DATA ANALYSIS OF AGRICULTURE CROPS

Nishank Shah

In Modern age smart farming concept are widely used which makes agriculture more efficient and effective with the help of machine learning. Machine learning is the scientific field in which machines are prepared to learn and made with strictly programmed. It creates new opportunities to farmers to quantify and understand data intensive process in agriculture operational environment.

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

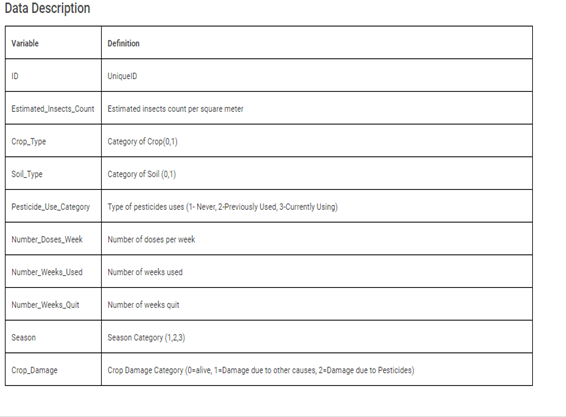
Farmer’s job is very hard and needs lots of determination and focus. Once the seeds are sown, he works days and nights to make sure that he cultivates a good harvest at the end of season. Harvesting depends on many factors to cultivate good quality of crops. Factors such as pesticide, water, soil fertility, preventions of insects etc. Weekly dose of pesticides should also be tracked because overdose can also spoil the harvesting.

I can across the Dataset which is completely based on crop harvested by various farmers at the end of the harvest season. Let’s get started.

**UNDERSTANDING THE DATA**

It is really important to understand the data first the data contains following columns. We will be using Python and environment is jupyter notebook you can download python by [Clicking here](https://www.anaconda.com/products/individual).

In this dataset we will be predicting whether the crop is damaged or not.



**IMPORTING DATA AND REQUIRED LIBRARIES.**

The first thing we will be doing is importing libraries these are initial libraries which is important to import data. We will be importing more libraries later. The libraries I will be importing is Pandas , Numpy, Mathplotlib, seaborn this are all inbuilt libraries in python which we will be using to import data and visualization of data.

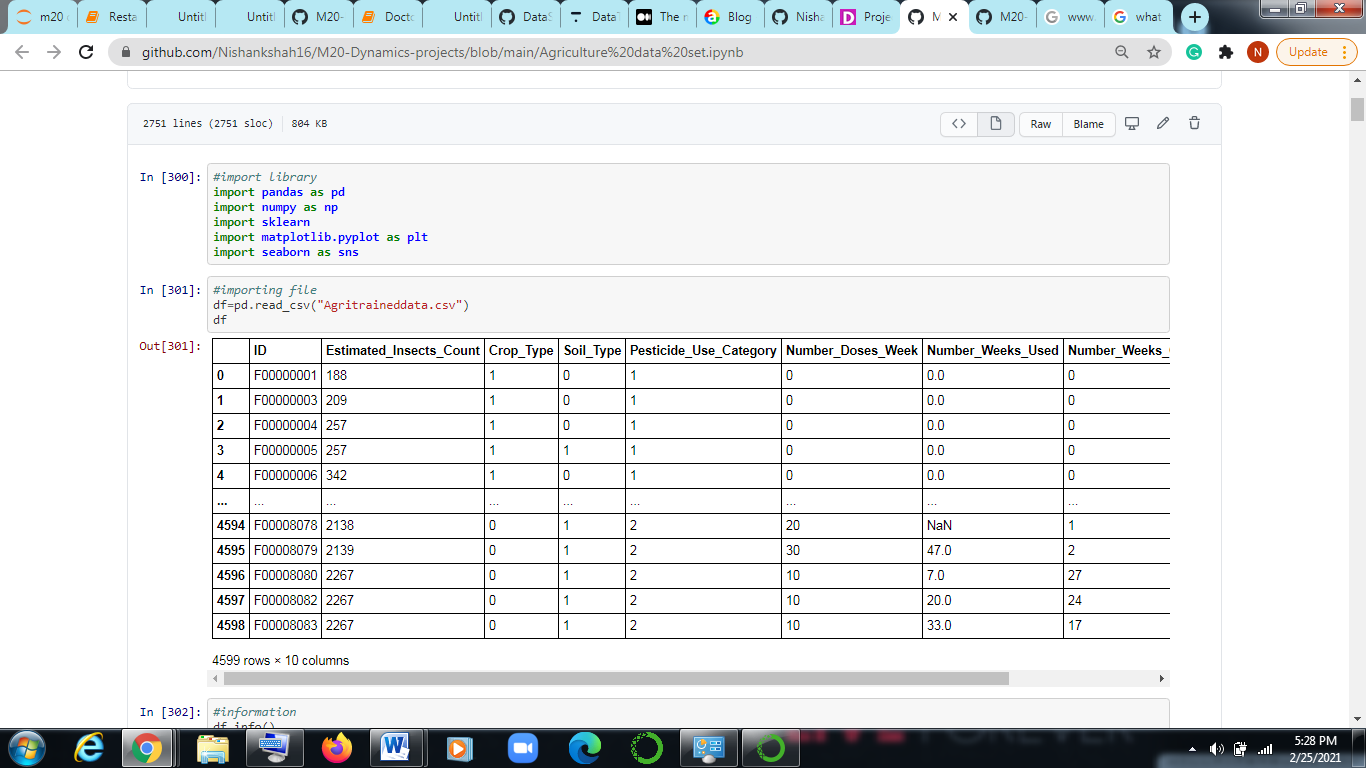
Lets see what this library do

**Pandas** is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language

**NumPy** is a Python library that provides a simple yet powerful data structure: the n-dimensional array.

**Matplotlib** is a comprehensive library for creating static, animated, and interactive visualizations in Python.

**Seaborn** is library that provides various charts and visualizations of data.



As above we can see that we have imported all required libraries by using **import** function.

We imported the file by the command **pd.read\_csv(“Agritraineddata.csv”).** Which has created a nice dataframe called by **df**. df is nothing but we have stored our data in df variable.

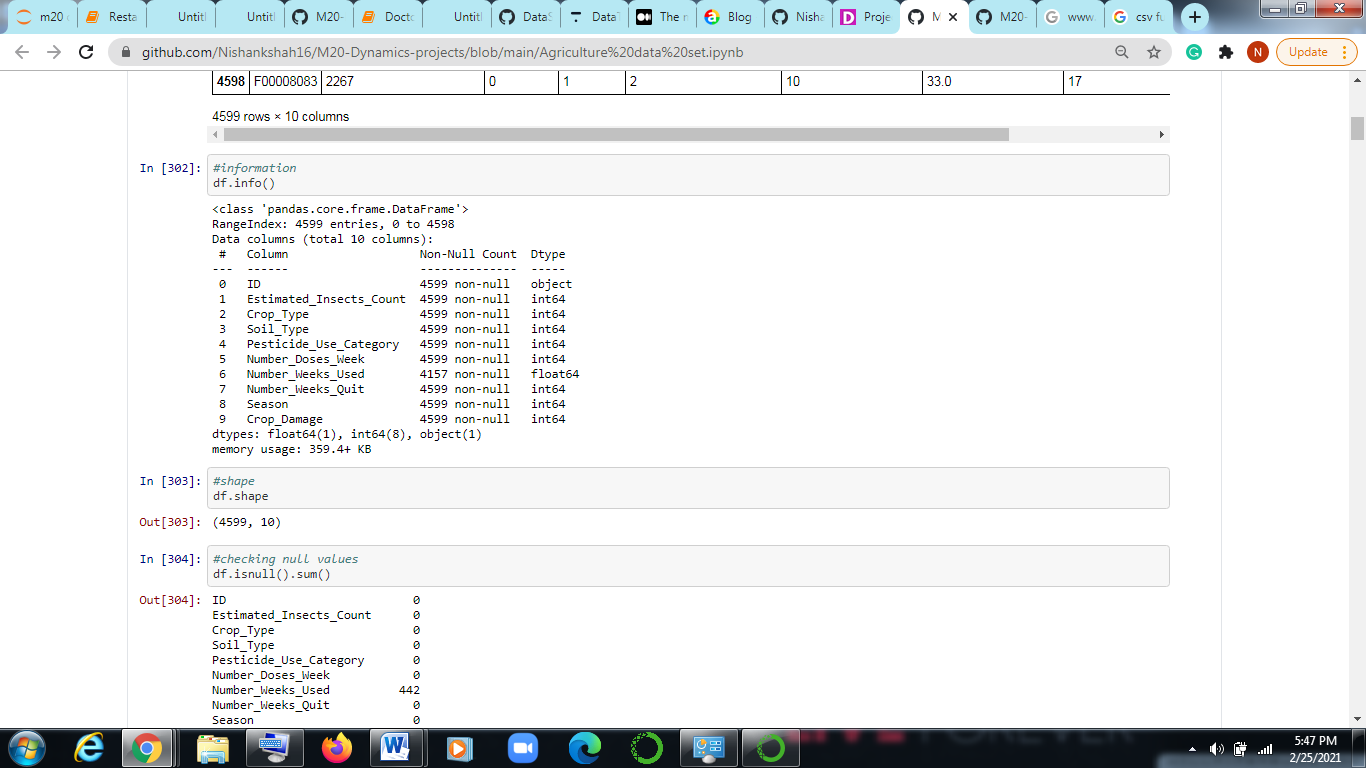
Here, csv is comma separated values.

