****

**INTRODUCTION:**

This blog post is about building a machine learning model which predicts whether the crop will be alive and healthy or whether the crop will be damaged. If the crop is damaged it might be due to 2 reasons. One being improper usage of pesticides and second might be due to other reasons. The prediction is done by considering other aspects like type of crop, type of soil, pesticide usage, dosage of pesticides used per week, season.

**PROBLEM DEFINITION:**

A farmer's job is 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 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.

**DATA ANALYSIS:**

There are two different datasets each for training and testing the model.

Data cleaning and preprocessing will be done on both the training and testing datasets. The model will be trained on training dataset and predictions will be made on testing dataset.

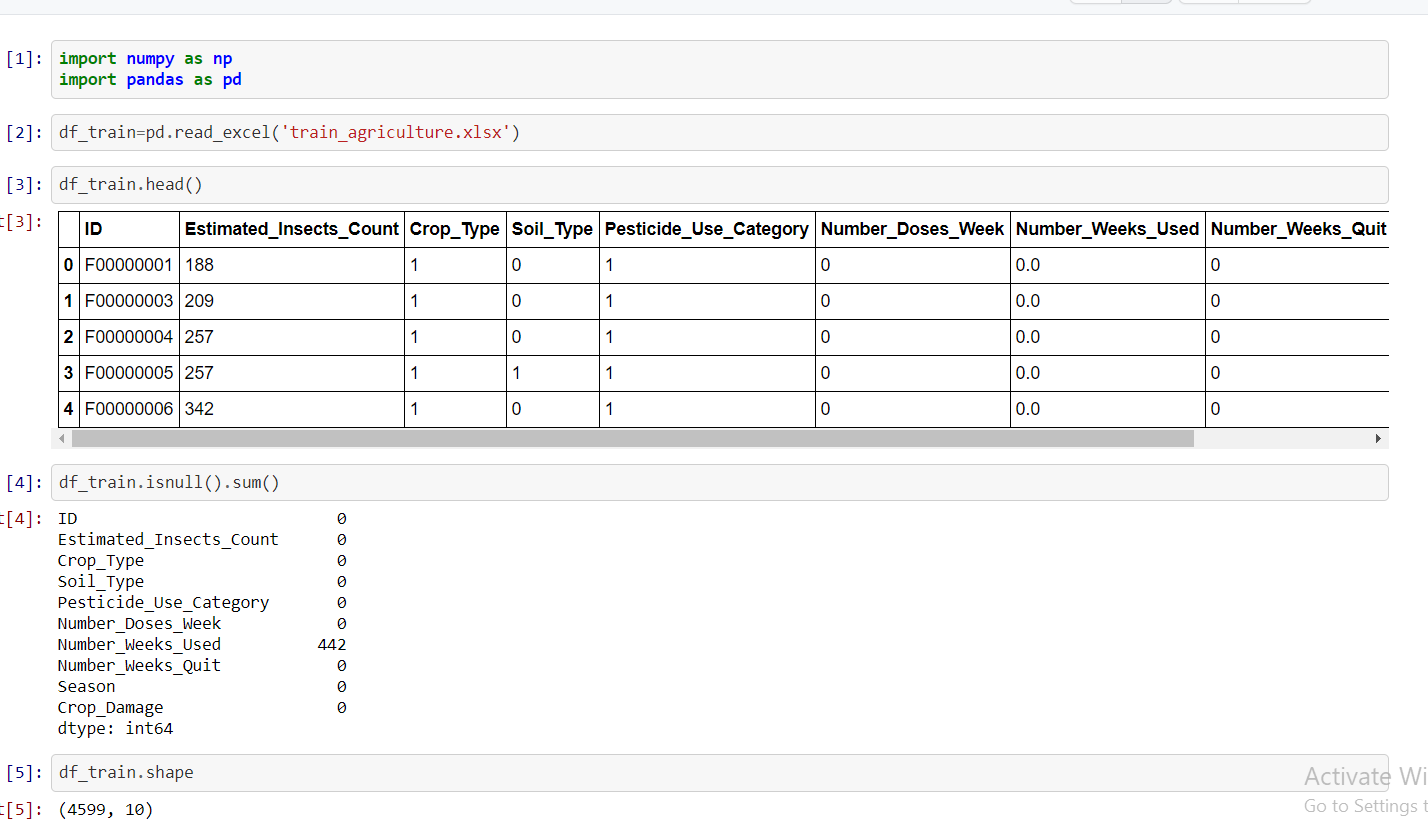
After importing the necessary libraries, loaded the data set.

Checking for the following

1) The first 5 rows of the dataset

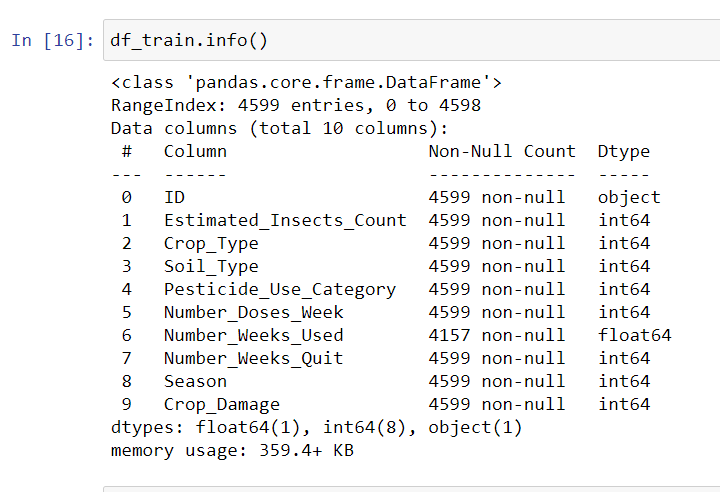
2) Verifying whether the dataset has any nulls.

