**Crop Damage Prediction on Agriculture dataset**



1. **Problem Definition:-**

In this project I have used agriculture dataset. Most important field in any part of the world is agriculture. Food is must for every living thing on the earth and for human it is mostly come from farming, UT 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 will use machine learning technique to build the model which will predict the outcome of the harvest season, i.e. whether the crop would be healthy (alive), damaged by pesticides or damaged by other reasons.

1. **Data Analysis:-**

* **Dataset Information:-**

As we know that the task is to predict the outcome of the Harvers Season. In the Dataset, we have to predict the target, 'Crop Damage'. The possible values are 0, 1, and 2. 0 implies alive, 1 implies Damage due to other causes and 2 represents Damage due to Pesticides. The dataset consist of many features like ID, Estimated\_Insects\_Count, Crop Type, Soil Type, Pesticide\_Use\_Category, Number Doses Week, Number Weeks Used, Number Weeks Quit, Season, and Crop Damage. I will use classification technique to predict the crop damage category. Here are the unique feature present in the dataset:-

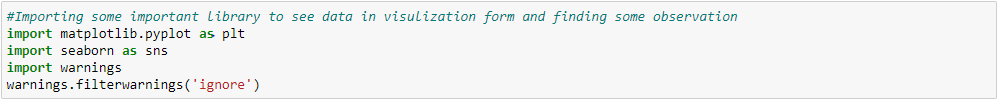
1. ID -> Unique ID
2. Estimated\_Insects\_Count -> Estimated Insect Count per square meter
3. Crop Type -> Category of crop (0,1)
4. Soil Type -> Category of Soil(0,1)
5. Pesticide\_Use\_Category -> Type of pesticides uses (1=Never,2=Previously used,3=Currently using)
6. Number Doses Week -> Number of doses per week
7. Number Weeks Used -> Number of weeks used
8. Number Weeks Quit -> Number of weeks quit
9. Season -> Season Category(1,2,3)
10. 10.Crop\_Damage -> Crop damage Category (0=alive,1=Damage due to other causes,2=Damage due to pesticides)

After observing dataset i found that most label are present in the dataset so I will use classification techniques to predict the target variable which is Crop Damage.

* **Data Import and Analyse:-**

Importing some necessary library-

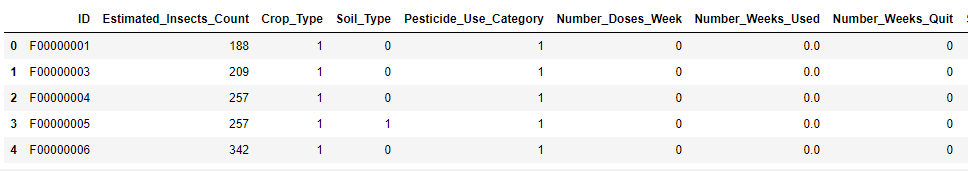


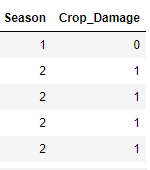


Importing the data-



The data is looking like:-



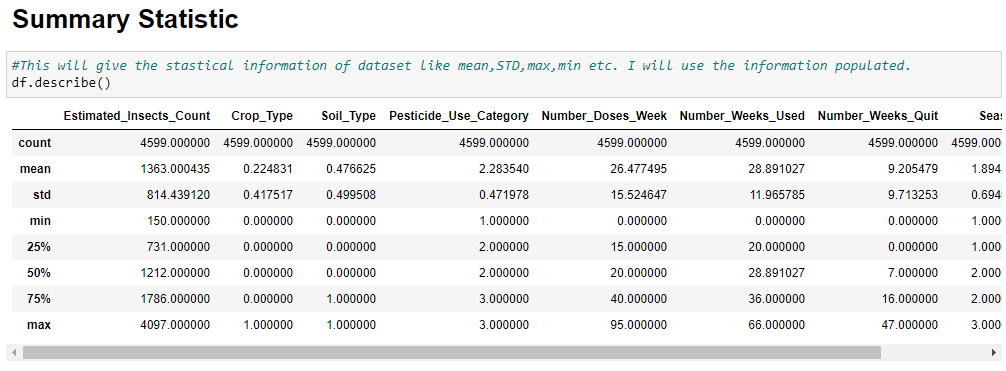


Shape Of the data:-There are 10 columns and 4599 rows present in the dataset. My target variable is crop damage which labelled so I will use classification technique to build the model and predict the crop damage scenarios.

Data types:- In this dataset we have only ID column which is in object data type and rest are integer or float so I don’t have to use label encoding as ID column is no useful and all are unique ID can’t decide the crop will damage or not. So we can drop the column ID from the dataset and rest features data type will not create any issue on building model.