Pd is nothing but shortform of pandas as we mention **import pandas as pd**.

**Checking Information and shape of the data:**

Now I checked the info of the dataset and function used is **df.info()** in this we will get all details of the data like what are the data types , size of the data, and names of columns etc.

Shape is size of the data which is checked by simple function i.e **df.shape** no brackets are used at the end.

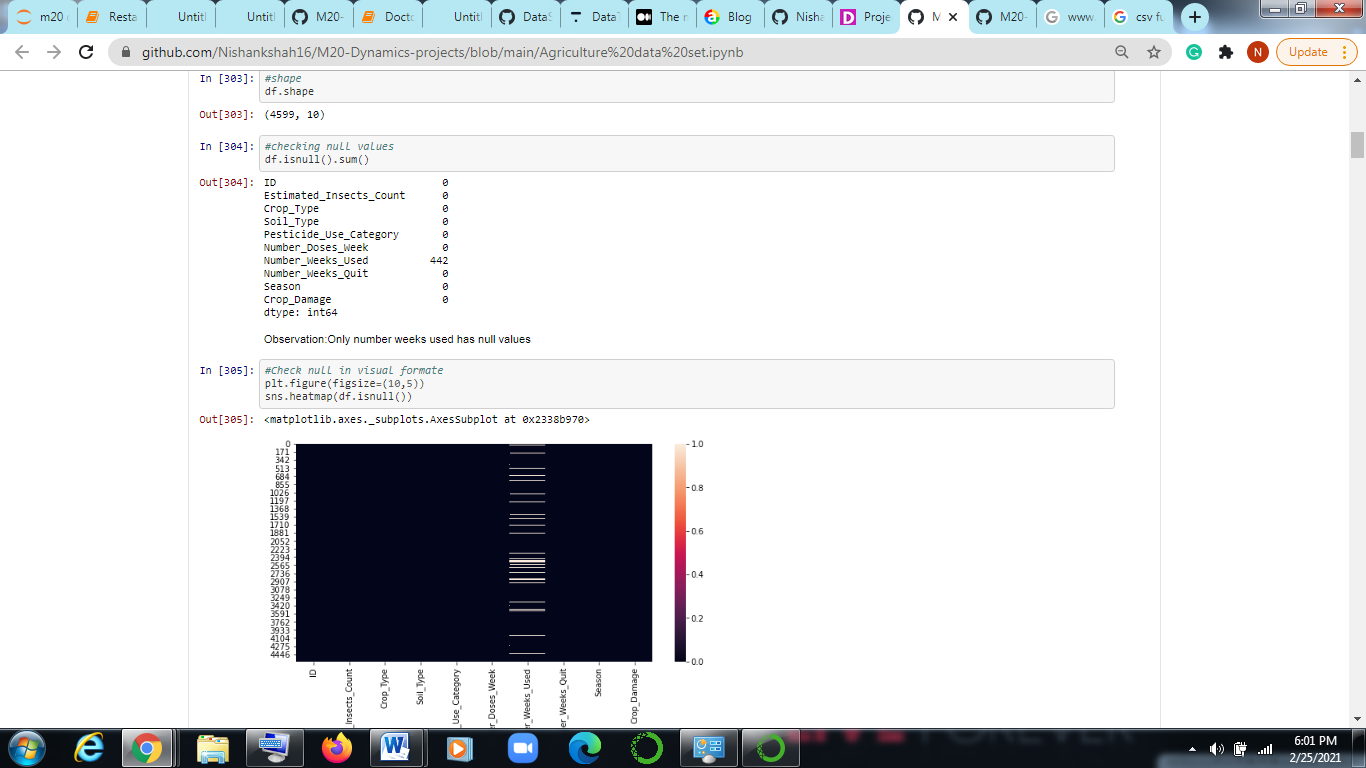


AS we can see all the information in the data set and in shape there are 4599 rows and 10 columns.

**Checking null and visualization:**

Null values are values which are missing in the given data set the can be seen as **NaN** . We can check null by using simple inbuilt function i.e **df.isnull.sum().** This means check is there any null value present in the dataset and to get the sum of the all the values extend the code with sum().

Similarly we can see null values with the help of visualization but here we will be using seaborn(sns) library which we have already imported initially. Let us check how we will be executing the code.



As we can see only number Weeks used has 442 null values. Similarly we can see in heat map the white horizontal bar represents null values in number of weeks**. Sns.heatmap(df.isnull())**. Where ‘plt’ is matplotlib.pyplot which is represented as plt. **Plt.figure(figsize=(10,5))** we can adjust the size of the figure.

**Handling the null values:**

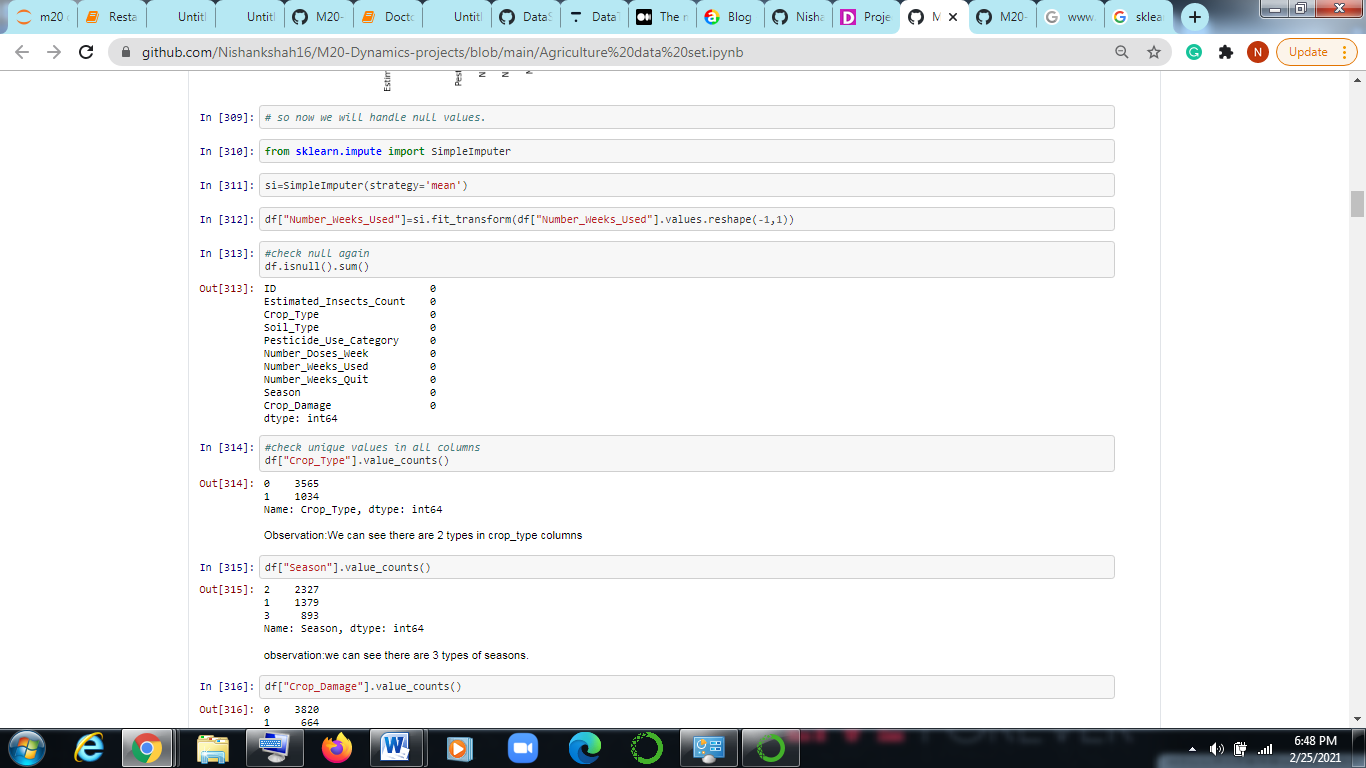
Null values should either be eliminated or should be fill with some value in our case let us see what we can do.

The only null column is Number\_weeks\_used there are two methods where we can use null values either we can use simple imputer of we can use mean of the rest of the present data.

Always remember if data is categorical or non-numeric always take mode of the present data. If the data is numeric then always use mean or median of the data.

In our case there is numeric data and data type is float. So we will be using Simple imputer which is inbuilt function in sklearn. Sklearn is Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, etc.

So from sklearn.Impute we will import simple imputer. In this function we have to set the strategy which means we can want data mean or median.