3) Checking the size (i.e no. of rows and columns)

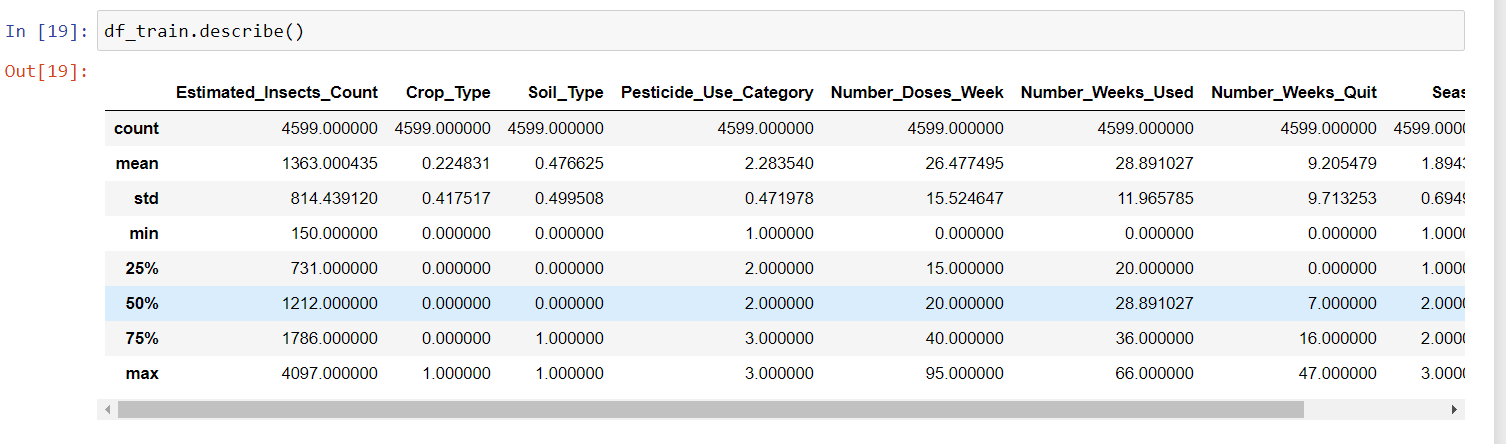


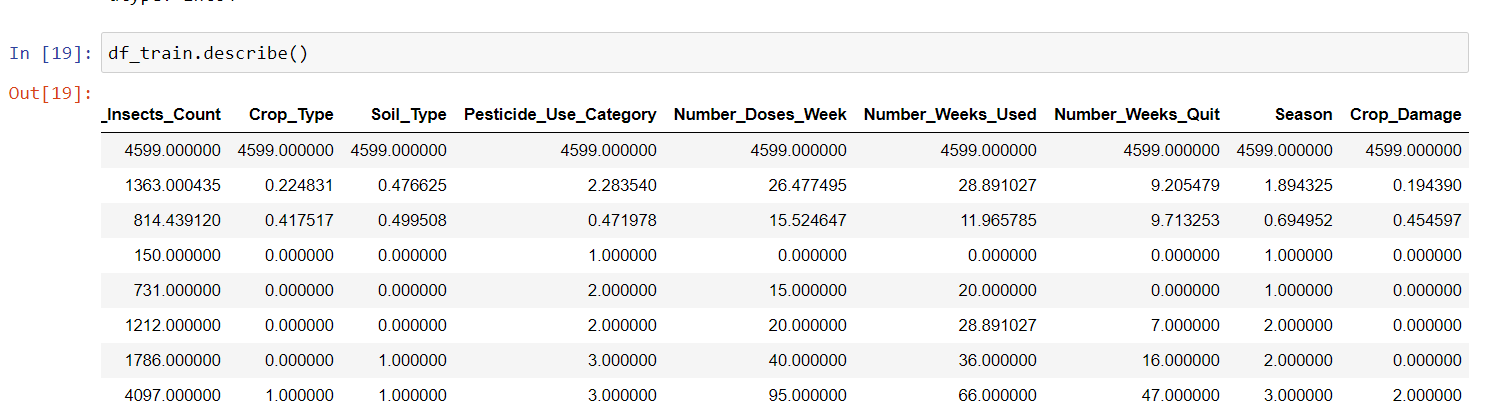
We can notice that Number\_weeks\_used column has null values; these null values will be dealt later.

The training dataset has 4599 rows of data corresponding to 10 columns. Out of these 10 columns, crop\_damage column is the target variable, rest are independent variables.



We can observe the data types of each column and their corresponding row count.



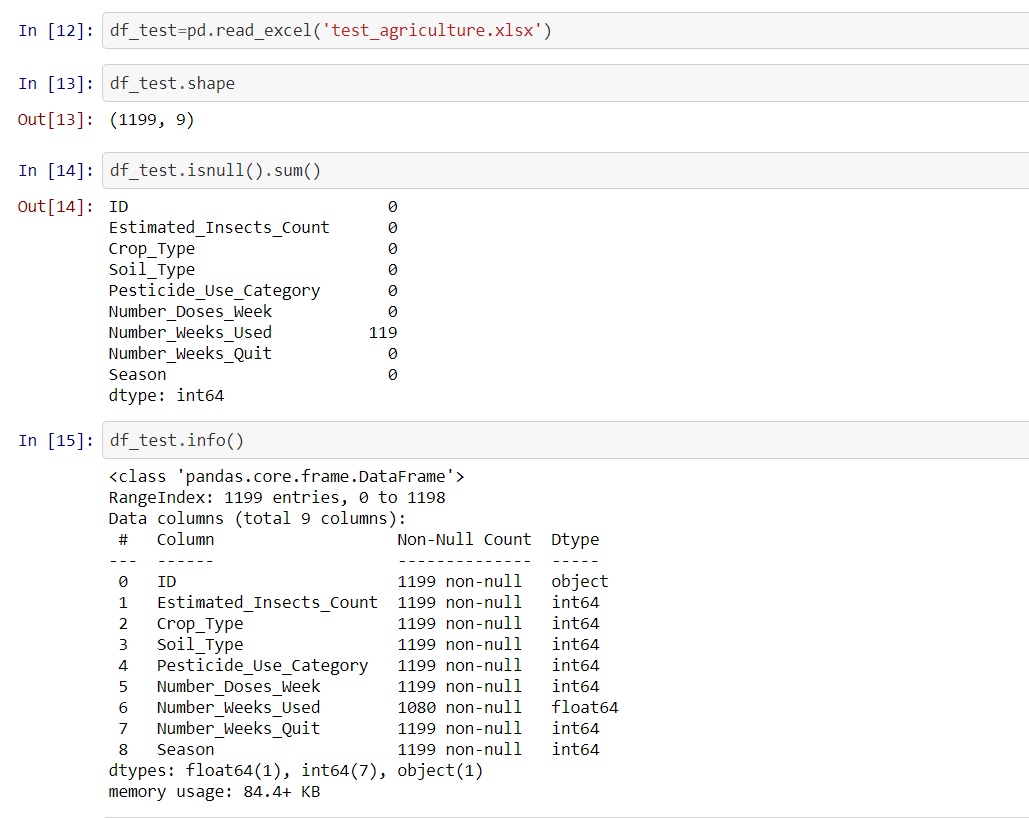


Crop\_Damage, Season, Pesticide\_Use\_Category, Soil\_Type, Crop\_Type have cateogerical values(i.e 0,1,2,3)

No.of doses of pesticides per week has minimum of 0 doses and maximum of 95 doses per week.

