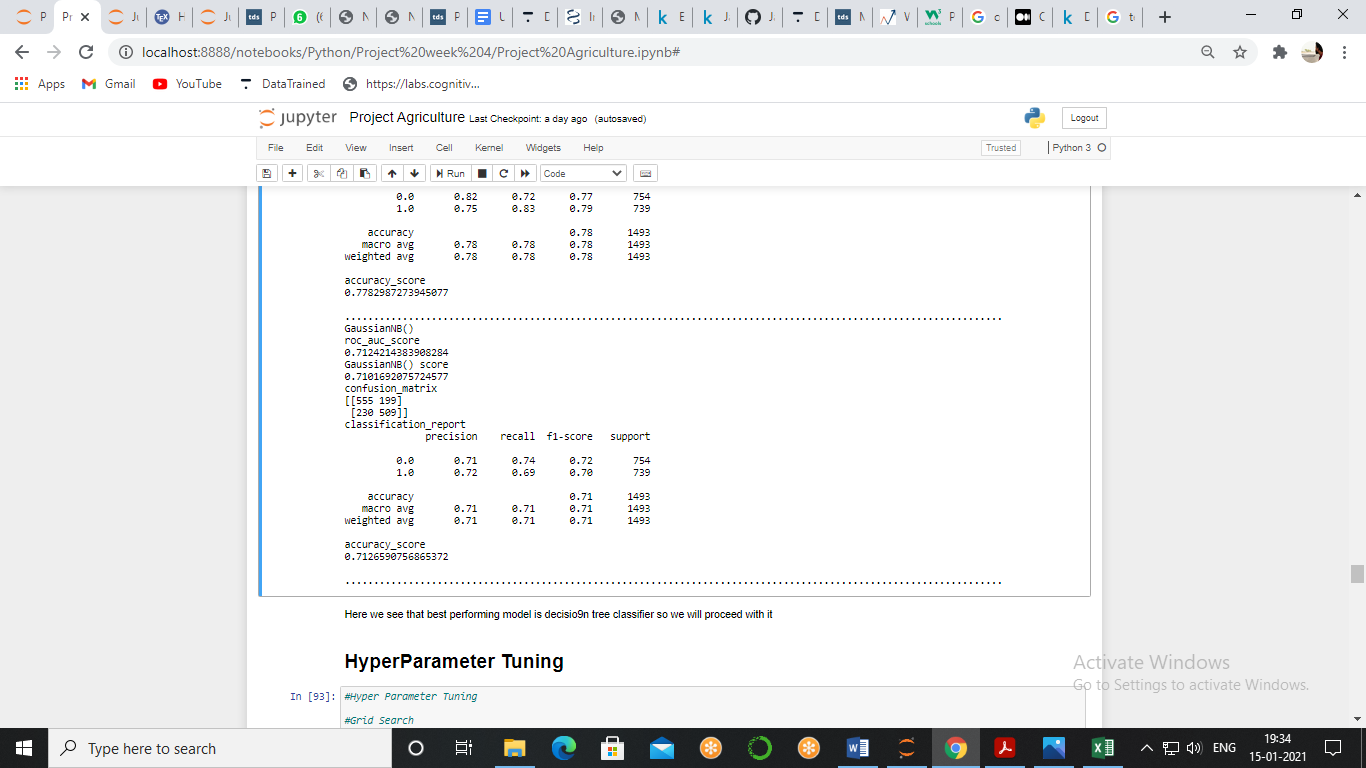
**Decision Tree Classifier**



**K- Neighbours Classifier**

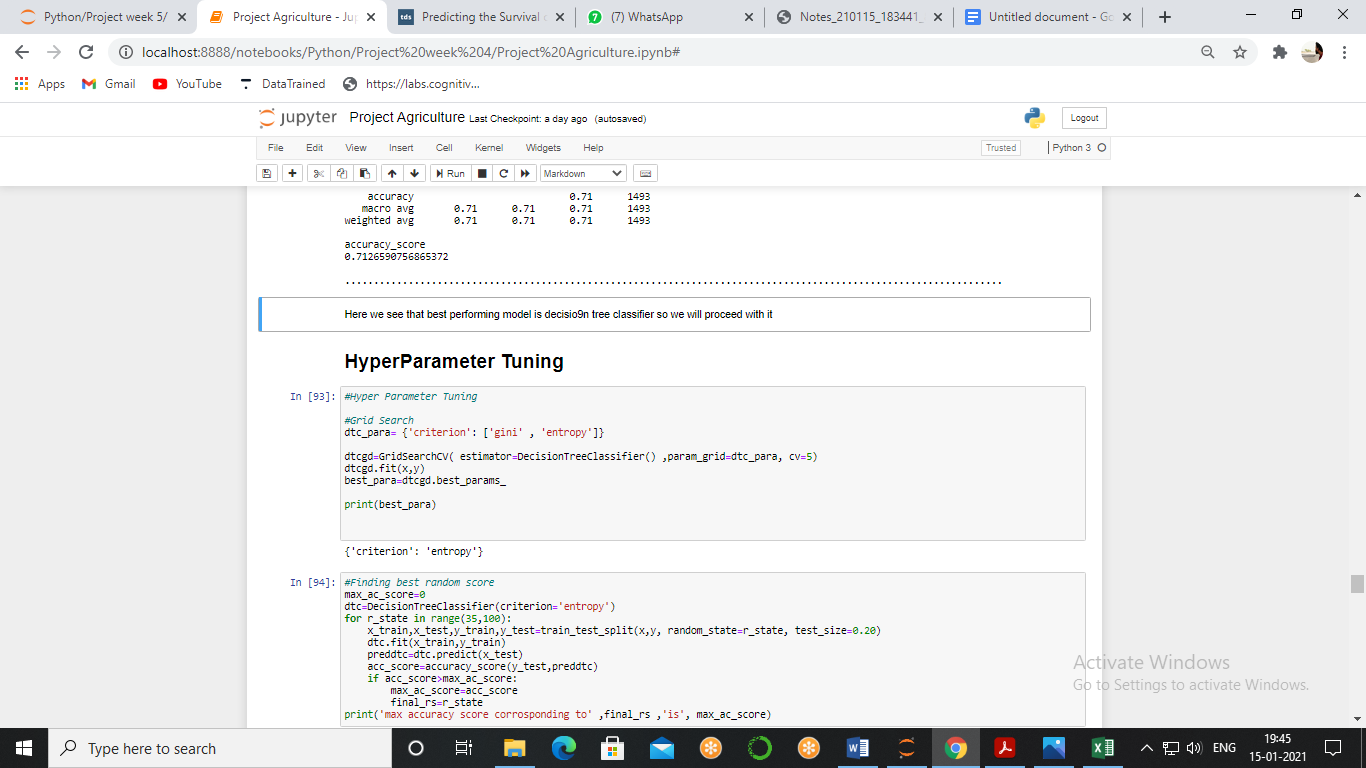


**Gaussian Naive Bayes**

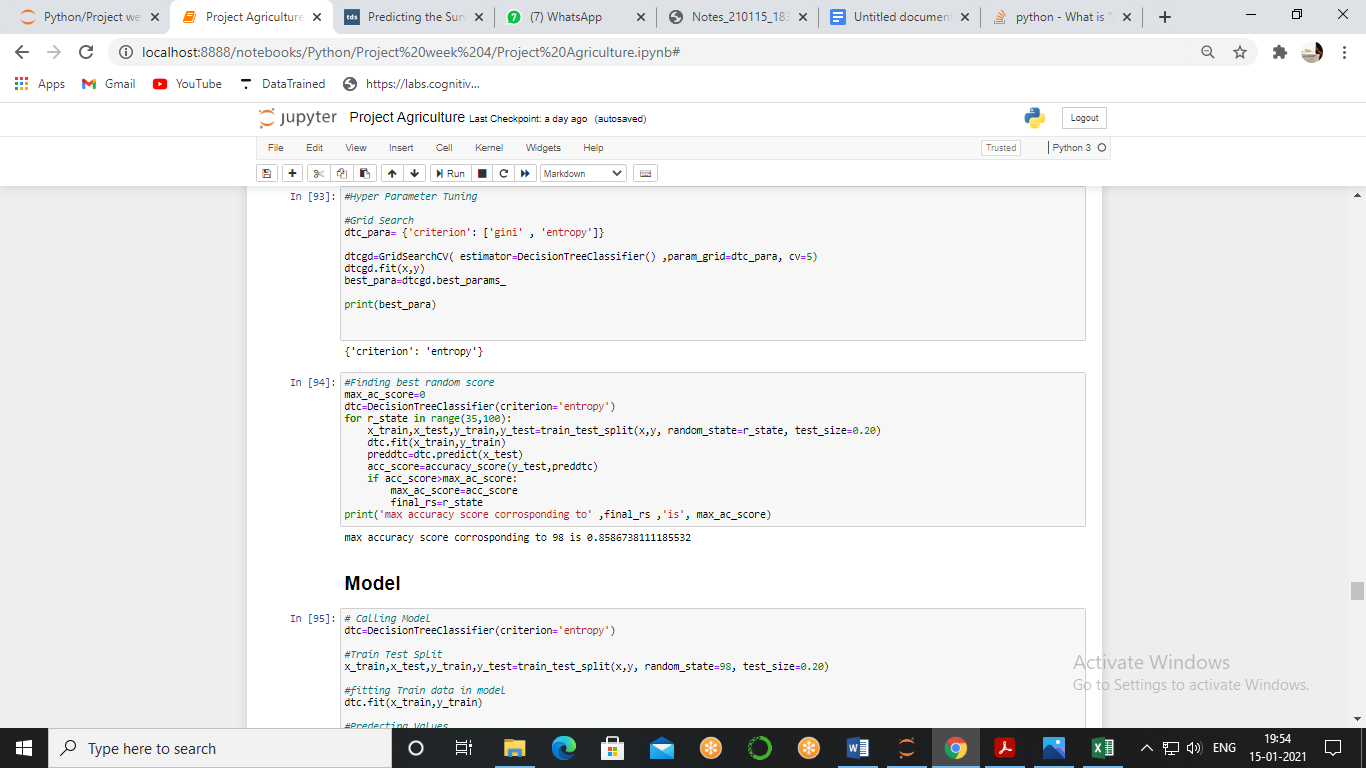


Where we see that we get the best accuracy score, auc - roc score as well as F1 score in decision tree classifier hence we will proceed with this model.