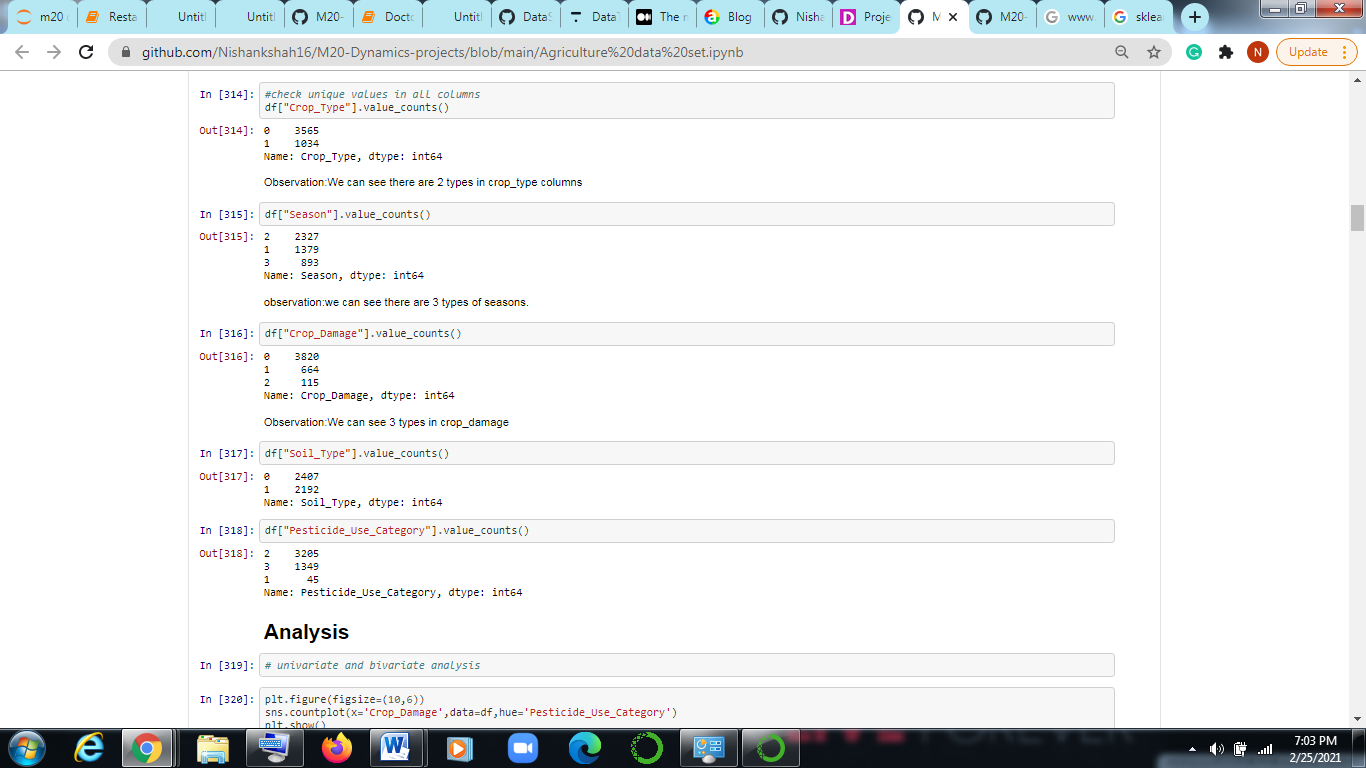
As we can see first we imported simple imputer from sklearn.impute then we assigned the strategy as mean because values are number or integers in simpleImputer(), now we will write a code to fill null values which is **df[“ ” ]=si.fit\_transform(df[“ “].values.reshape(-1,1)** this nothing bit we are fitting simple imputer and transforming the given null columns and filling the values.

After that we will check again for null values are here is the result we have successfully handled the null values without missing any data, Its always better to fill values rather than droping the values as we don’t lose any data.

**Checking data and uniqueness:**

In this we will be checking uniqueness in the columns with the intention of making categorial value in numeric. This will give more info about the data in columns. It is not necessary for all Columns but doing this good practice.

The function we will be using is **df[“ ”].values.counts()** this command will show you uniqueness entities present in the columns.



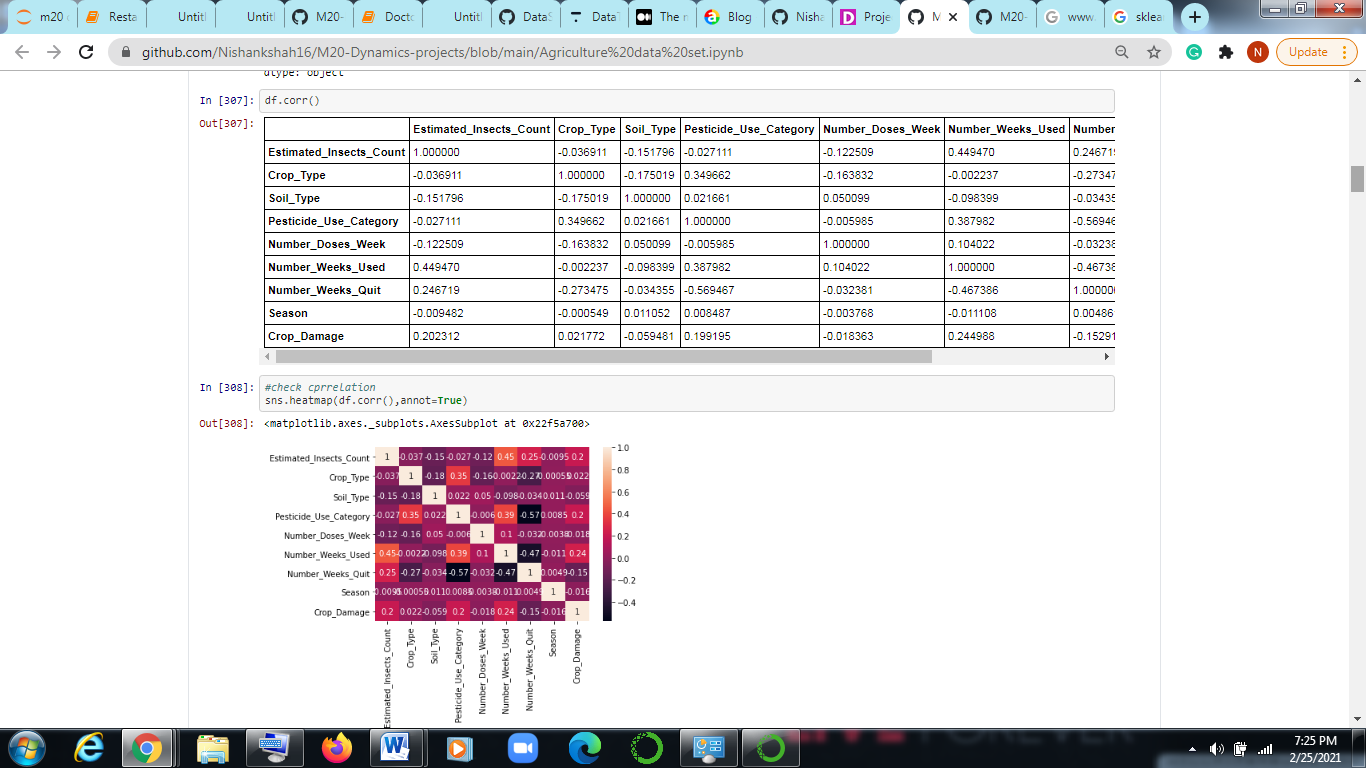
We can see in above image the unique value for crop type is 0 , 1 in which 0 counted 3565 and 1 is counted 1034 and also data type is float.

Similarly other columns such as season, crop Damage, Soil type etc can be checked.

Now far we have seen how we can handle data and how can we handle null values and finding the unique values , there is one more thing we have to do and that is finding the correlation.

**Finding the correlation and Describe method:**

Now in python there is inbuilt function which is call corr( ) which means correlation we can find the correlation by running this code df.corr() and similarly we can visualize it using heat map. Which we use already.



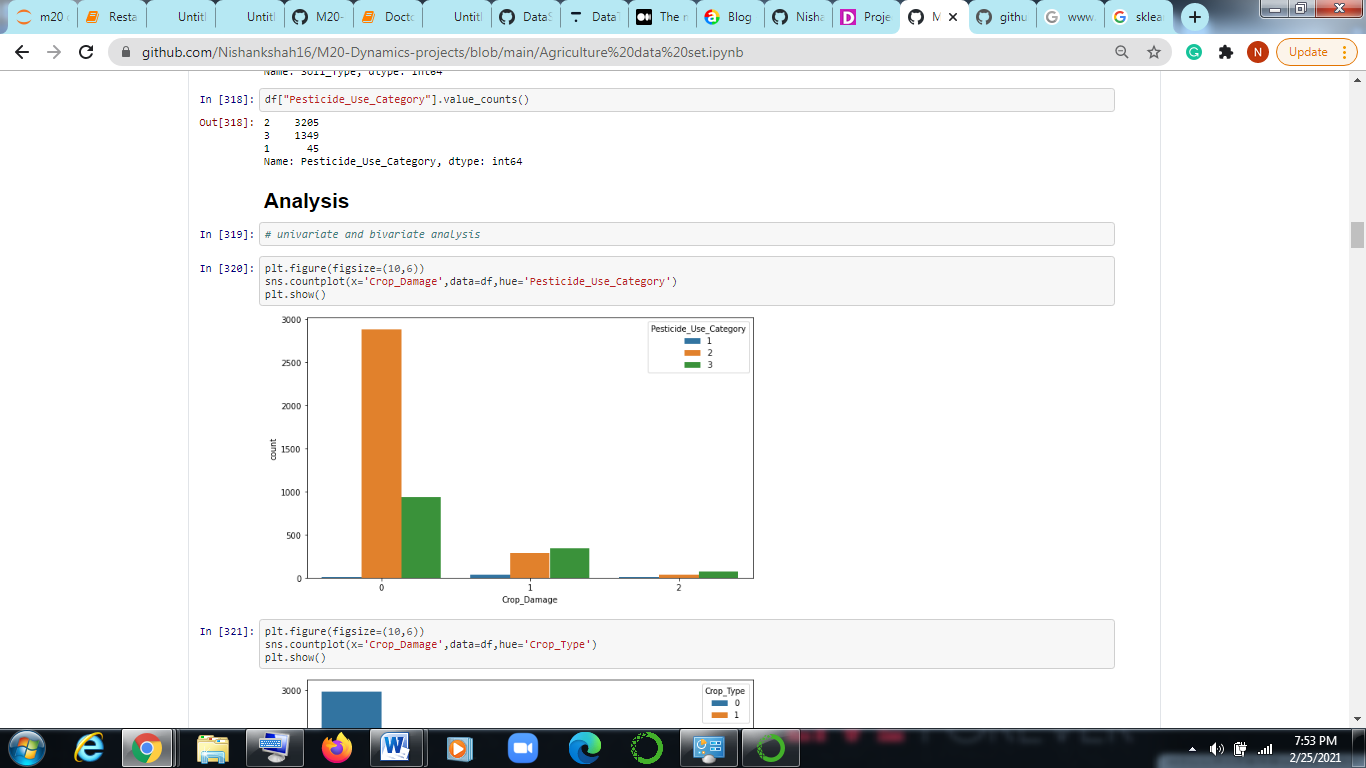
As we can see df.corr() and and we can see heatmap the lighter color represent they are highly correlated and dark part represents the values are negatively correlated.