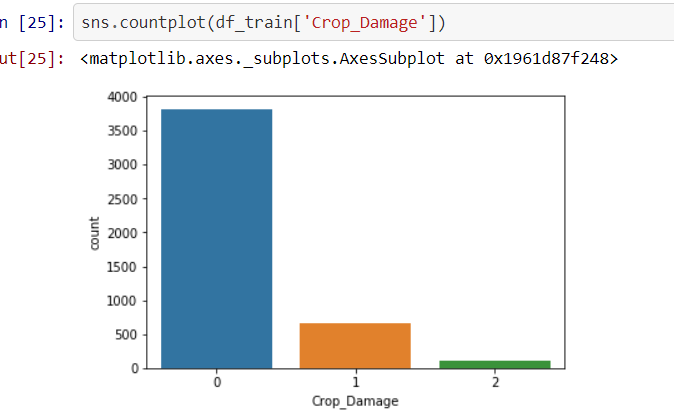
No.of weeks of pesticide usage varies from 0 to 66.

No.of weeks of pesticide quitting varies from 0 to 47 weeks.



The test dataset has 1199 rows of data corresponding to 9 columns. For these 9 columns the model will predict crop damage cateogery.

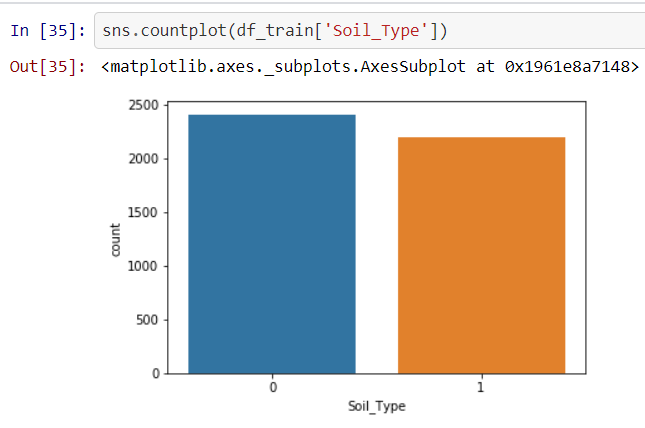
1. CROP DAMAGE



Crop damage of type 0 means the crop is alive,we can observe most of the crop is alive.

Crop damage due to pesticide(type 2) is less compared to the damage created by other reasons(type1)

2)SOIL TYPE



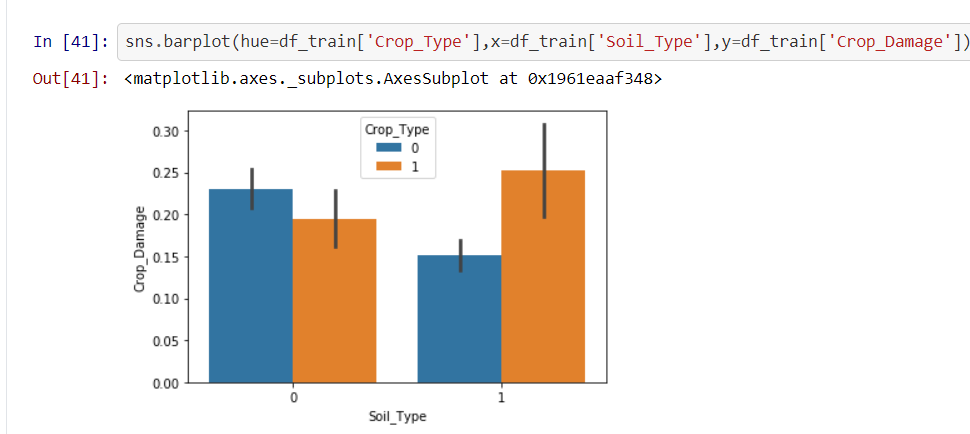
The collected data has relatively high information about soil type:0 when compared to soil type 1

3)Damage due to Insects:



There are two crop types [0 & 1], in both crop type’s likelihood of living crop is low with increased insect count.

4)Damage in different soils:

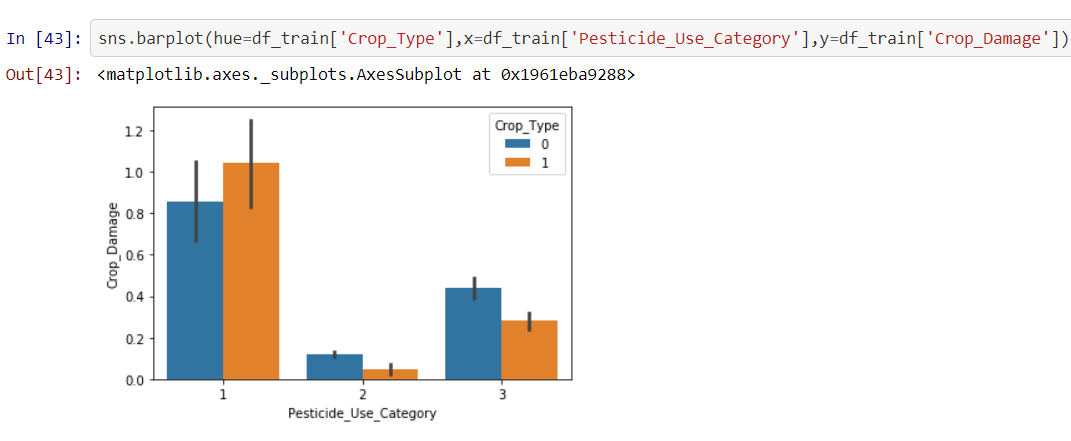


In type-0 soil, damage is higher in type-0 crop,

In type1 soil, damage is higher in type1 crop.