Summary Statistics of the data:-

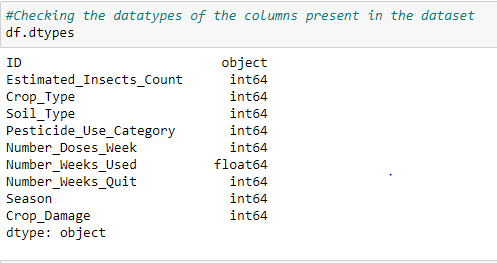
Summary statistics gives us statistical information about the data like mean, median and mode etc. The information is for all features present in the dataset. Descriptive Statistical Analysis helps us to understand your data and is a very important part of Machine Learning. This is due to Machine Learning being all about making predictions. On the other hand, statistics is all about drawing conclusions from data, which is a necessary initial step. Below are the visual of statistical summary.



Key Observation:-

1. We can see that there is 4599 values available across all columns so we can say that there is no missing value present in the dataset.
2. Standard deviation is very high in the column Estimated Insects Counts so data in this columns is spread and range is high.
3. STD is also high in columns Number Does Week, Number Does Used and Number Weeks Quit so range is high also in these columns and data is not normally distributed in these columns.
4. Mean is higher than median across all columns accept of Season. It means skewnees is present in these columns and columns are right skewed or positive skewed.
5. In season column median is high than mean so it is left skewed or negative skewed.
6. Difference between min and max is high in most of the column which shows high range and spread of data.
7. Difference between 75 percentile and max is high in columns like Estimated Insects Count, Number Does Week, and Number Does Used Number Weeks Quit. So by this observation we can say that there is outliers present in these columns.

Checking data types of all columns present in the dataset:-



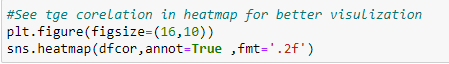
After observing the data types of the column present in the dataset I observed that the all columns data type is integer accept of ID which is object and Number weeks used which is float. I will drop ID column as it is no impactful on decide the crop damage conditions.

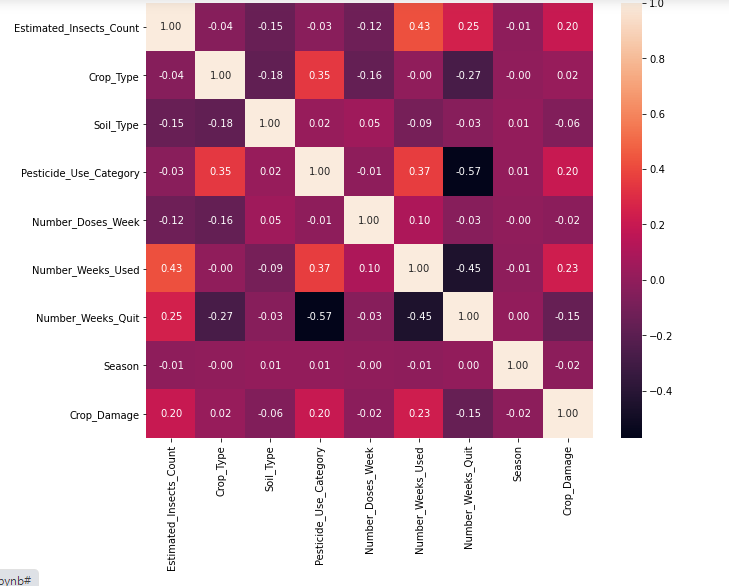
1. **Exploratory Data Analysis:-**

I have many features which may be useful to decide the crop damage labels as they are positive correlated or negative correlated with the target variable crop damage. Here I will check the correlation of all columns with my target variable crop damage and see which feature is more useful and which is less.

* Checking correlation of all features with our target feature Class.

Here we are checking the correlation of all features with the target class by which we will decide that which feature is good correlating with class and which is bad. It will also give use information that which feature is positively correlated with the target variable and which feature is negatively correlated with target feature.





Observation:-

By plotting correlation heat map we can see that the how other columns is correlated with our target column Crop Damage. Below are the insights i get from above correlation heat map. ->

1. The columns Estimated Insects Counts, Pesticide Use Category, Numbers Weeks Used, Crop Type are positive correlation columns with my target variable.

2. Soil Type, Number Doses Week, Number Weeks Quit and Season are negative correlated with the target variable.

3. Highest +ve correlated column = Number Weeks Used

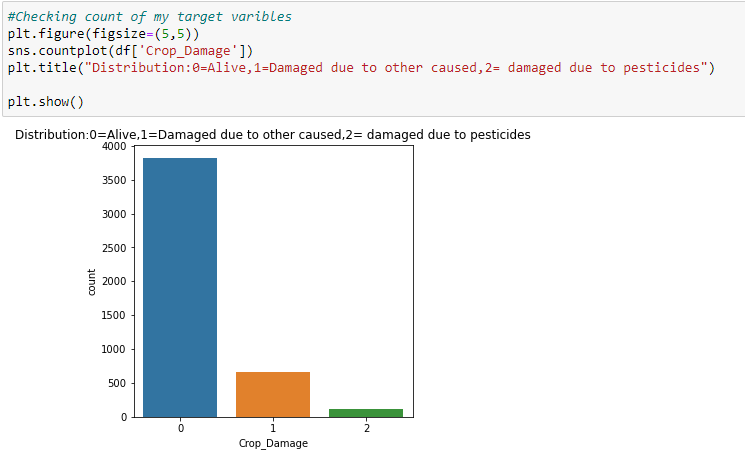
4. Highest -ve correlated column = Number Weeks Quit

All columns are participating in correlation with target variables and both -ve and +ve correlation exist.