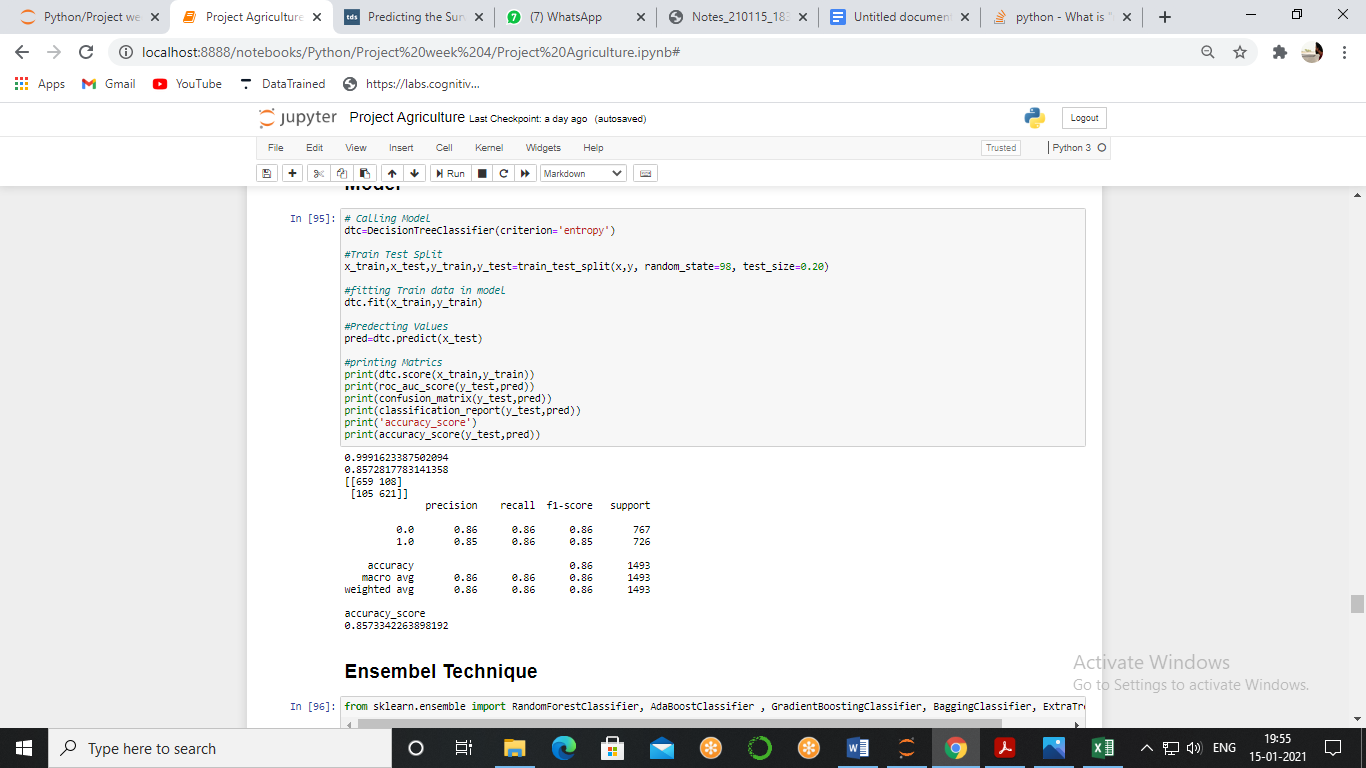
There are so many parameters which we apply on a model to get the best result and selecting that parameter is called hyper parameter tuning and for that the best tool we have is grid search it will internally test all the parameter given by and give us the best parameter