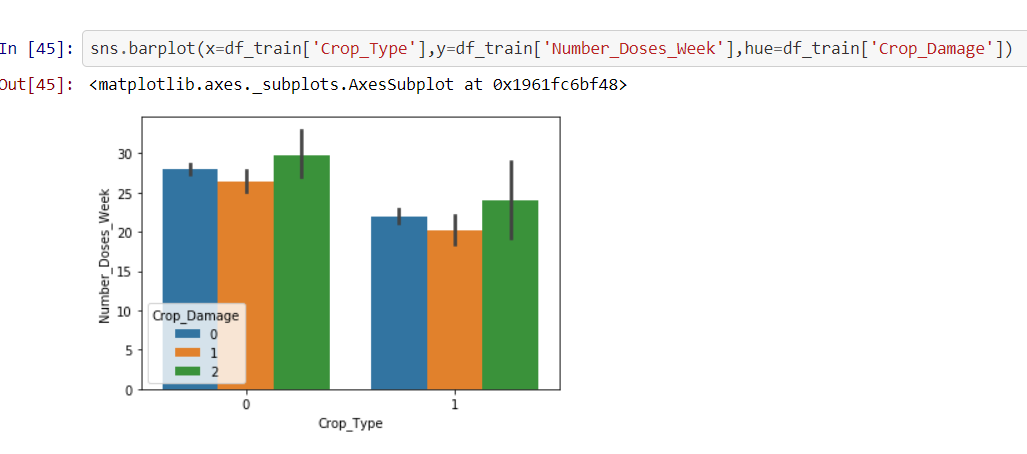
Type 0 soil is suitable for type1 crop. Whereas Type1 soil is suitable for Type 0 crop.

5)Damage in different crops due to pesticide Usage:



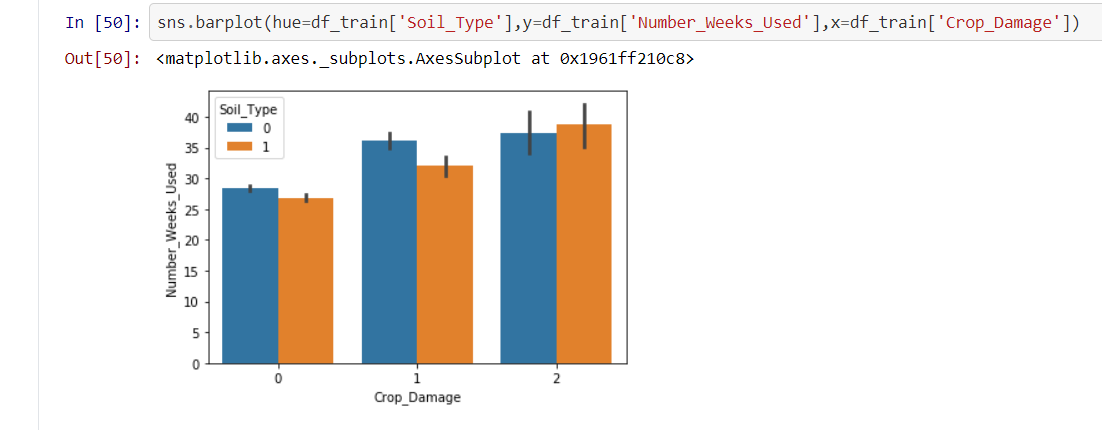
Crop damage is very high when pesticides has never been used. Damage is relatively low when pesticides are not currently used.

6)Damage due to dosage of pesticides per week:



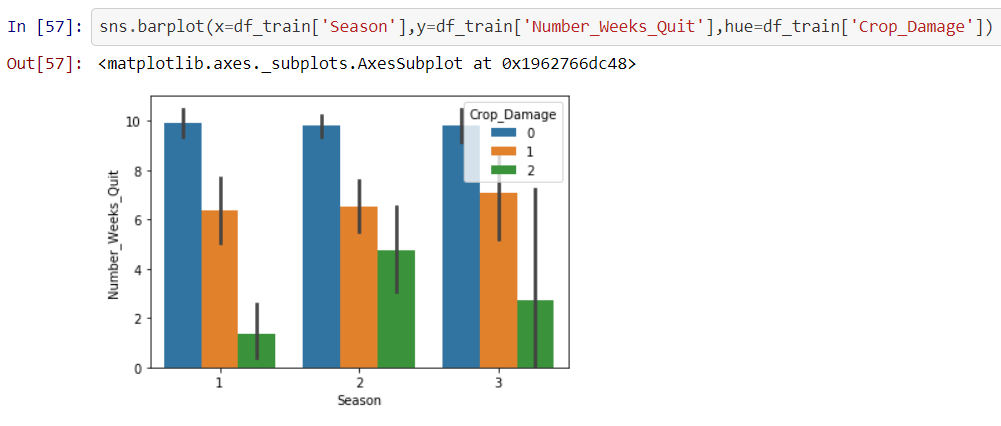
With increase in dosage of pesticides per week, damage due to pesticides is higher in both crop types.

7) Crop Damage in different soils based on weeks of pesticide usage:



With increase in usage of pesticides per week, crop damage due to pesticides is higher.

8) Crop damage in different seasons:



After withdrawing the pesticides, the likelihood of crop survival is very high in all the seasons.

In second season damage due to pesticides is high after withdrawal.

In first season damage due to pesticides is low after withdrawal.

**DATA ANALYSIS CONCLUSION:**

* Likelihood of living crop reduces with increase in insects.
* Type 0 soil is suitable for type1 crop,Type1 soil is suitable for Type 0 crop.
* Crop damage is very high when pesticides has never been used. Damage is relatively low when pesticides are not currently used.
* With increase in dosage of pesticides per week, damage due to pesticides is higher in both crop types.
* With increase in usage of pesticides per week, crop damage due to pesticides is higher.
* After withdrawing the pesticides, the likelihood of crop survival is very high in all the seasons.
* In second season damage due to pesticides is high after withdrawal.
* In first season damage due to pesticides is low after withdrawal.