I have some missing values present in the dataset which I have replaced by taking mean of the column where missing values present.

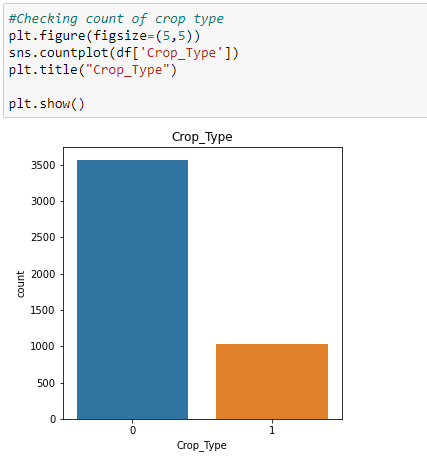
Data Visualization:-

Checking Count of all classes present in the dataset:-



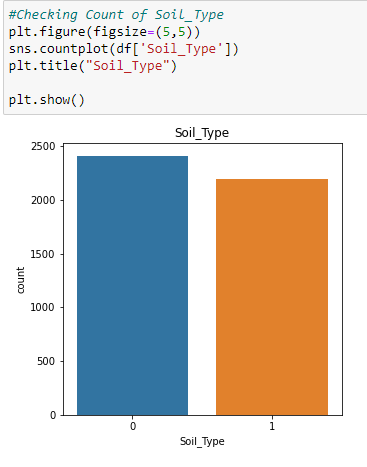
We can see here the more data is for alive. However the crop damaged due to other caused is higher than the crop damaged due pesticides.

Checking which type of crop present in the dataset and how much:-



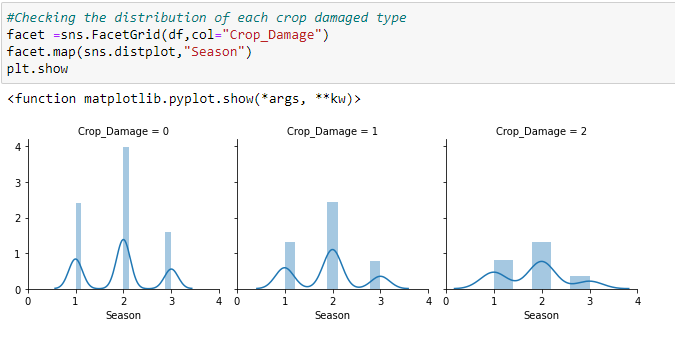
Crop Type 0 is present more than 1 in the dataset.

Checking Soil Type present in the data:-



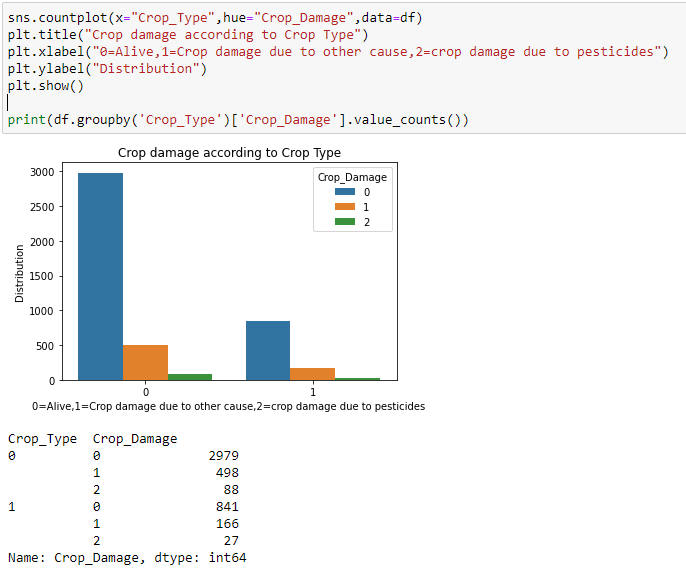
Soil Type 0 is present more than 1 in the dataset.

Crop damage according to season:



Here from above observation we can see the Crop damage season wise. The Crop is survived and alive high in season 2 then 1 and in season 3 it is very less. Crop damaged by other cause is high in season 2 and then season 1 and lowest in season 3. Crop damaged by pesticides is high in season 2 then season 1 and lowest in season 1.