We see that Criterion – Entropy is the best parameter to make the model but there is still one more parameter, Random State, simply sets a seed to the random generator, so that your train-test splits are always deterministic. If you don't set a seed, it is different each time, to get the best random score we pass range of score in for loop and get the best score.



Now, we got best parameters so it's time to build our model with best parameter

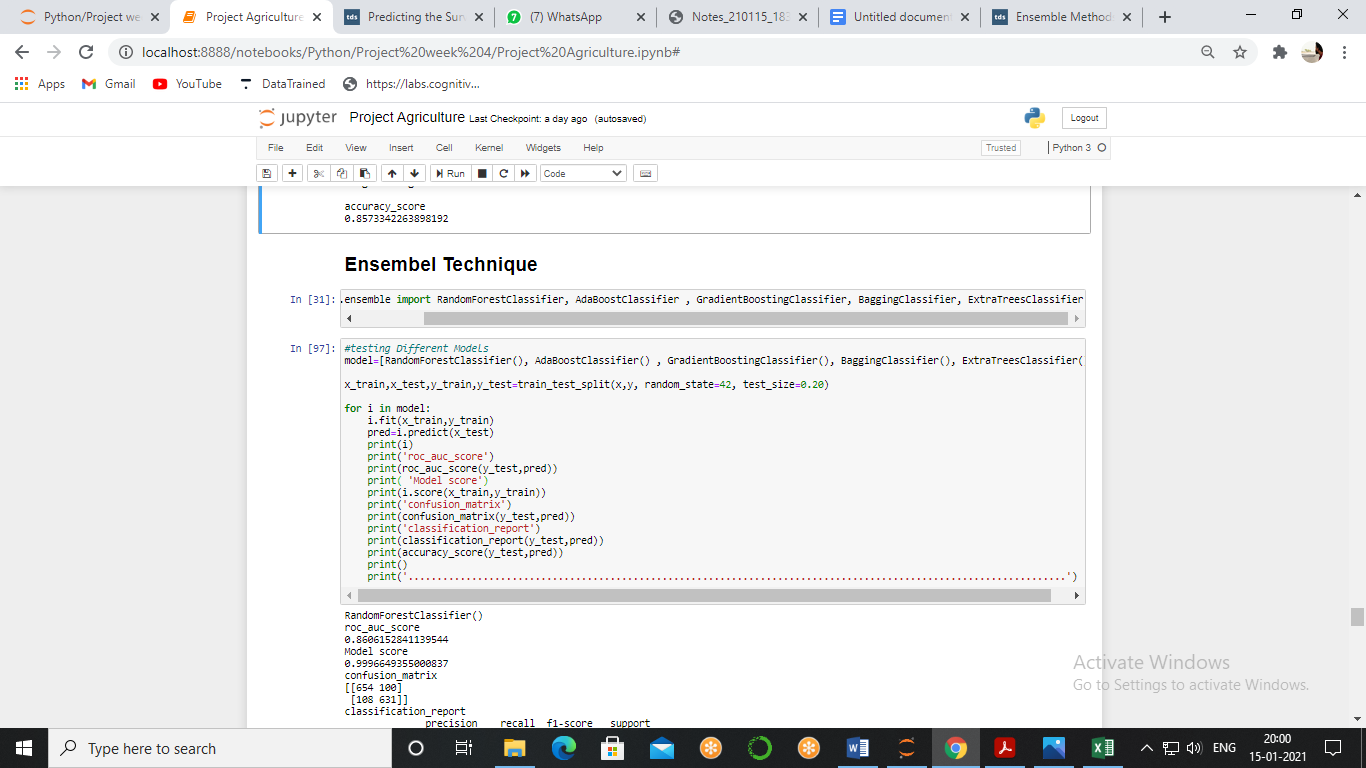


When we made a model with a decision tree classifier with Criterion = Entropy and Random State = 98, then we got Accuracy Score = 85.73%, Auc\_Roc Score = 85.72%, Precision = 85%, Recall = 86%. Although we got a good matrix score still there is a room of improvement even if there is not, still we can check the ensemble methods and see how much power performance can increase.