**DATA PREPROCESSING:**

We usually receive data in the raw form, data pre-processing implies cleaning the data into a suitable form such that it can be used by a machine learning model to perform predictions. It involves variety of methodologies like

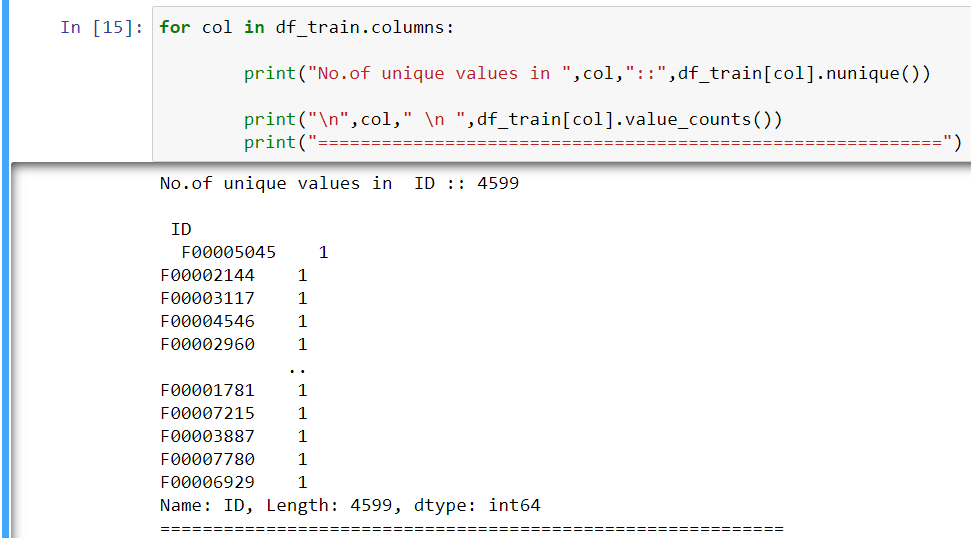
1. Dealing with missing values,

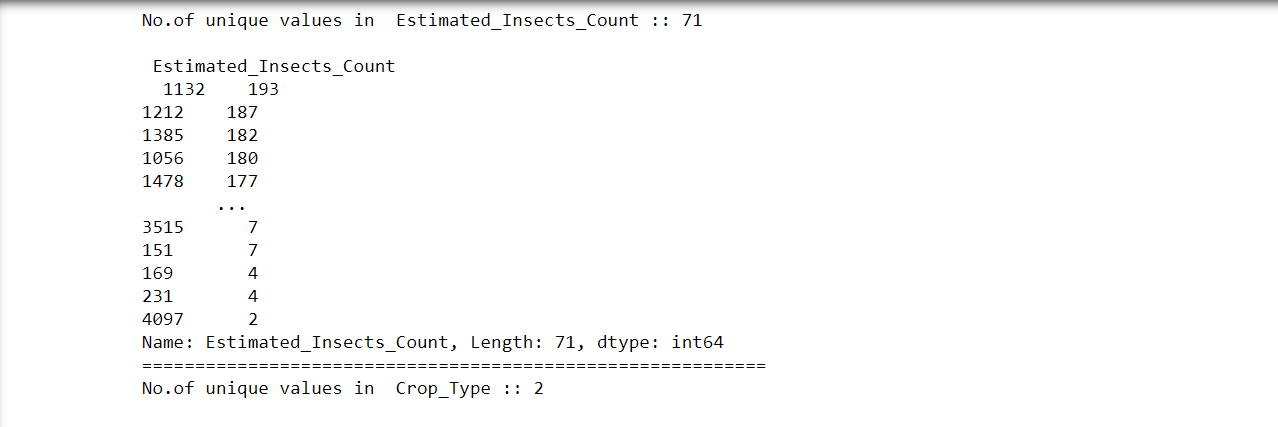
ii) Conversion of object data type column into numeric through variety of techniques like one hot encoding or label encoding. This data set has almost all numeric features so there is no need of encoding the categorical data.

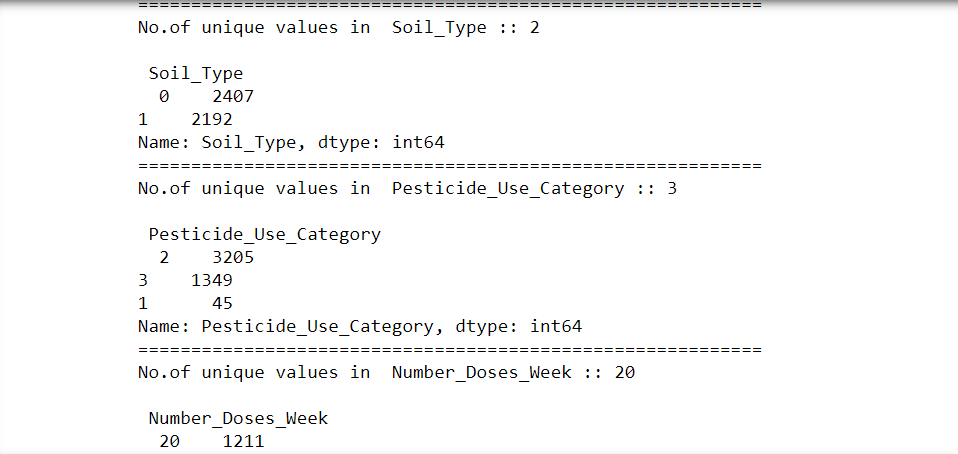
iii) Splitting data into training and testing sets.