Crop damage according to crop type:-



As per above observation one answer we can find:-

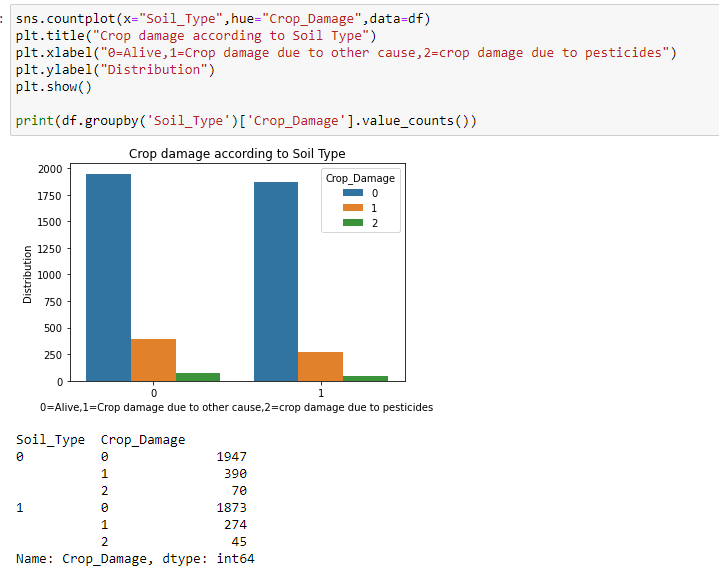
In Crop Type 0 -> Maximum values are present of crop is alive with 2979 values. Then crop damage due to other reason with values 498 values. Least crop damage due to pesticides with values 88 values.

So by above observation I can say that on the basis of this dataset when coil type is 0 there are maximum chance that crop will not damage due to any reason.

In Crop Type 1 -> Maximum values are present of crop is alive with 841 values. Then crop damage due to other reason with values 166 values. Least crop damage due to pesticides with values 27 values.

So in crop type 1 there is less data present as compare to crop type 0. But the same result here also in coil type 1 the chance of crop is alive higher than it will damage due to other reason and then damage due to pesticides.

Crop damage according to soil type:-

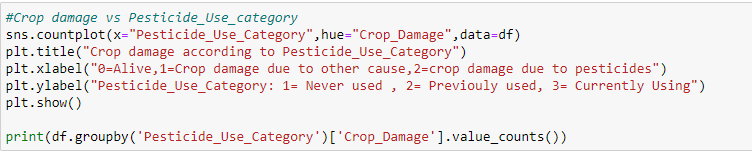


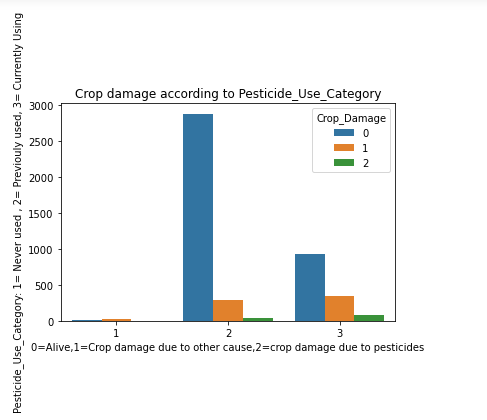
As per above observation we can see that there is two types of soil present in the dataset.

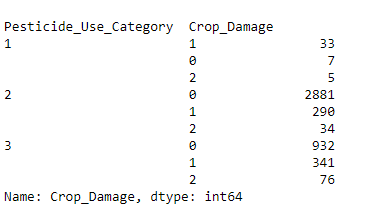
Soil Type 0 -> in this soil type we can see that there are 1947 times the crop is alive and then 390 values present when crop is damaged due to other cause whereas 70 values present when crop is damage due to pesticides. So we can see that there is maximum values present in the dataset for soil type in the case when the crop is alive as compare to damage.

Soil Type 1:- In this soil type the crop is alive 1873 times and damaged due to other cause 274 times and then damage due to pesticides 45 times. So we can see that the same like soil type 1 the crop is alive maximum time when soil type 1.

Crop damage according to Pesticide use category:-







Here from above observation we can see that there are 3 category in Pesticide\_Use\_Category.

Pesticide\_Use\_Category (1=Never used) -> So in the scenario where pesticide never used we can see that the crop damage due to other cause with 33 values. Then the crop is alive 7 times and damage due to pesticide 5 times. So we can easily see that no use of pesticide is not good option as maximum time the crop is damaged.

Pesticide\_Use\_Category (2=Previously used)-> In the scenario when pesticide used previously we can see that there is maximum chances to survive the crop as here crop is alive 2881 times and then crop is damage due to other cause 290 times and then crop is damage due to pesticide is very less 34 times only. So this is best option to use pesticide here we can see that.

Pesticide\_Use\_Category (3=Currently Using)-> In this scenario we can see that from above observation there is maximum scenarios when crop is alive which is 932 times. Then crop damage due to other cause 341 times and the crop is damaged due to pesticide is 76 times. So in this option also the chances to survival of crop is very good.