annot=True is nothing but displaying values on the heat map.

**DATA ANALYSIS :**

Now it is very important to analyze the data. It depends on one’s analytical skill that how do they observe the data. Analyses can be made in univariate and bivariate analyses.

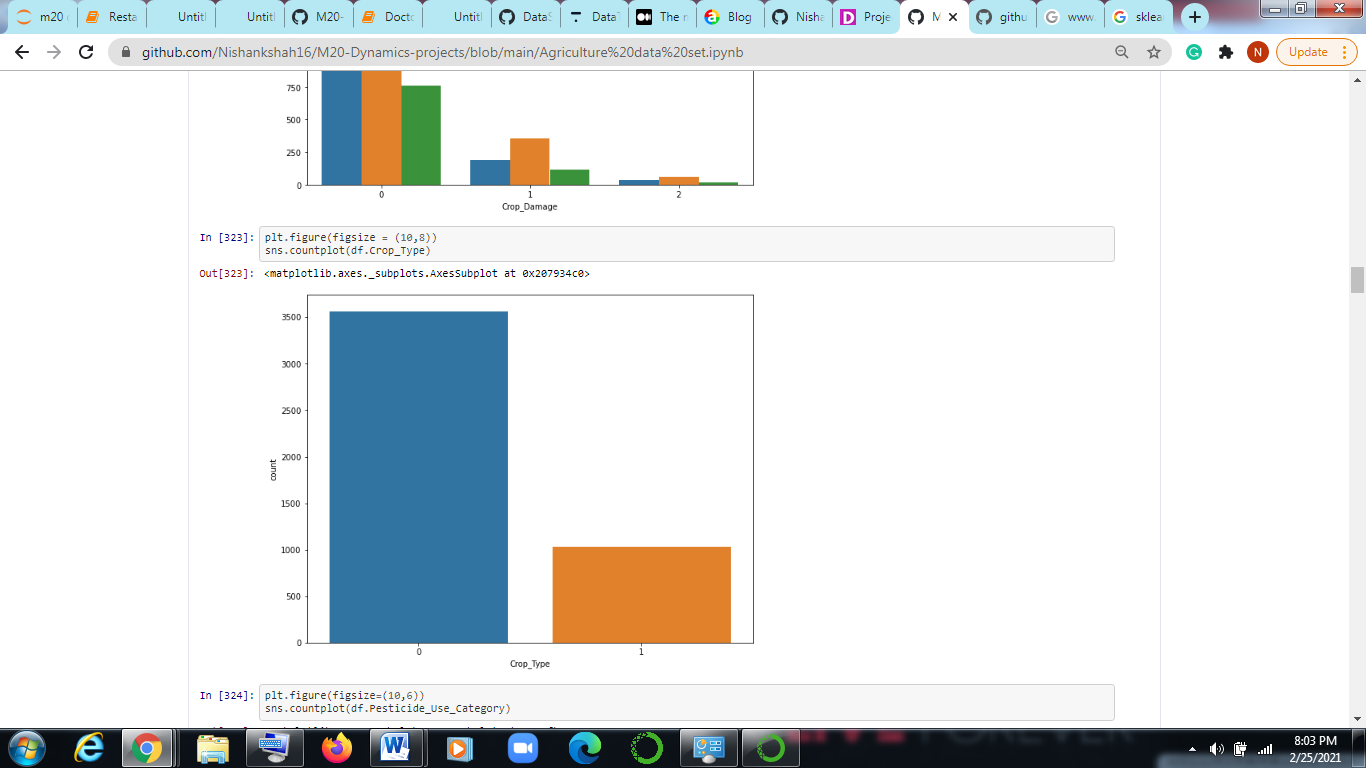
Univalent means analysis done on single attribute and bivalent is analysis done by two attributes.



Here we can see we used Count plot which is basic visual method we use bivariate analysis as there are two attributes Crop damage and pesticide\_use\_category.

Our aim in this analysis is to see which crop is damaged more by using which pesticides.

So we can see in the graph that most damage is occurred by using 2 type of pesticide and 0 type of crop is more damaged.



This type of analysis is univarient analysis here we have used only one attribute i.e is crop type.

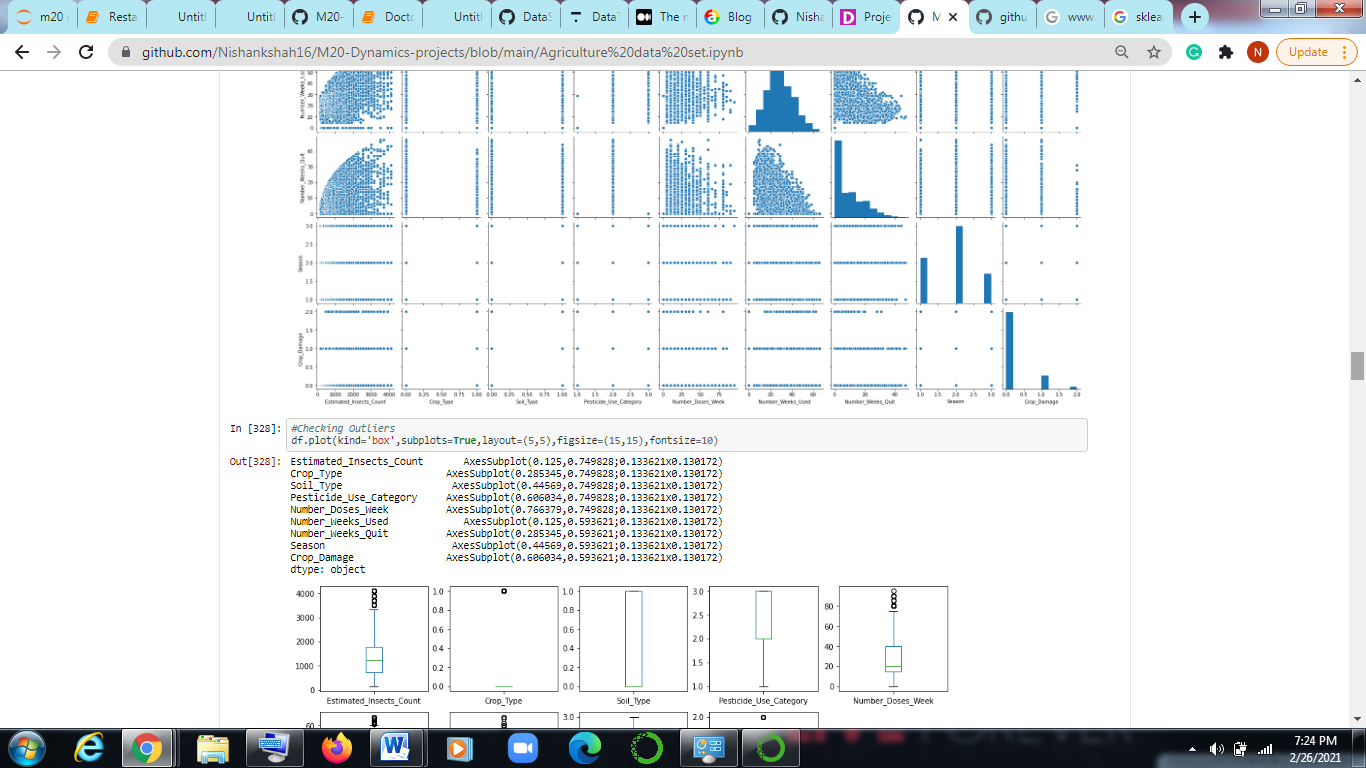
Similarly we can do rest of the analysis.





There is also a function called pairplot which is shows all the relation between all the attributes and predict the observation.The code is written as **sns.pairplot(df).**





As we can see all graphical relationship between the attributes.