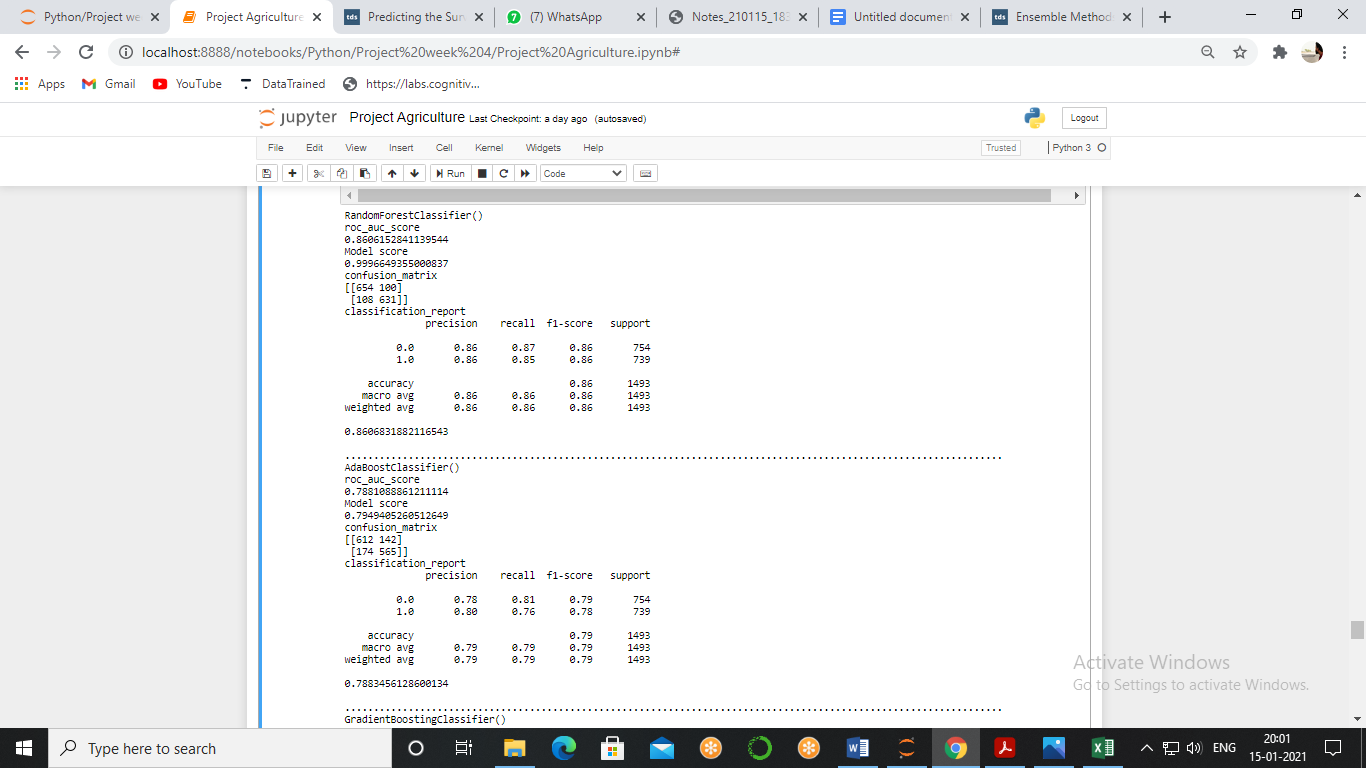
Ensemble methods

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

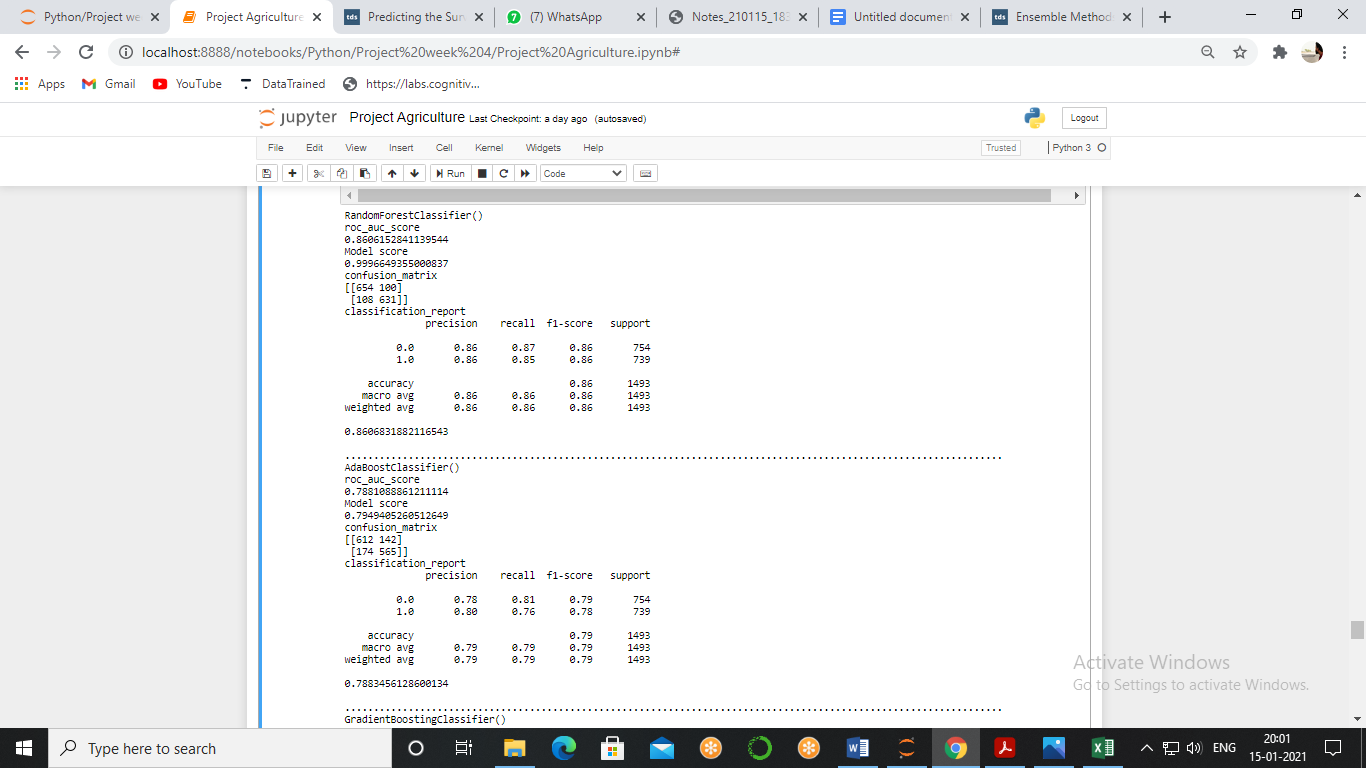
We will test different methods using for on our data and will compare the matrices