By observing all scenarios I can say that the chance to crop survive we must use pesticide.

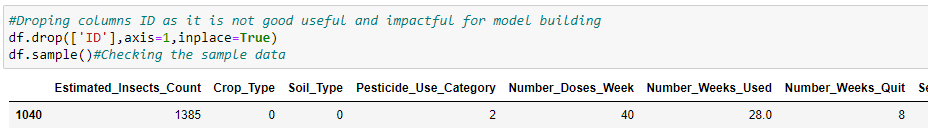
By above observation we can understand that there is outliers present in the dataset and we have to remove the outliers from the dataset to make sure that model will perform well and give good result.

1. **Pre-processing Pipeline:-**

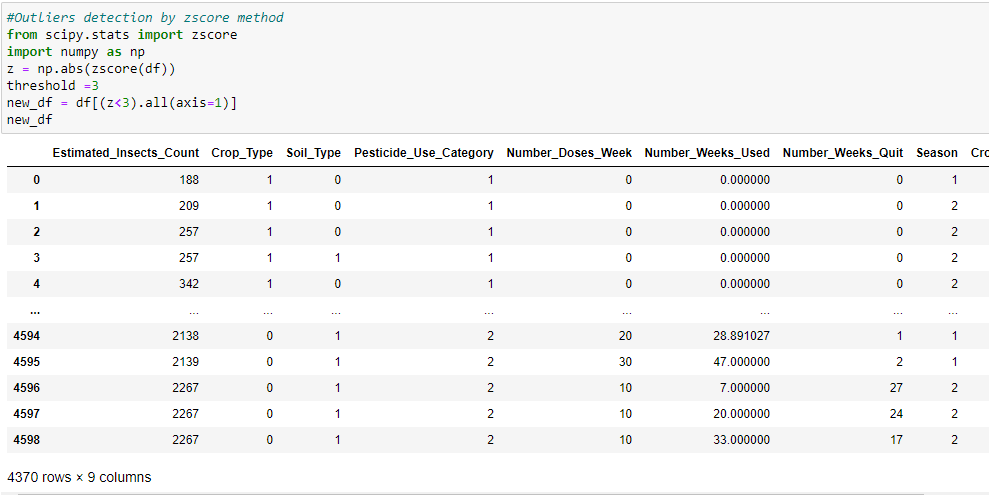
We have to pre-process data before the data is ready for build the model. If there is outliers, skewnees and any missing value present in the data then the model will not perform well and give correct prediction. I will use certain steps to prepare the data for model building mentioned below:-

**Step1:-**

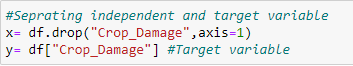
I have dropped ID column.



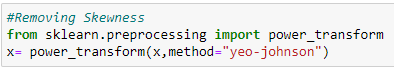
I will remove outliers by using zscore method by taking threshold value 3 and I will remove those values which zscore is above 3 and select only those data which zscore is under 3. Here for the data frame df I have created new data frame called new\_df. In this new data frame we have all data within the threshold value 3. Below is the code for that. After removing the outliers the shape of new data is 4370 rows × 9 columns.



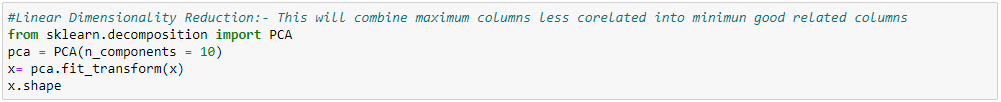
**Step2:-** After removing outliers I will separate new data frame into two part X and Y. In Y I am taking target variable class and in X I have taken rest columns.



**Step3:-** In this step I will remove skewnees from the data so my model will behave well and give good accuracy. Power transforms are a family of parametric, monotonic transformations that are applied to make data more Gaussian-like. This is useful for modelling issues related to heteroscedasticity (non-constant variance), or other situations where normality is desired. Currently, Power Transformer supports the Box-Cox transform and the Yeo-Johnson transform. The optimal parameter for stabilizing variance and minimizing skewnees is estimated through maximum likelihood. I will use power transform with “yeo-Johnson” method to transform the data and remove skewnees. I will remove skewnees in only x variable as the skewnees and variance is necessary in target variable.



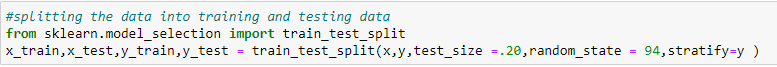
**Step4:-** Principal Component Analysis (PCA) is used to explain the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality-reduction technique. The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables. I will use PCA to minimize the number of columns to make less and more impactful and effective which is good correlated with the target variables.



