Machine Learning in Agriculture

Problem Definition:

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.

In this article I will try 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 using machine learning techniques.

Data Analysis:

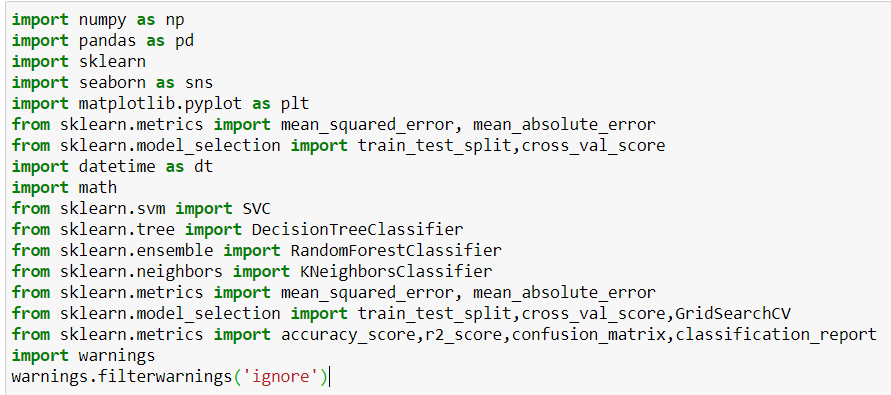
In this article we will be using data which is based on crops harvested by various farmers at the end of harvest season.

We have separate data for training and testing purpose in this example.

Following are the columns in our data and its description:

1. ID: Unique id for each entry in the dataset. This column is of object data type.
2. Estimated\_Insects\_Count: Estimated Insects Count per square meter. This column is of integer data type.
3. Crop\_Type: Category of crop. This column is of integer data type and has only 0 and 1 value in it.
4. Soil\_Type: Category of soil. This column is of integer data type and has only 0 and 1 value in it.
5. Pesticide\_Use\_Category: Type of pesticides uses. This column is of integer data type and has values: 0 – Never used, 1 – Previously used and 2 – Currently using.
6. Number\_Doses\_Week: Number of doses of pesticides per week. This column is of integer data type.
7. Number\_Weeks\_Used: Number of weeks pesticides used on crops. This column is of floating data type.
8. Number\_Weeks\_Quit: Number of weeks pesticides are not used on crops. This column is of integer data type.
9. Season: Category of season. This column is of integer data type and has values 0,1 and 2.
10. Crop\_Damage: Category of crop damage. It is of integer data type and has values: 0- alive, 1- damage due to other reason, 2-Damage due to pesticide usage.

Here **Crop\_Damage** is our target variable which we will try to determine in our machine learning model. Now since our target variable can have either of 3 values (0, 1 and 2) this is a clear case wherein we will have to use classification type of algorithms.

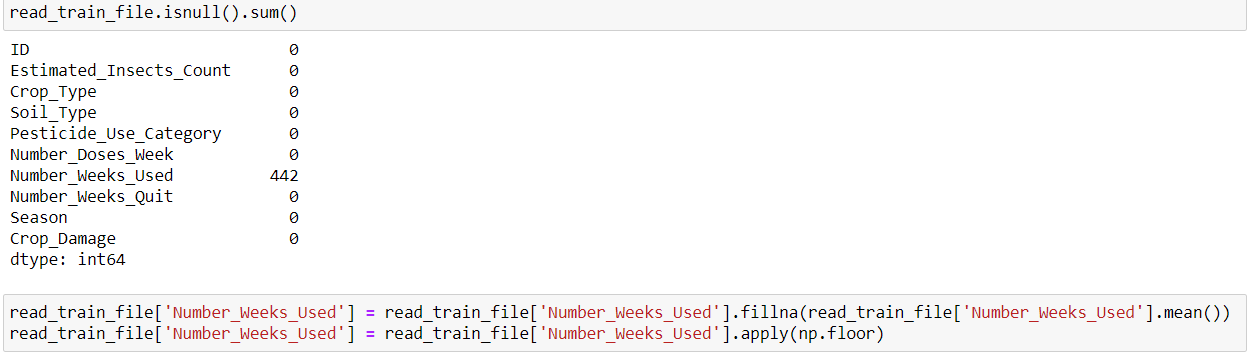


Here we are importing all the necessary libraries and classes which we will be using in our model building process.

Since this is a classification type of problem we are importing various classifier libraries and all the matrices libraries which are used for this type.