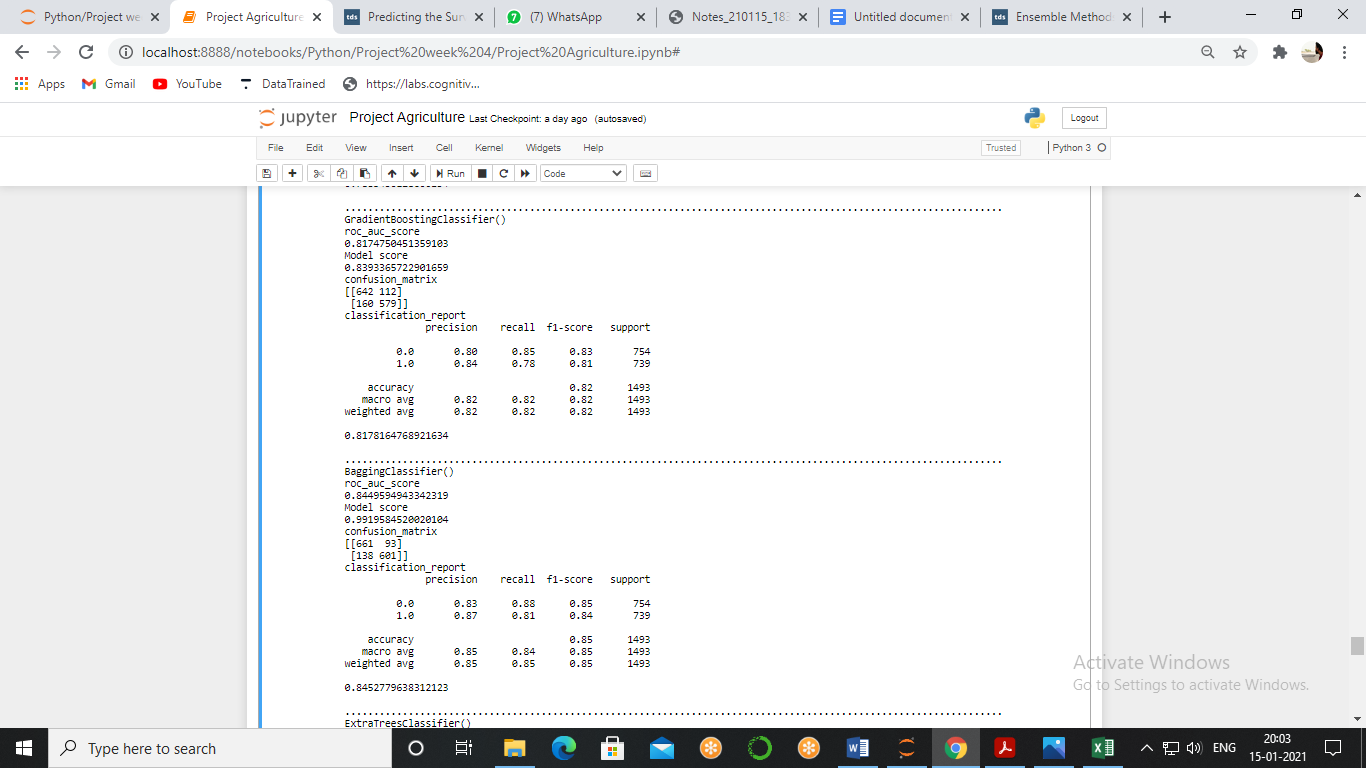
**Random Forest Classifier**



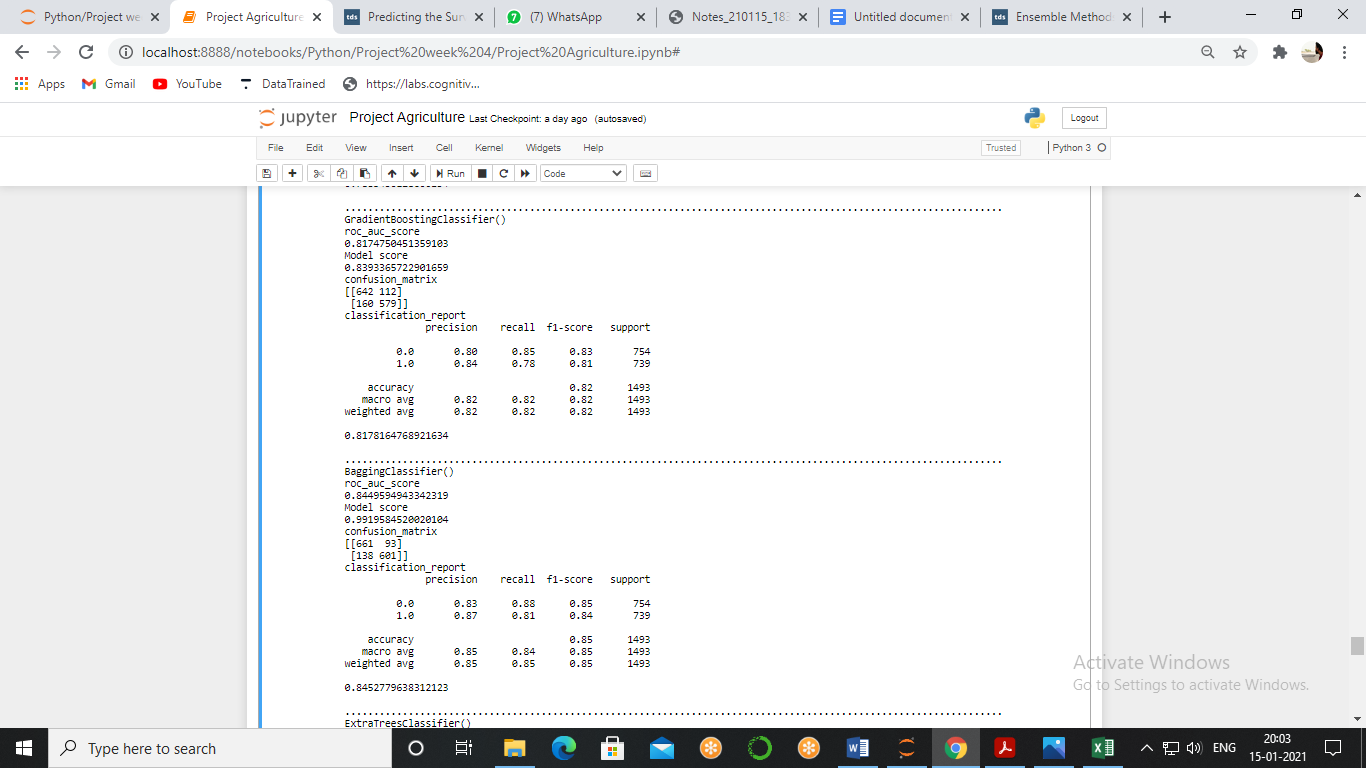
**AdaBoost Classifier**



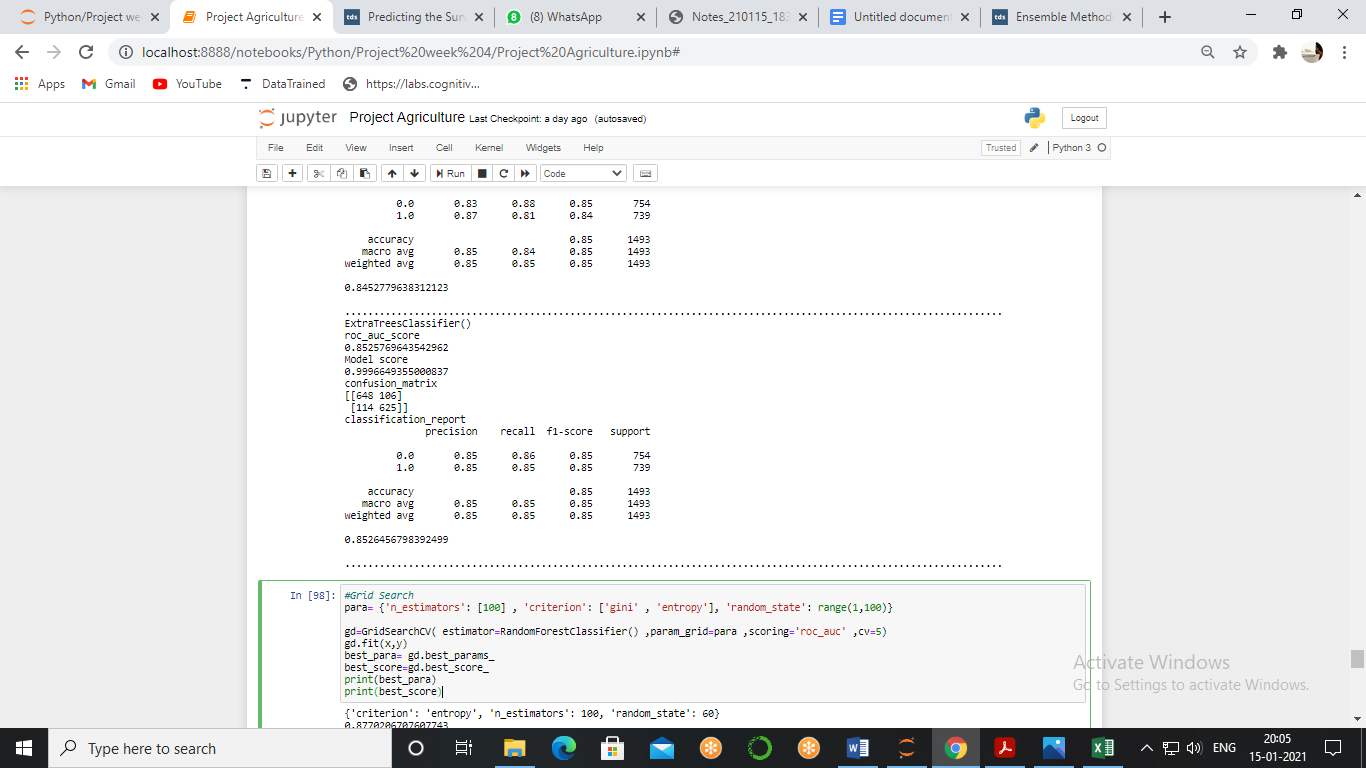
**Gradient Boosting Classifier**



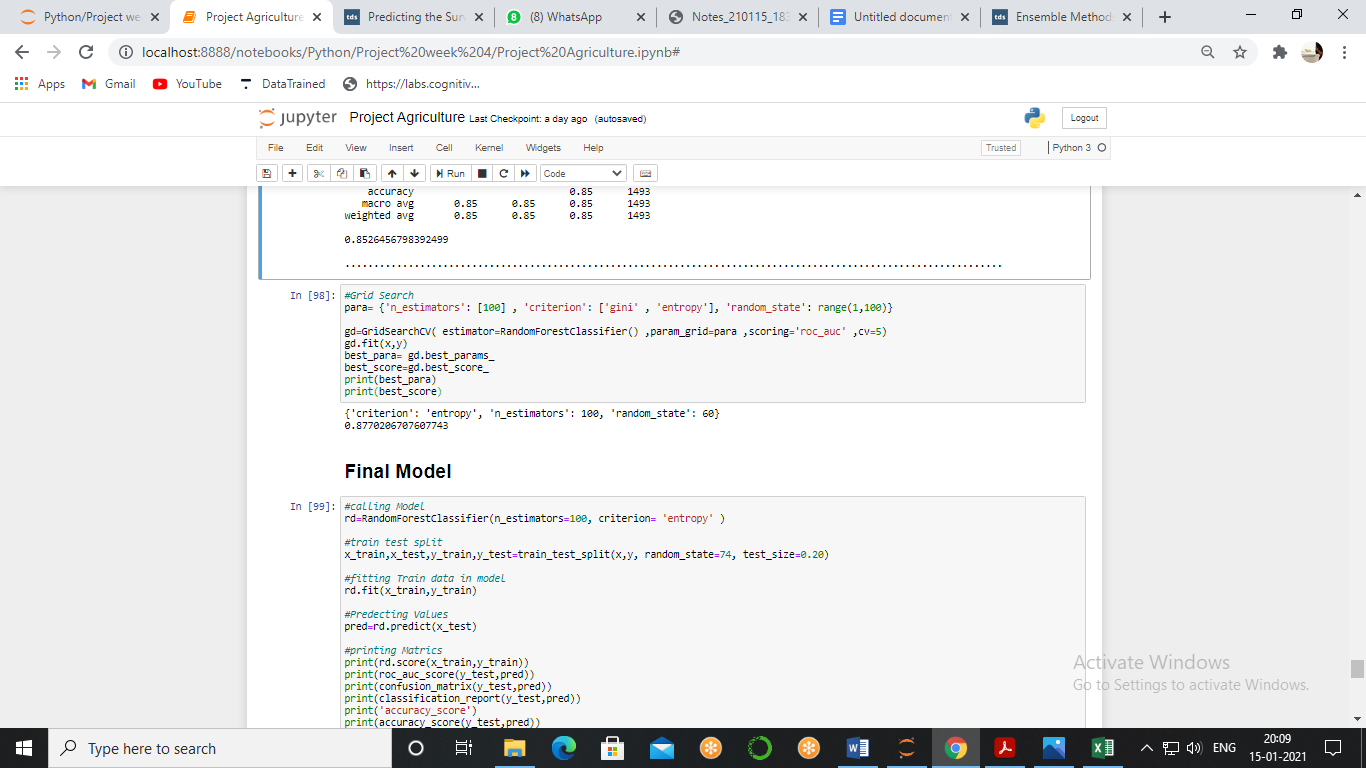
**Bagging Classifier**



**Extra Trees Classifier**



We got almost similar results in many model but we will choose Random Forest Classifier as its TN and FP values are less as well as its accuracy score is higher than others, now, we will have to do the hyper parameter tuning using grid search.