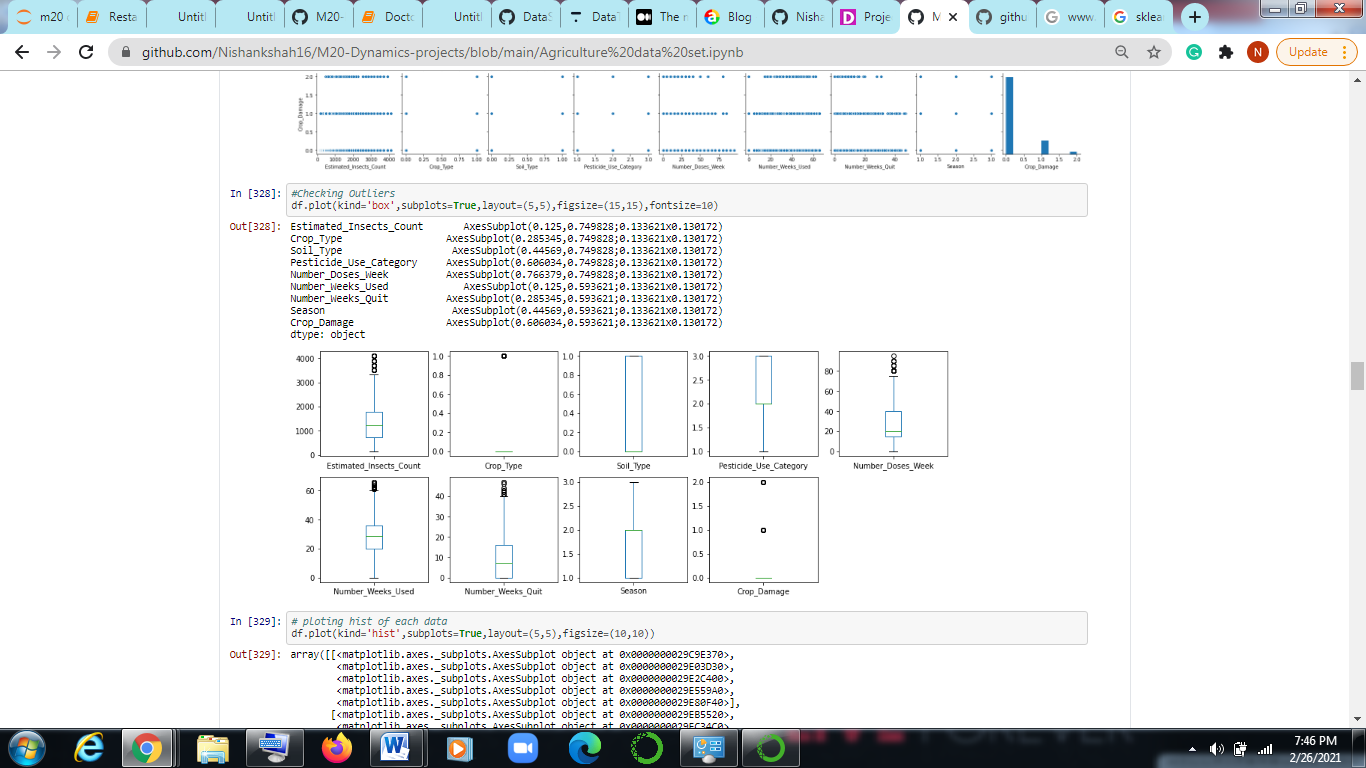
**Checking for outliers:**

Now we will check for outliers ,outliers are nothing but values which are exception or values which varies more rather than other values present which may affect your mean and result.

So it is important check the outliers.

To see the outliers we will use boxplot where we will be seeing all the outliers in one frame. Remember outliers can only be plot for numerical values we can’t plot outliers for string values.

We will be using matplotlib which we have denoted as plt, so code we will be writing is **df.plot(kind=’box’, subplots=True, layout=(10,10), figsize=(10,10))**.

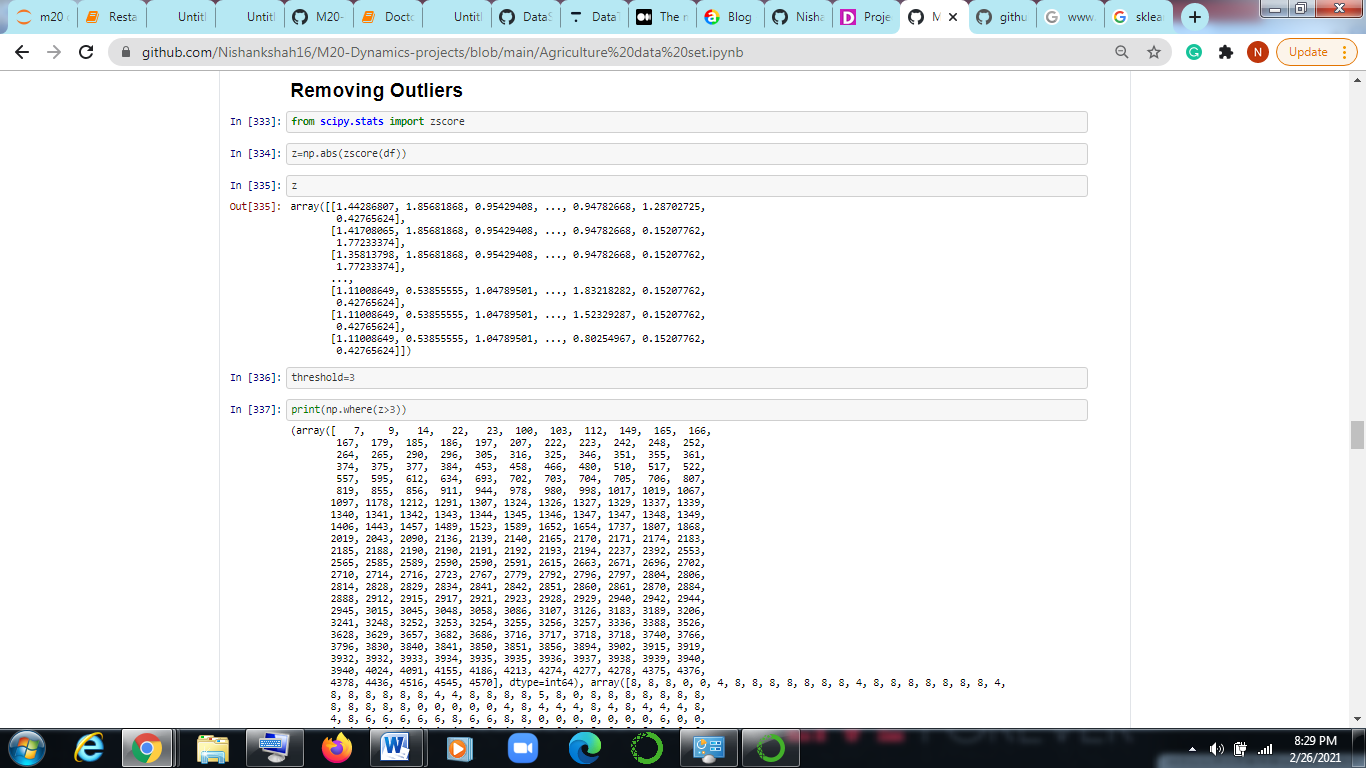


So we can see all the box plot in one frame. The circle which are above are outliers. There are few outliers present inEstimated insects count, number of weeks used , number of weeks quit , number of doses week.

**Removing the outliers:**

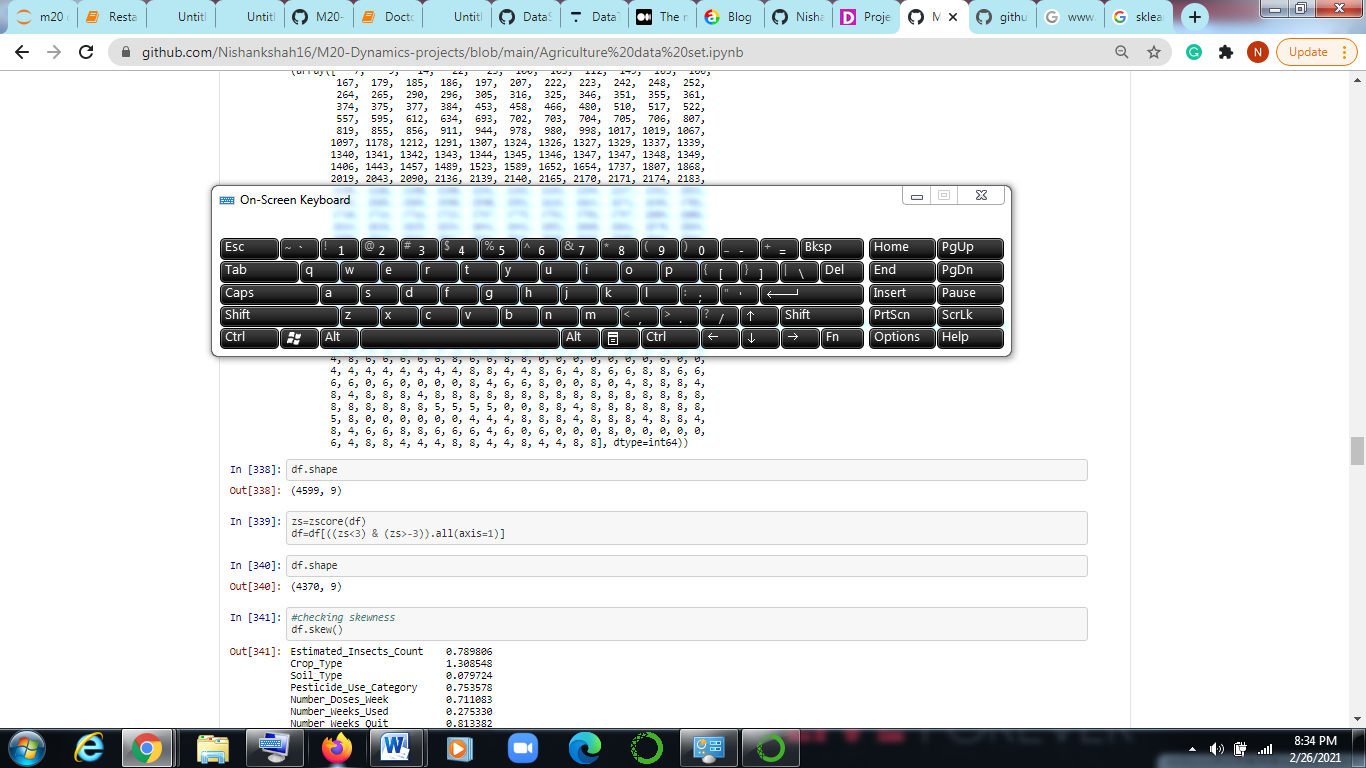
Now we will remove outliers by using z score , zscore is inbuilt function which is removes the outliers where we mention threshold it is imported from scipy.stats library.