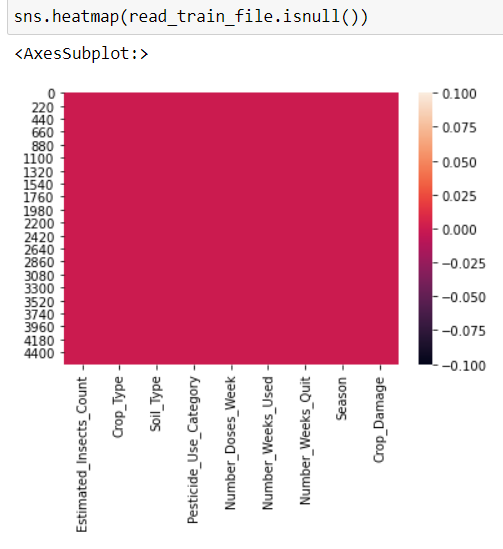
Over here our input file/data is in xlsx file. We first load data in a variable and then use .to\_csv method of pandas to convert it into csv format.



Here we can see that column ‘Number\_Weeks\_Used’ has 442 null value entries which we are replacing with mean value of this column. Now since this column is of integer data type, we are using ‘np.floor’ method to extract integer part of mean value.

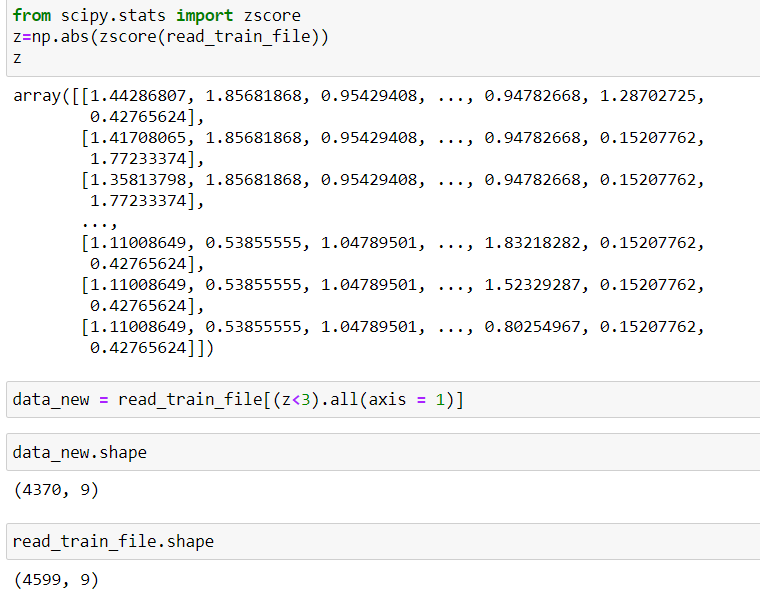


Since ID column has unique value for each data entry, we are going to drop this column.

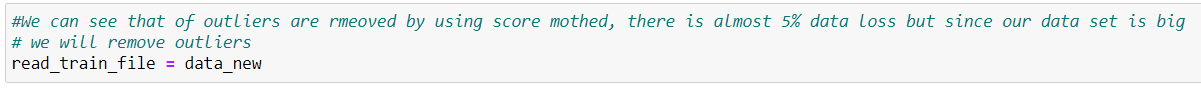


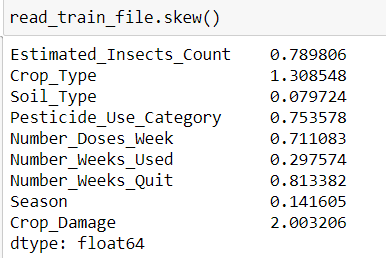
Heatmap is indicating that there are no null values present in our dataset.

By using box plot method we can see that there are outliers present in 'Estimated\_Insects\_Count','Number\_Doses\_Week', 'Numbner\_Weeks\_Used' and in 'Number\_Weeks\_quit' for which we need to work upon.