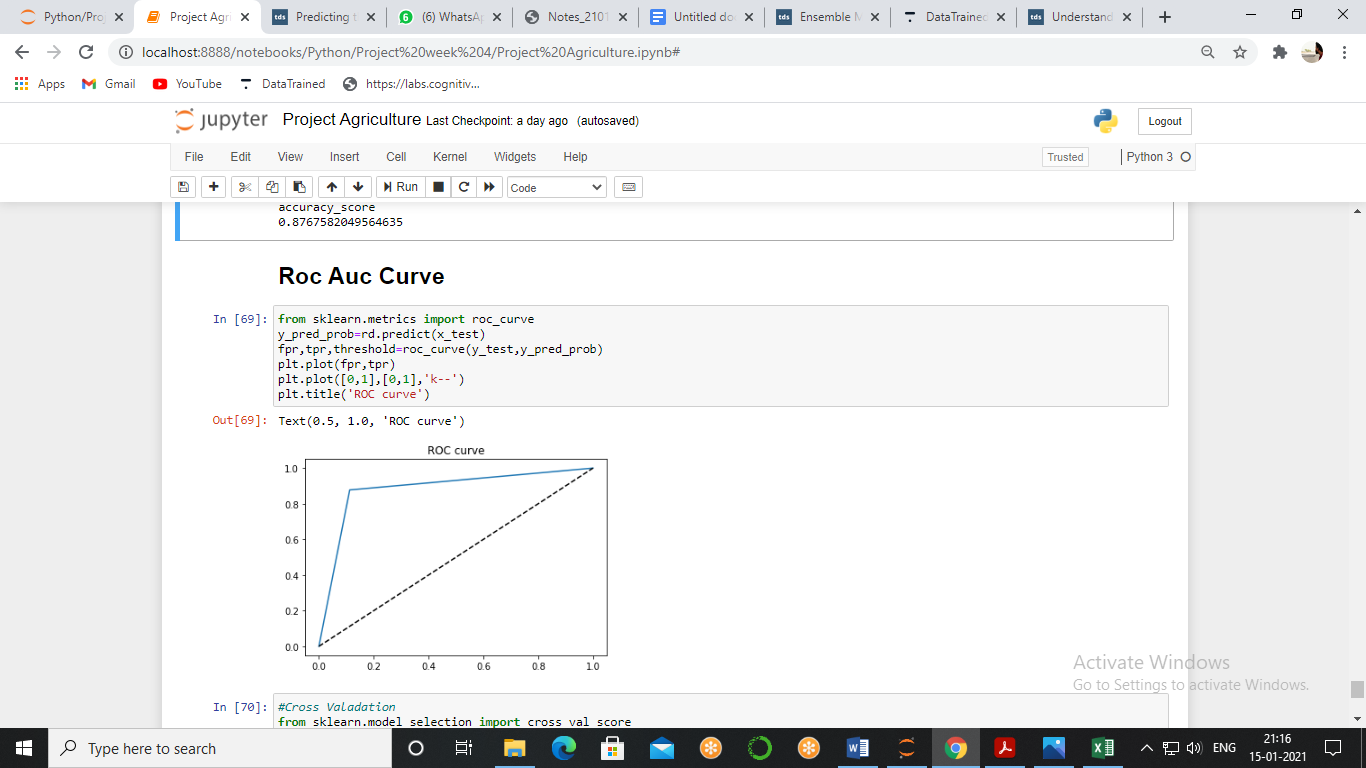
we can clearly see that this model is giving us the best score and confusion matrix as well, so, we will make our final model using **Random Forest Classifier** and the parameters that are given by grid search.

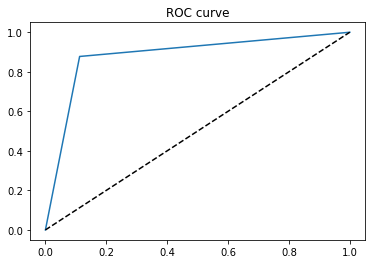


Our final model is **Random Forest Classifier** with N Estimator = 100, Criterion = Entropy and Random State = 60, then we got Accuracy Score = 87.67%, Auc\_Roc Score = 87.64%, Precision = 87%, Recall = 86%.

ROC Curve

It is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting.