So we have to import the library first , then we will write a code as follow.



Now we will check the shape before removing outliers.

For checking shape we write df.shape



As we can see the number of rows has get decreased to 4370. Thus we have removed the outliers.

**Removing skewness**



To find skweness we will use inbuilt function i.e is **df.skew().**

For removing skewness we will use log transformation as we can see in the figure above.

**Model Making:**

Now we will make a model.

First step is we have to import required libraries, libraries required are train test split, min max scaler, standard scaler logistic regression Decision tree classifier , svc K neighboursclassifier , GaussianNB.

So train test split is a model selection in which we define data for testing and data for training.

The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and **standard** deviation of 1.

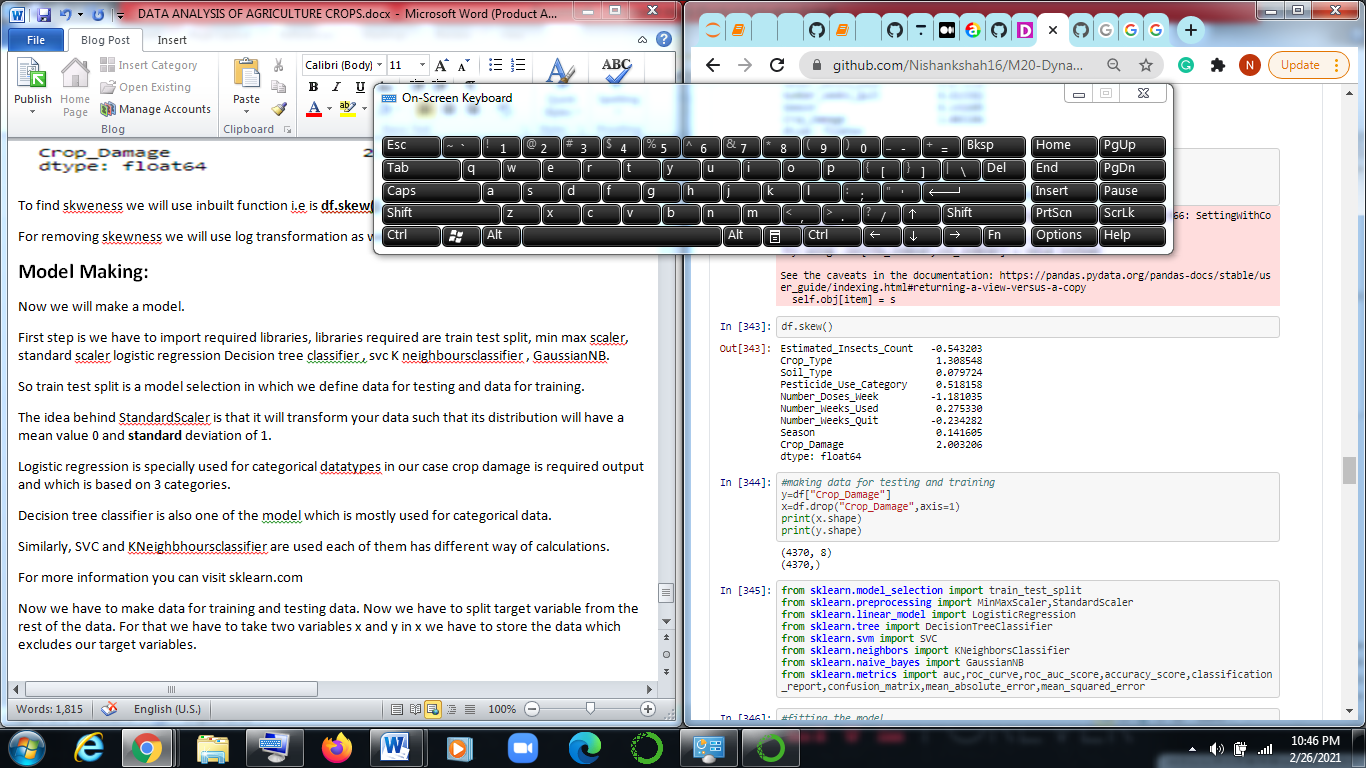
Logistic regression is specially used for categorical datatypes in our case crop damage is required output and which is based on 3 categories.

Decision tree classifier is also one of the model which is mostly used for categorical data.

Similarly, SVC and KNeighbhoursclassifier are used each of them has different way of calculations.

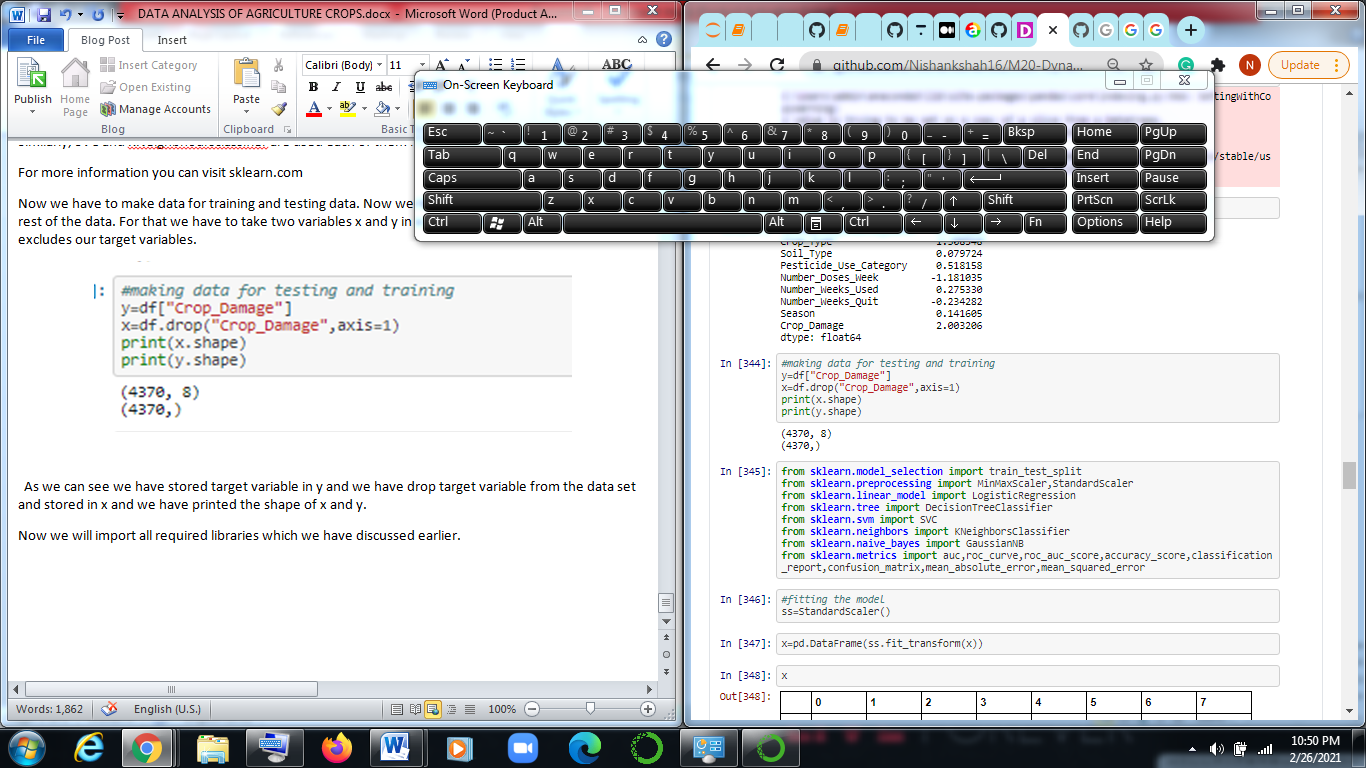
For more information you can visit sklearn.com

Now we have to make data for training and testing data. Now we have to split target variable from the rest of the data. For that we have to take two variables x and y in x we have to store the data which excludes our target variables.