1. **Building Machine Learning Models:-**

Now I will start building model and the whole process is divided into some important steps listed below:-

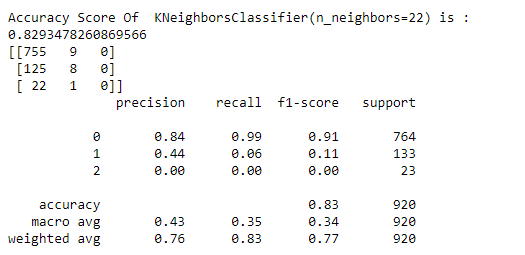
**Step1:- In** this step I will split the data into training and test data. By training data I will train the model and later I will check the performance of model by using test data. I will use train\_test\_split function to split the x variable into train and test data.



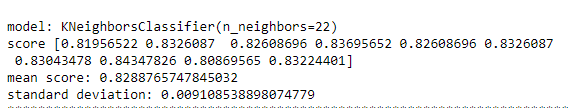
**Step2:-** I will use multiple algorithms and see the model performance and accuracy. As in this project I am using classification techniques so for checking the model performance I will use some important metrics. I will use classification report, accuracy score, confusion matrix and F1 score to see the model performance and decide which algorithm I will select and finalise for model.

Below are the different models used and their performance report:-

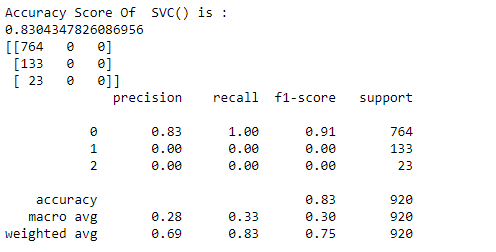
1. KNeighborsClassifier:-



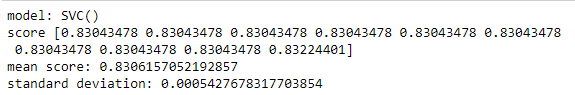
Cross Validation Score:-



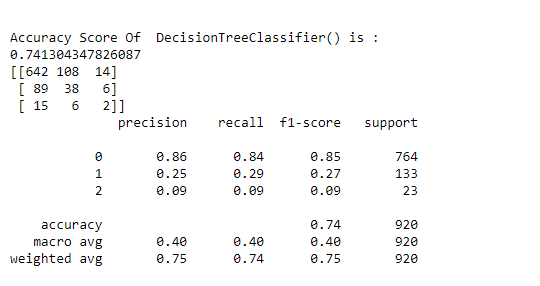
1. Support Vector Classifier :-



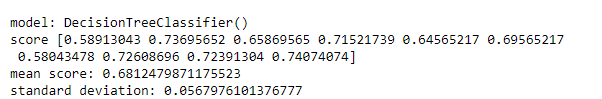
Cross Validation Score:-



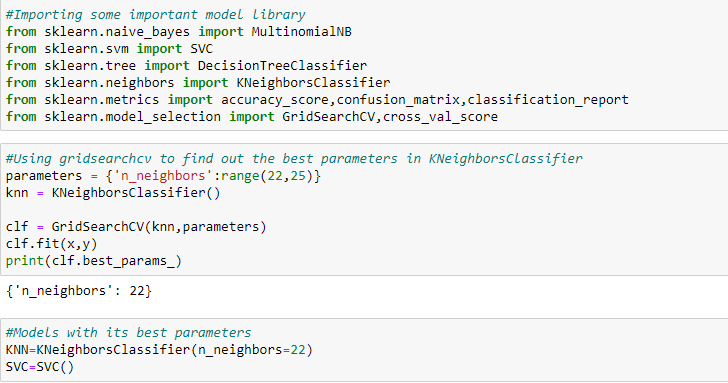
1. Decision Tree Classifier:-

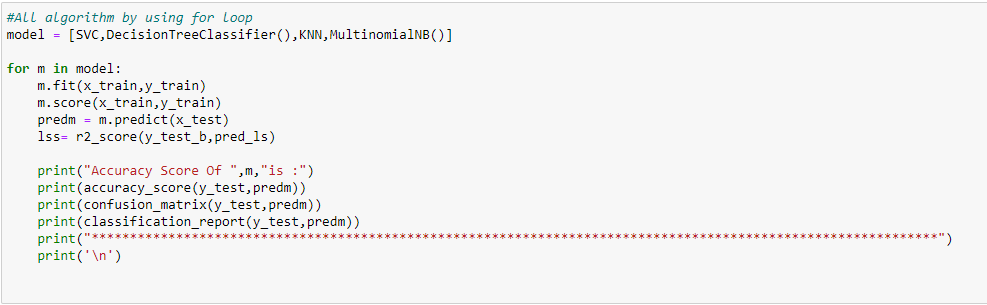


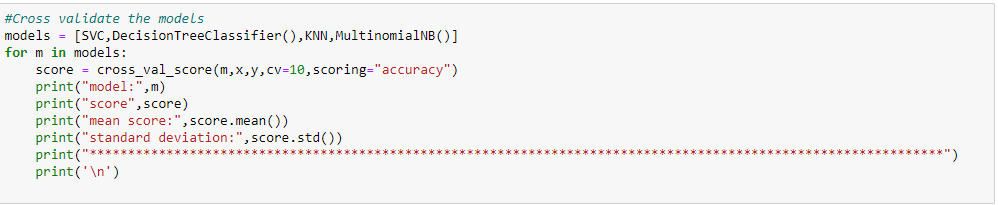
Cross Validation Score:-



Code I have used for get these results are:-

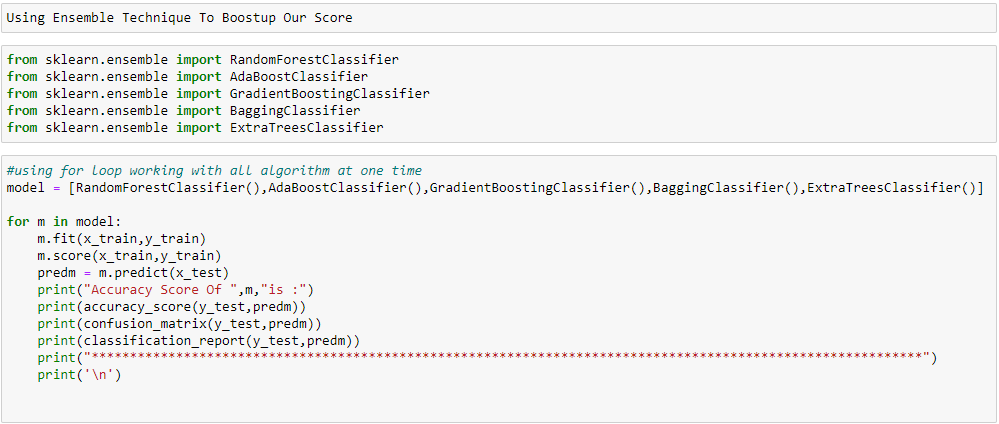


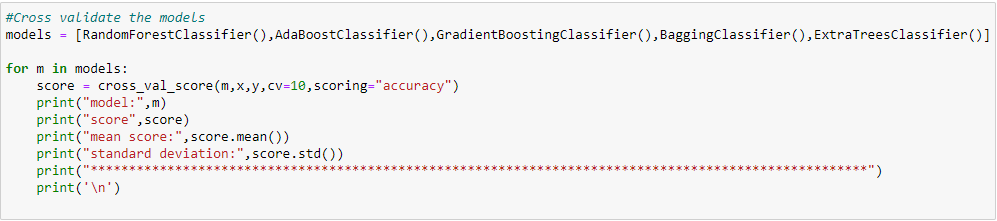




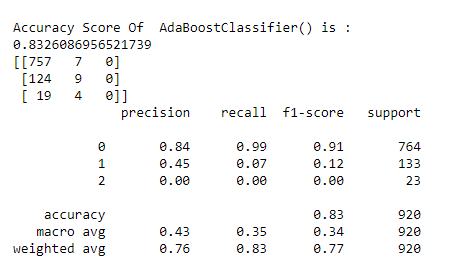
From above all observation I found that SVC giving best score with accuracy score of 83.04% at the same time cross\_val\_score is 83.06%. Now I will use some boosting technique to boost up the score and then I will decide that which I will select and save.

**Ensemble Technique to boost up the score:-**

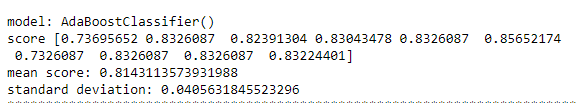