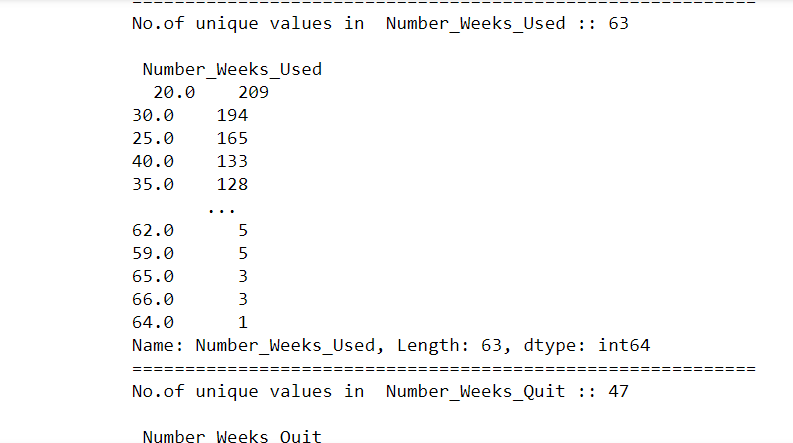
iv) Feature Scaling

First checking the values unique entries in each columns

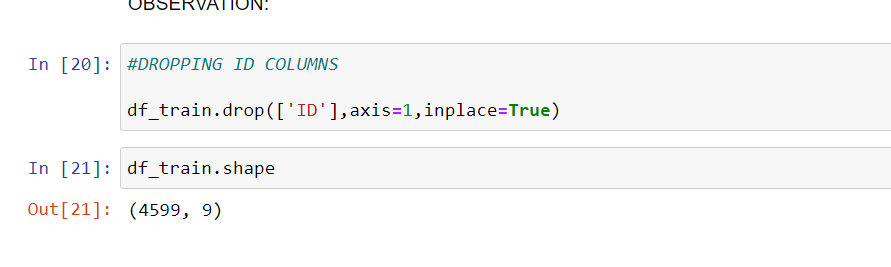




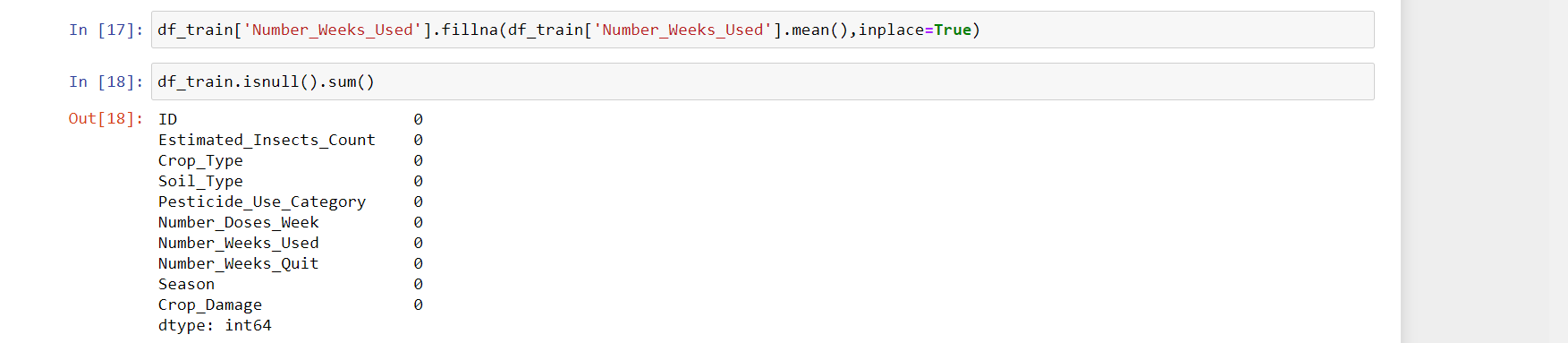




We can notice that ID column has 4599 unique values and as ID doesn’t play much importance in prediction of crop damage, dropping the column.

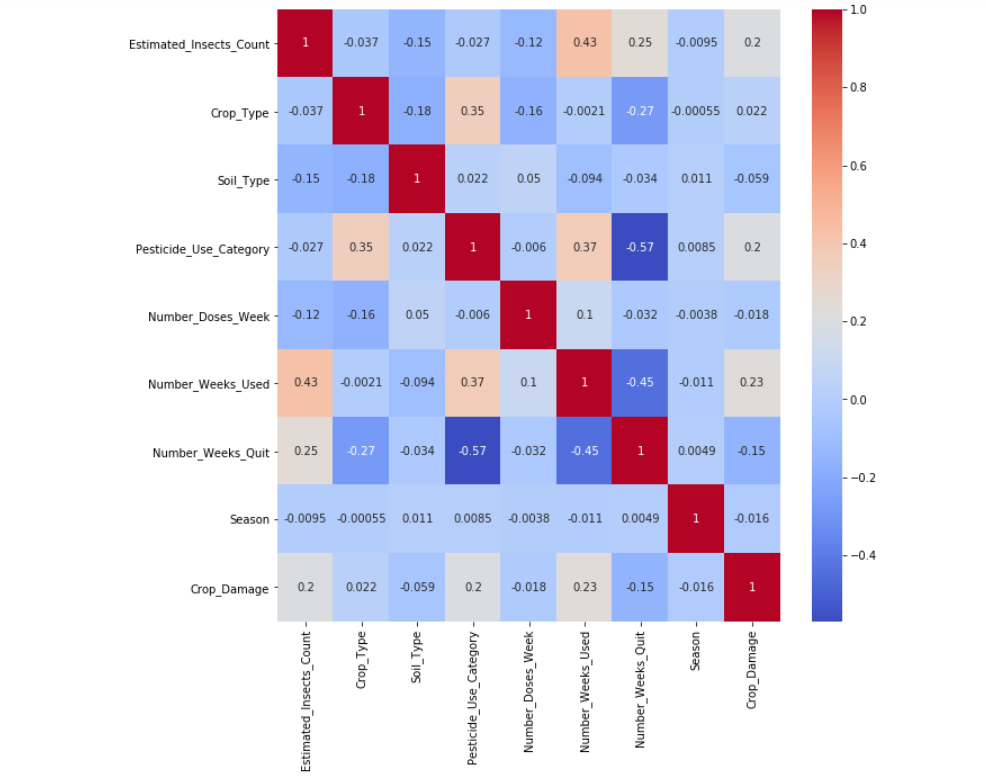


Number\_weeks\_used column has missing values, so replacing them with average no. of weeks used.



Correlation determines how variables depend on the target variable. Correlation is an important factor in building machine learning models. Correlation can be either positive or negative.

**Plotting Correlations:**



Crop damage is positively correlated with:

1 )pesticide use category

2) number of weeks the pesticides used

3) Insects count

4) crop type

Crop damage is negatively correlated with:

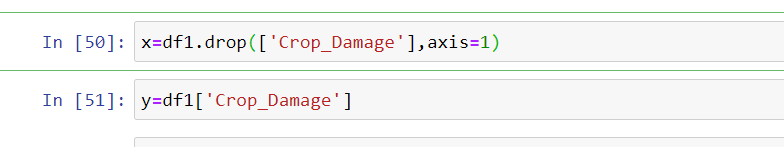
1) soil type

2) Number of weeks since pesticides are quitted

3) season

**Splitting data into training and testing sets:**

Before performing splitting of data into training and testing sets, we need to determine the x and y variables. X will be the list of independent variables where as Y will be the dependent variable on which prediction is done.



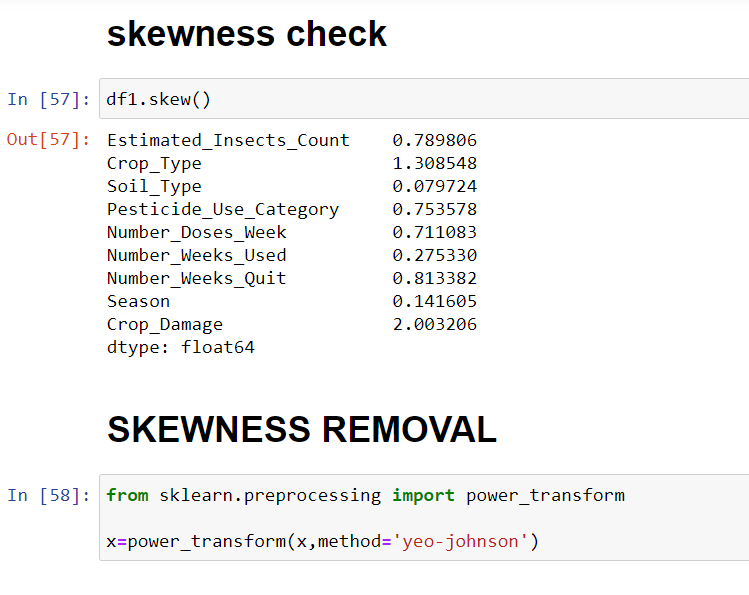
Before proceeding with the train and test split, feature scaling is done. Feature scaling is done in cases where the columns have values of different magnitude.