As we can see we have stored target variable in y and we have drop target variable from the data set and stored in x and we have printed the shape of x and y.

Now we will import all required libraries which we have discussed earlier.

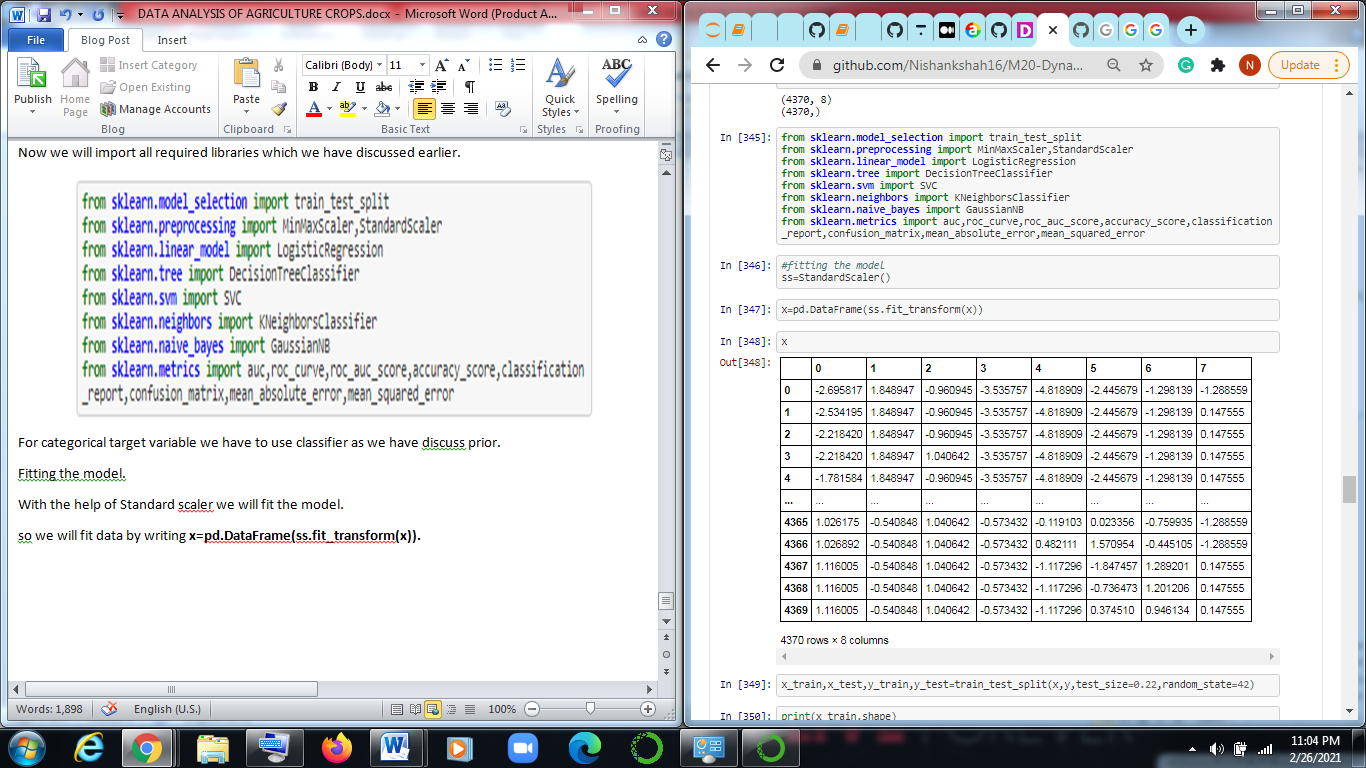


For categorical target variable we have to use classifier as we have discuss prior.

Fitting the model.

With the help of Standard scaler we will fit the model.

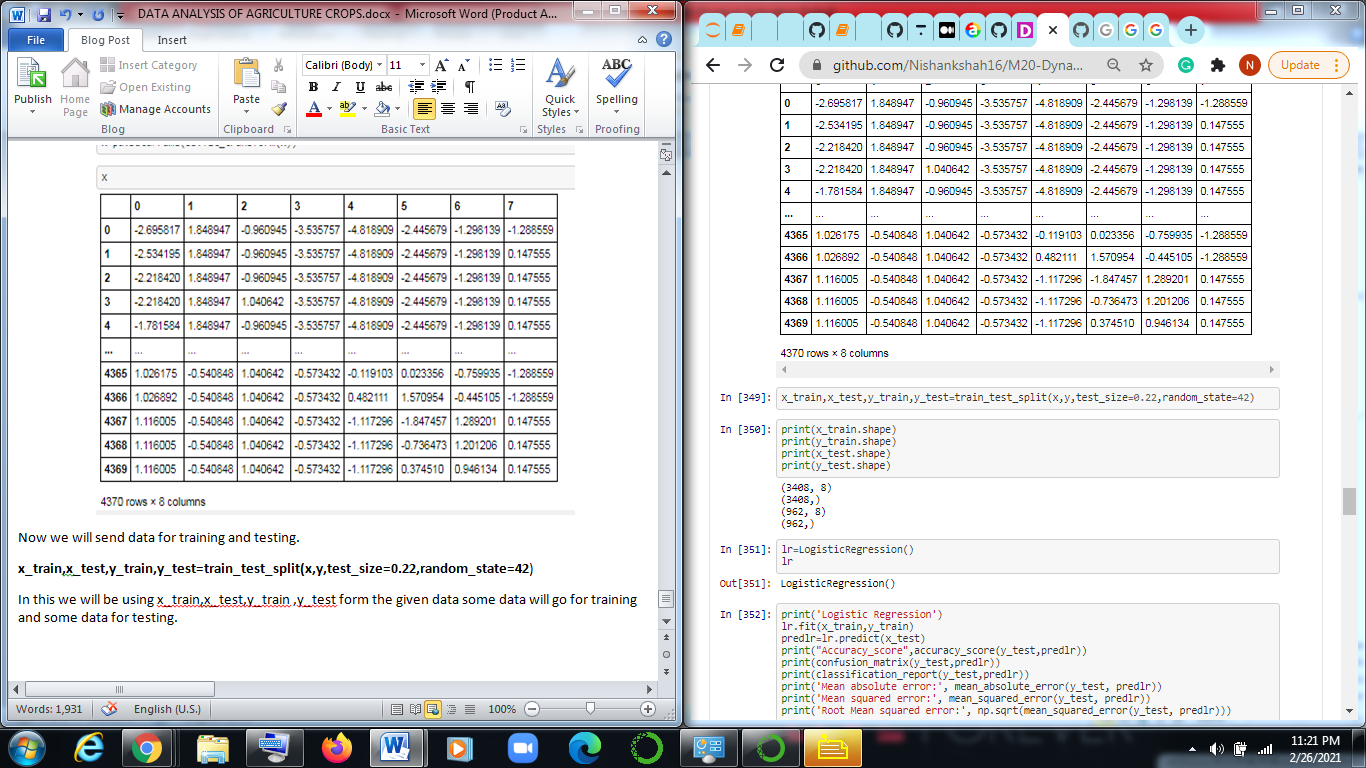
so we will fit data by writing **x=pd.DataFrame(ss.fit\_transform(x)).**



Now we will send data for training and testing.

**x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.22,random\_state=42**)

In this we will be using x\_train,x\_test,y\_train ,y\_test form the given data some data will go for training and some data for testing.



Now we make use of various models like Logistic regression and Decision tree classifier.