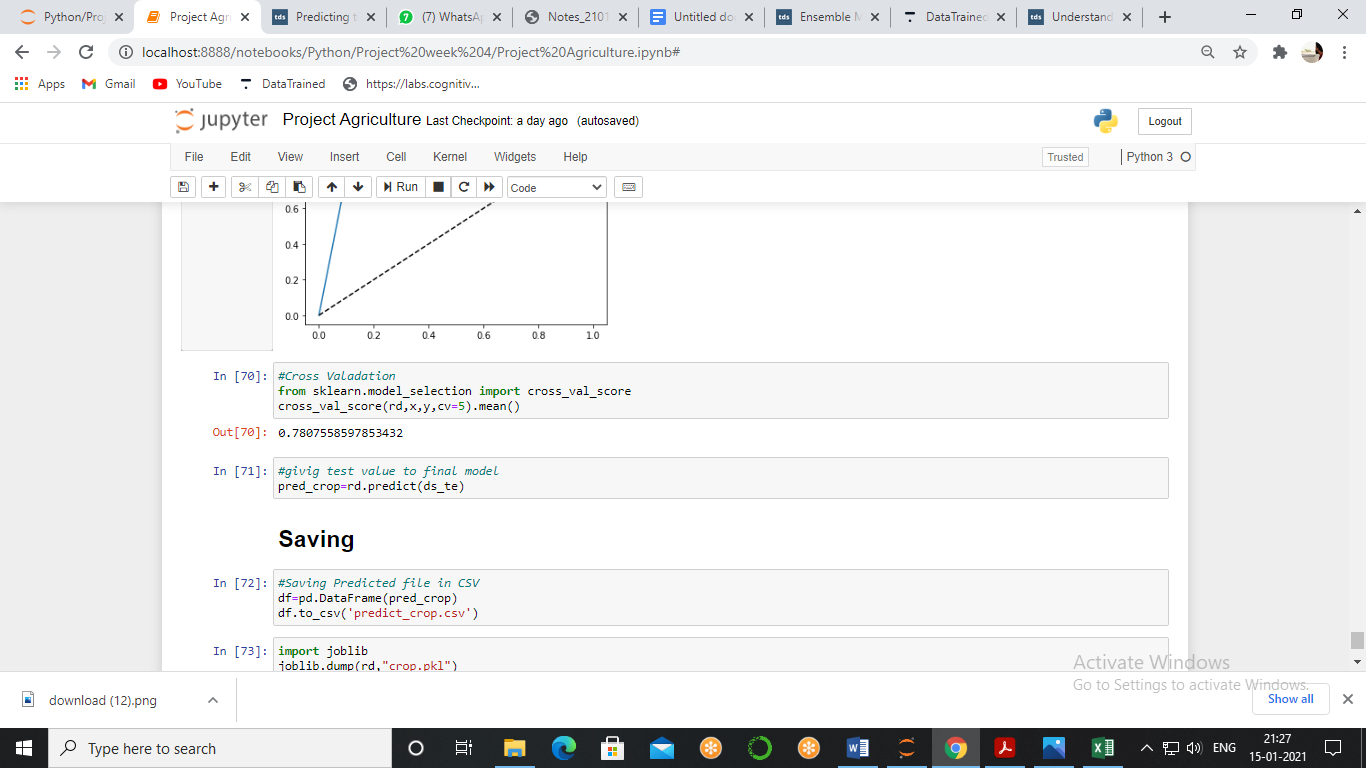




Red line in the middle represents a purely random classifier and therefore your classifier should be as far away from it as possible. Our Random Forest model seems to do a good job.

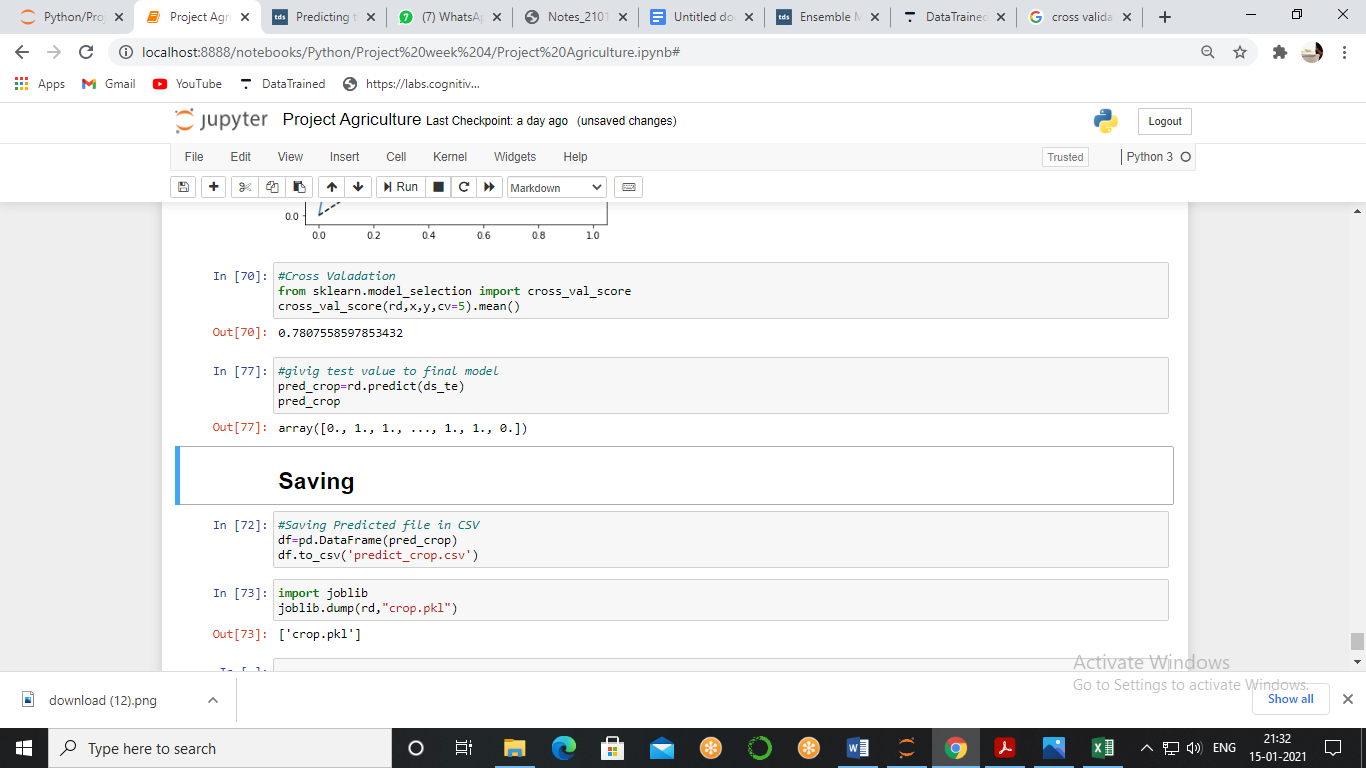
Overfitting

Now, our model is ready and we are getting a fair amount score but this model is fitted perfectly because we learned that when model learn everything it shows good score but when we introduce the new data then prediction all go wrong and to see that we see cross validation score.



We got a score of 78% which is still a good score so we can proceed with this model.

As we are done with model making still we left with one task to pass the test data and find the crop damage level.



We got our result and we got a good model with a good score.

Conclusion

Use of pesticides is very critical for crops environment as well as people hence it should be used to get the maximum benefit from the minimum use, with the use of data processing and visualisation we came to know when how and at what amount we should use a pesticide, as well as what are the factors which can damage the crop.

With the use of machine learning Technology, we made a model which can predict the damage on the crops so the farmer can find out the real cause of damage, as well as to protect himself from the future damage by entering the values of every decision he takes and can get the result beforehand, this thing will lead us to sustainable farming, increase the production which will help farmer to get better crops as well as they can support the economy of nation.