In this dataset we can notice that, few columns like season , crop type, soil type and crop damage have single entries like 0,1,2 and 3.Whereas rest of the columns have normal values. Performing feature scaling:



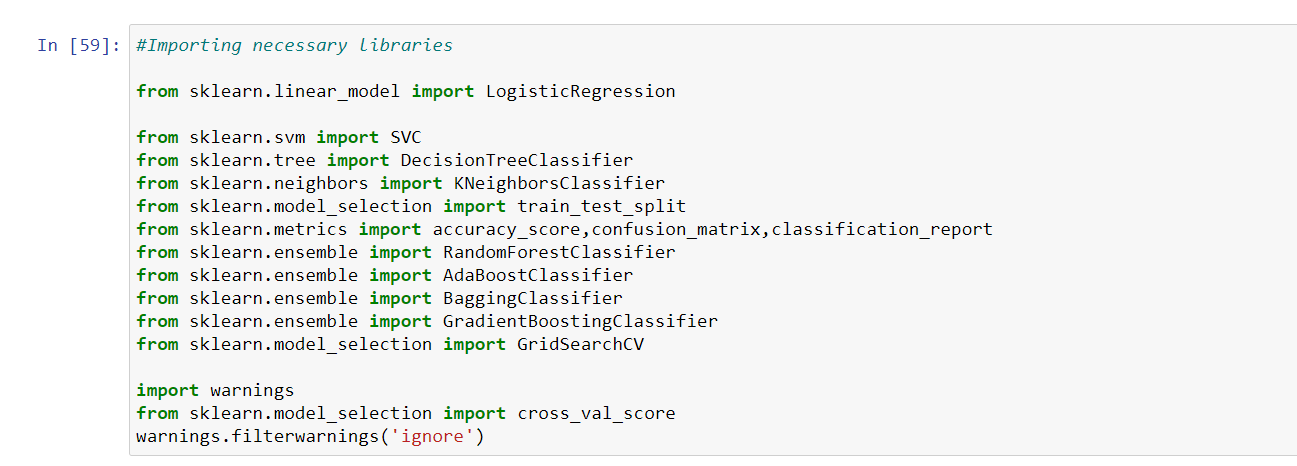
Before performing train and test split, skewness removal is done.

Skewness is a measure of the symmetry in a distribution. Presence of skewness is checked through skew() function.

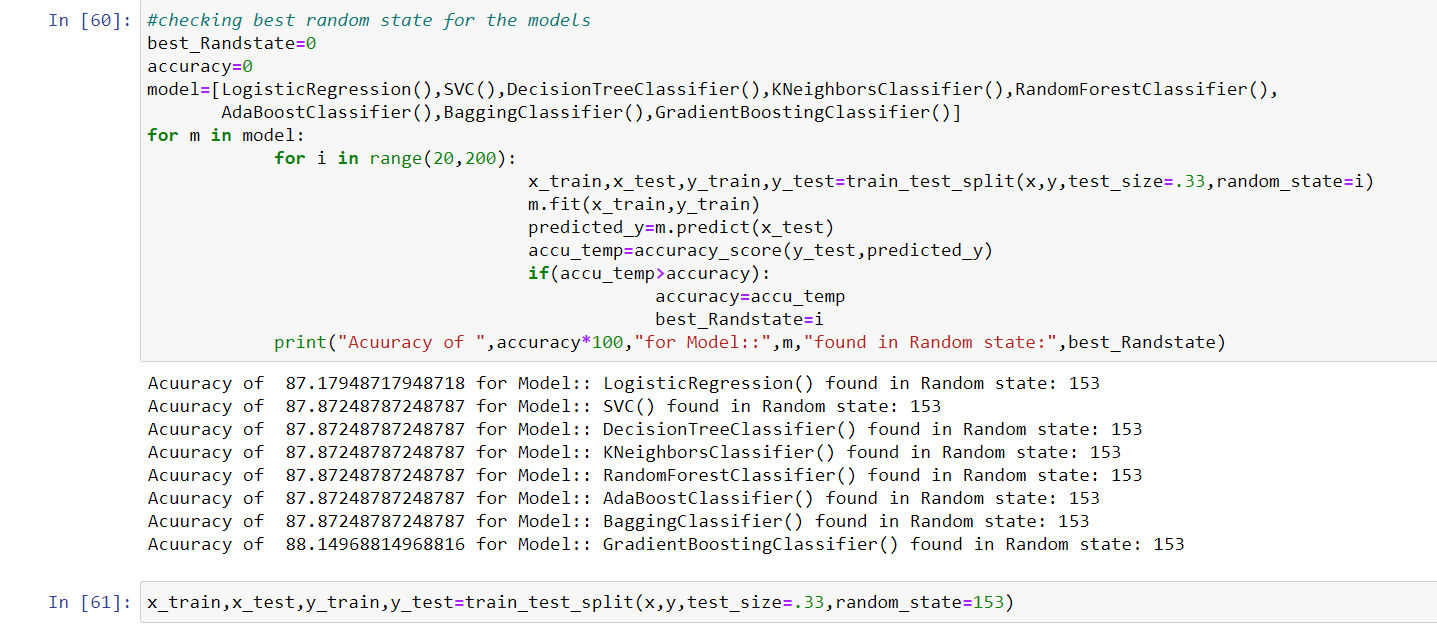


Skewness removal can be done through variety of techniques like box-cox, yeo-johnson, log transformation. I preferred yeo-johnson because it works for both positive and negatively skewed data.