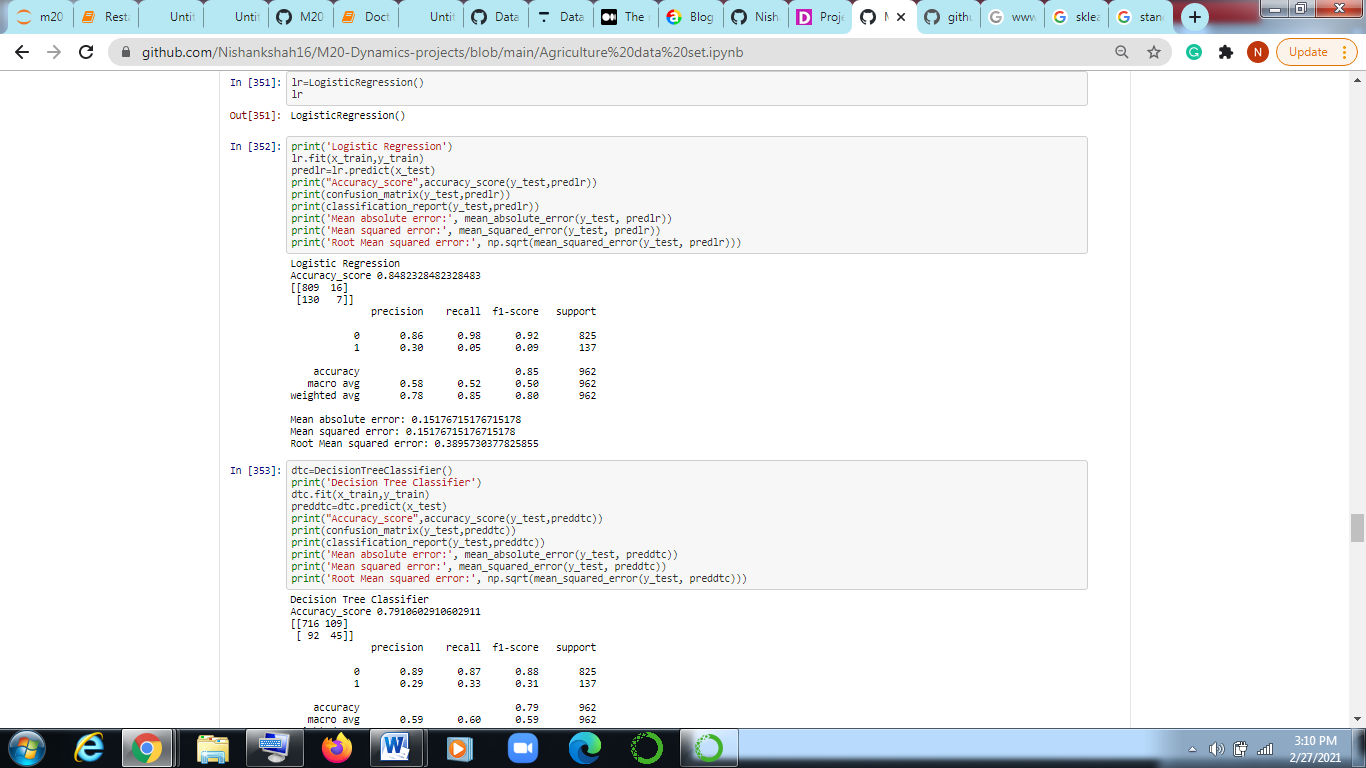
We will call logistic regression function and store it in lr.

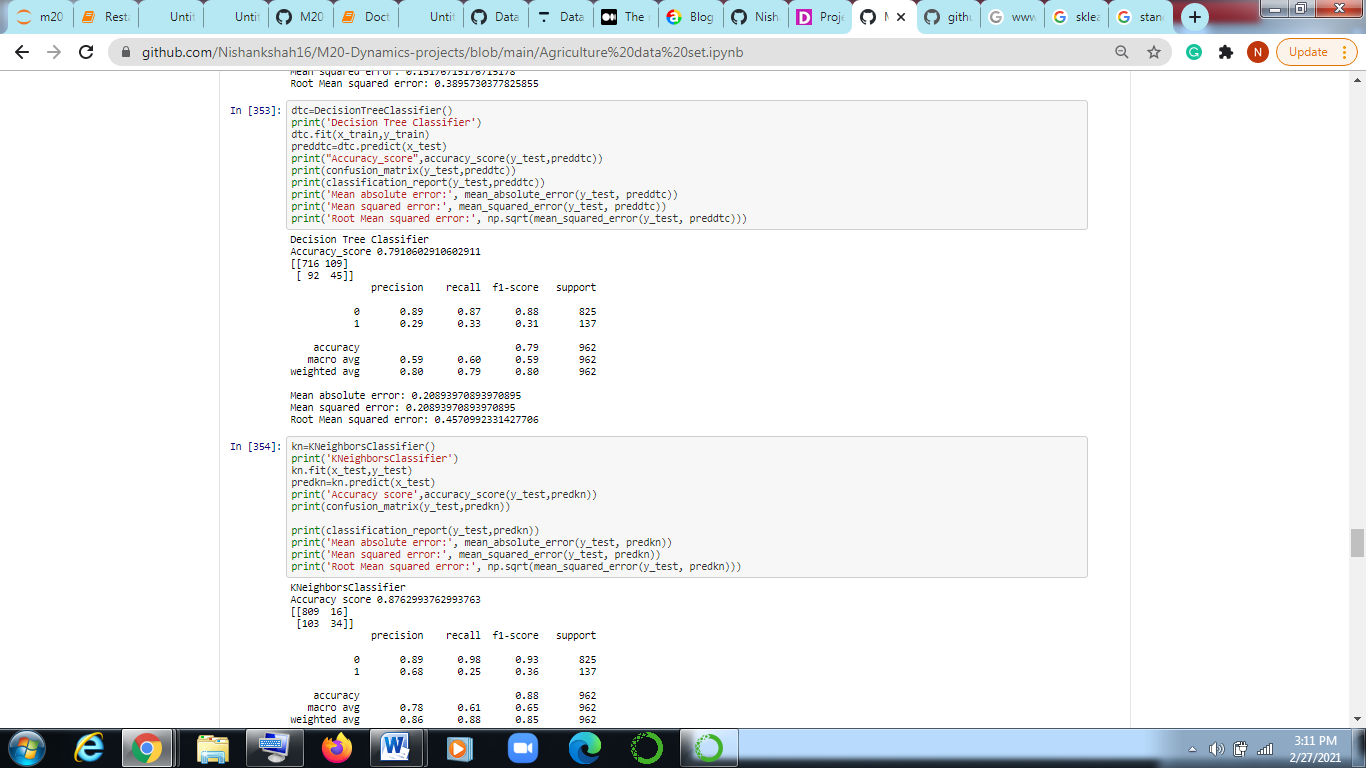
Now we will fit x\_train, y\_train data in logistic regression model.

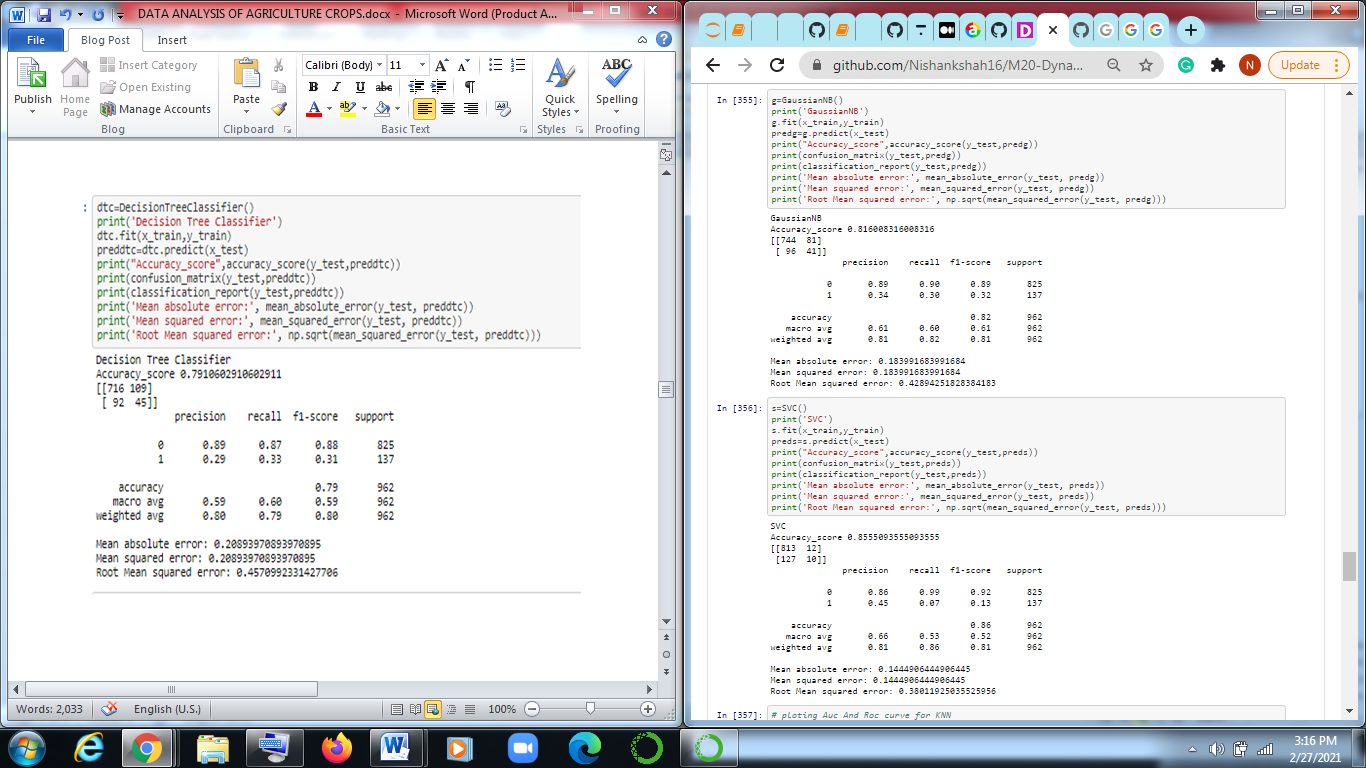
We will define predlr model and predict x\_test

After that we will find the accuracy score, classification report, Classification Report, mean absolute error, mean squared error, Root mean squared error.

We will be finding this for each of the models and we will get accuracy score and by comparing all the models, and we will be choosing the best models which will be performing better.







Now we will select the best model according to their accuracy score and check which model is performing well. As in above figure will now save the model for that we have to import joblib in which we can dump and load the model.

So we have chosen Support vector Classifier model as it has highest accuracy score. Now we will be saving our model in obj file.

Thus now we can put any number of data and predict the value for given dataset.

For more information you can visit datatrained.com which may provide best courses on machine learning.