1. Adaboostclassifier:-



Cross Validation Score:-



After using all techniques I found that the there are two models are working well with both accuracy score and cross\_val\_score and minimum difference between both scores. Models are = Support vector classifier (SVC) and Random Forest classifier.

SVC: - Accuracy Score = 0.8304347826086956 cross\_val\_score = 0.8306157052192857 difference = 0.00018092261059010628

ABC: - Accuracy Score = 0.8326086956521739 cross\_val\_score = 0.8143113573931988 difference = 0.01829733825897506

I will use Adaboost Classifier () as the difference between accuracy score and cross Val score is lower than the difference in Support vector classifier. Also Ada Boost classifier as it is encoded model also minimize the chances of underfiting and overfiting.

What is Adaboostclassifier:-

Adaboost, short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapiro, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance.

Reason for Choosing the random forest algorithm:-

An Adaboost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. This algorithm not only gives good results for binary classification, but it also can be extended for multiclass classification. It does not require feature scaling. The algorithm is capable of capturing nonlinear information inherent in the data. Additionally, it can capture feature interactions. Also we have already seen that this gives best accuracy score and cross validation score.

What is Accuracy Score:-

Accuracy score is the score when we test the predicted result of x test data with actual y test data and see how much percentage it is predicting good. As we splitted train and test data into x\_train, y\_train, x\_test and y\_test. So we train the model by using x\_train and y\_train and test with x\_test. When the model give the result then we test it with actual y\_test data and see that how much percentage it has predicted correctly and that score is accuracy score.