**Importing necessary libraries for model buiding:**



Choosing the best random state for splitting the data:



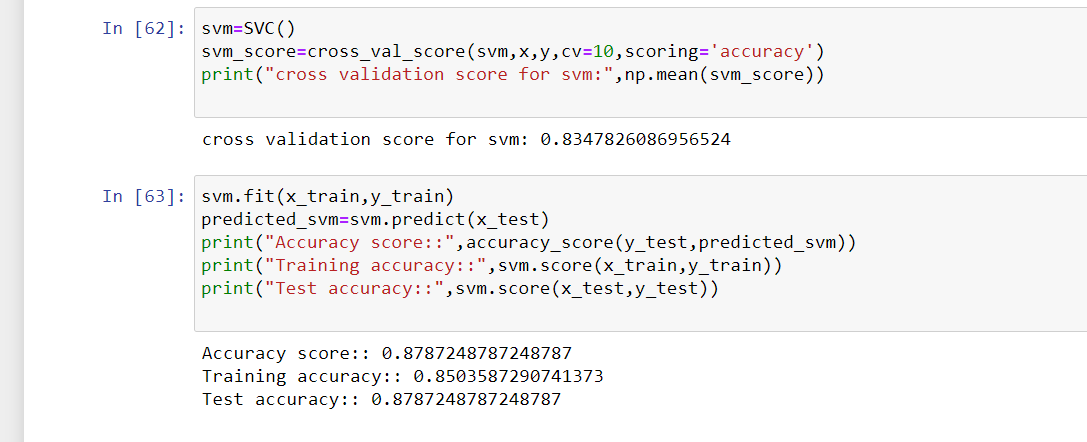
For splitting the data the usual test size is 30/70 or 20/80. Out of all the metrics to evaluate performance of the classification models, I chose accuracy score as the evaluation metric.

We can see that accuracy of 87.8 is achieved for all most all models at random state of 153.

I chose Support Vector Classifier.

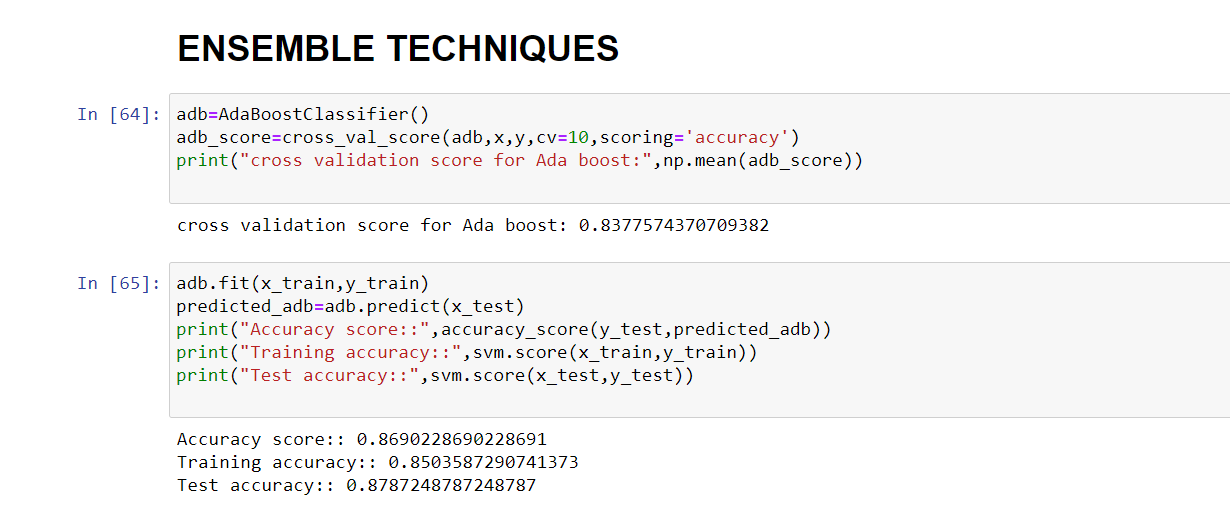
**MODEL BUILDING:**

Cross Validationisa very useful technique for assessing the effectiveness of your model, particularly in cases where you need to reduce overfitting**.**



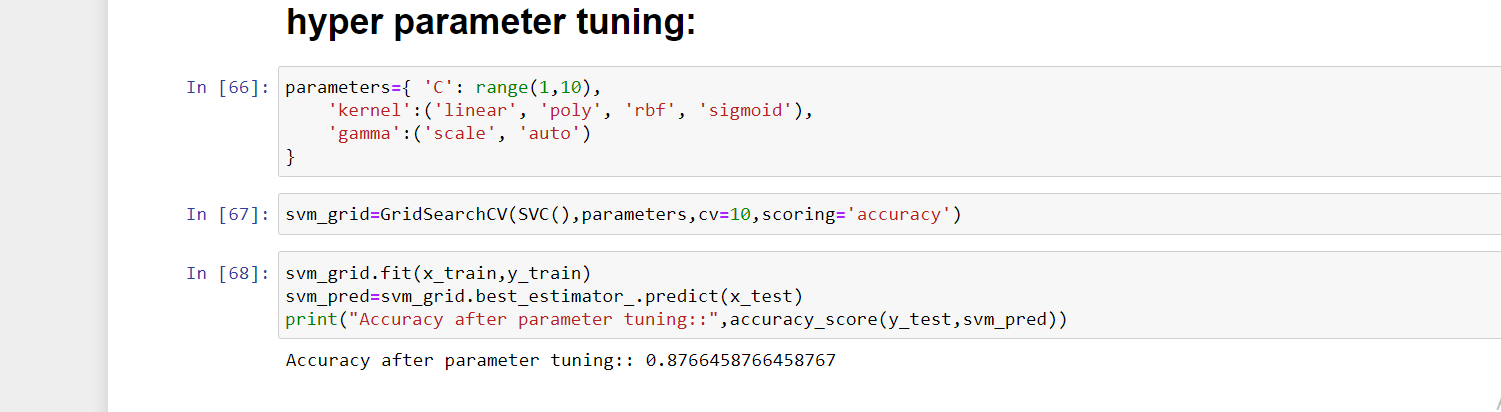
**ENSEMBLE TECHNIQUES:**

Ensemble methods are algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance, bias**, and thus**improve accuracy of predictions.**



**HYPER PARAMETER TUNING:**

A hyper parameter is a parameter whose value is used to control the predictions. Tuning hyper parameters for svc, because both svc and adaboost has same accuracy. Hyper parameter tuning reduces the overfitting. Accuracy has increased after hyper parameter tuning.

****

**SAVING MODEL:**



**CONCLUDING REMARKS:**

The built model predicts crop damage with 87% accuracy. The predictions on test dataset is saved in a csv file named as crop\_damagePrediction.