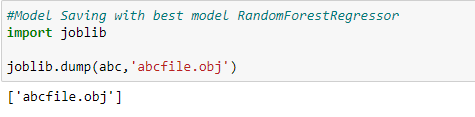
What is cross validation score:-

We use cross validation technique which is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. Cross validation score is the score in your validation set. We see that that our model is prevent from over fitting or under fitting we use cross validation technique.

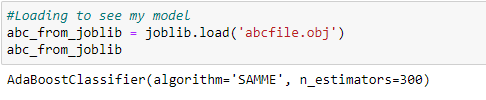
1. **Concluding Remarks:-**

Now I will save the model for production and also we will produce the result after using the same model. As in this case I have selected the Random Forest Classifier so I will save this. Let see in some steps:-

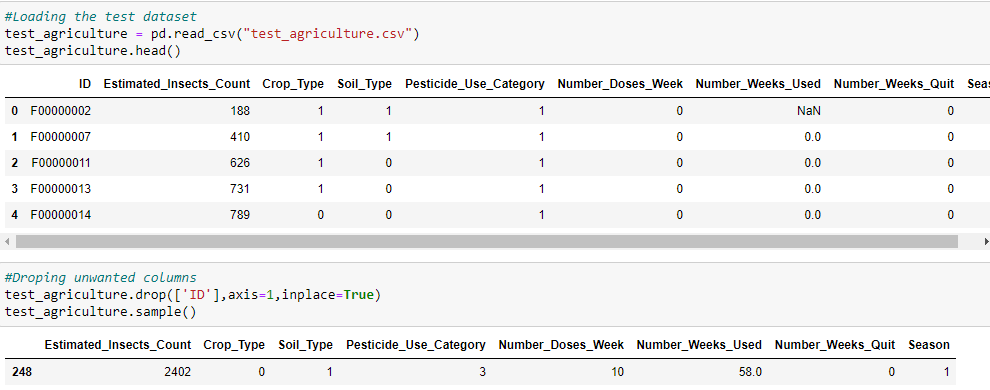
Step1:- In this I will save the model.

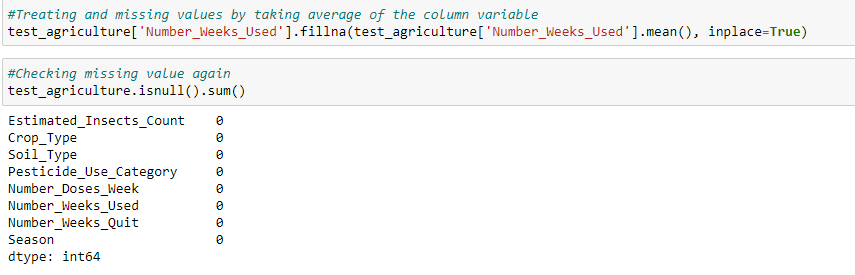


Step2:- Loading the Model for production

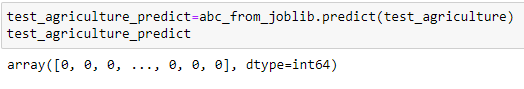


Step3:- Load the test data and pre-process the data using same pipeline use in Train data:-

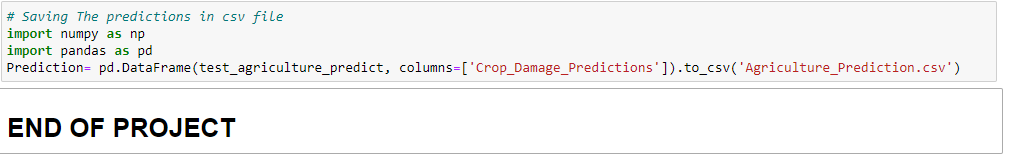




Step4:- Predicting the result:-



Step4:- Saving the result in CSV (Comma separated value) format.



**Conclusion:-**

The Adaboost algorithm worked well for Agriculture data. We ran the algorithm over the data using the Jupyter Notebook framework. We were able to classify the crop damage with an accuracy of 83%. It seems unlikely that any substantial gain in classification accuracy would result using any other algorithms or approaches on the same dataset, given the classical wisdom that ‘invariably, simple models and a lot of data trump more elaborate models based on less data. However, some significant gains could well occur if more spectroscopically classified data (training data) were available. However, given more such training data, better algorithms may also be possible.

I hope the article will help you to understand how I classified and predict the Sloan digital sky server dataset whether it is star, galaxy or quasar.

Thank You…….

Author:-

Vikash Kumar



The simplest way to describe me would be a technophile. I have an extremely curious mind with a knack to tinker with trending technologies and resolving real-world problems by trying to visualize the problem from a different perspective. Last but not least at the end of the day I really want my work to create a constructive impact on society for its upliftment.