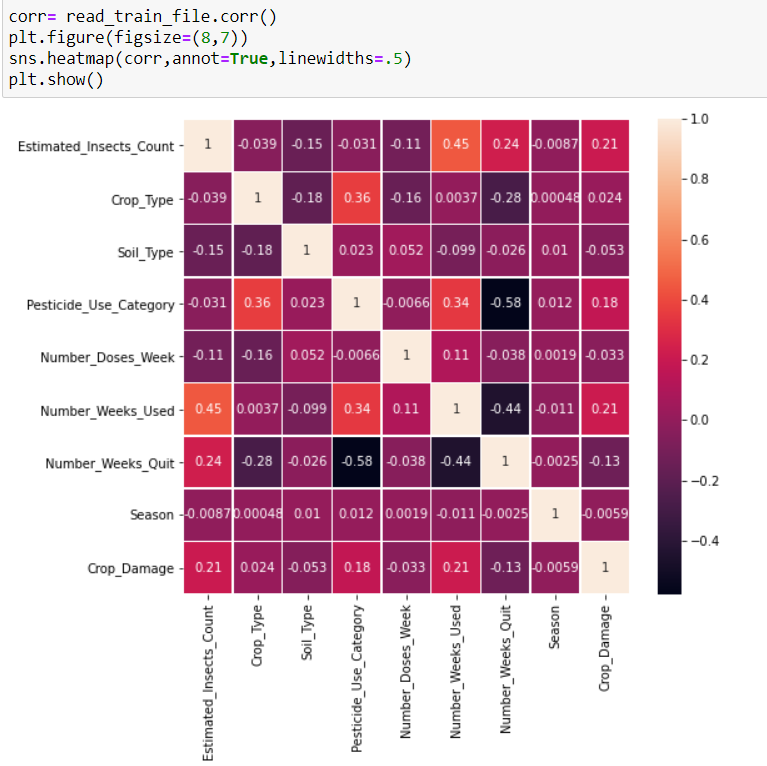


By removing outliers using z-score technique we can see that there is 5% data loss but since our data set is big we will remove outliers.

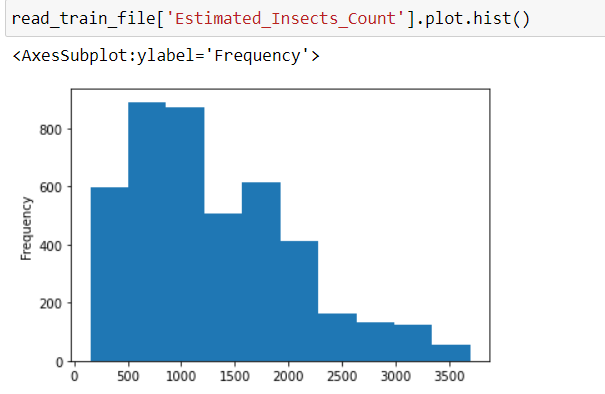




After removing outliers from data we can see that there is very little skewness in our data.



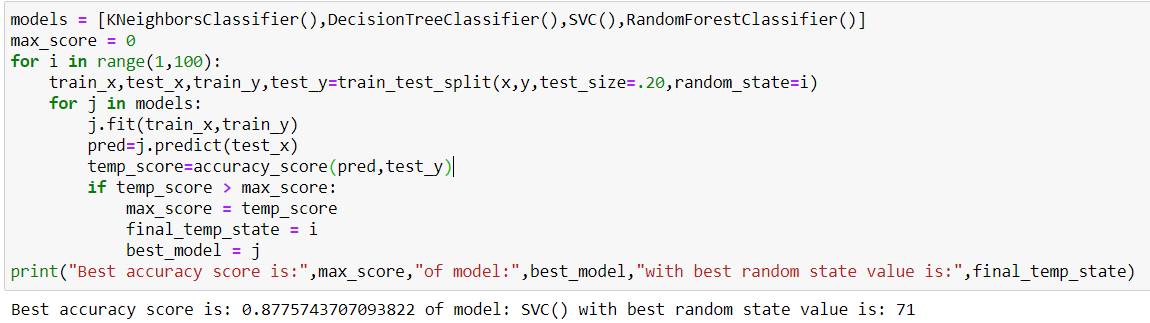
Above heatmap indicates that columns 'Number\_Weeks\_used' and 'Number\_Weeks\_Quit' are correlated with each other.



We can see 'Estimated\_Insects\_Count' column is right skewed and hence we are going to treat skewnesss.



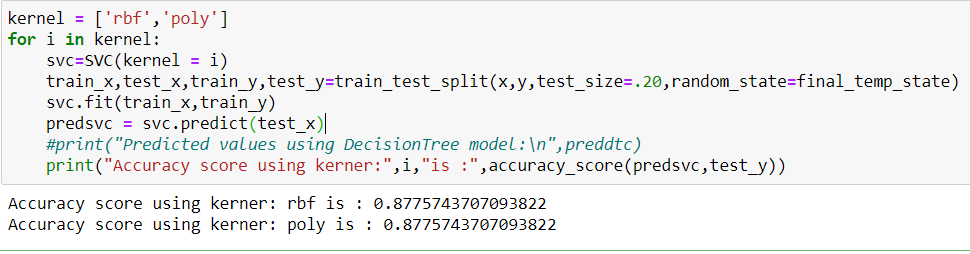
We are treating skewness by using ‘yeo-johnson’ method.



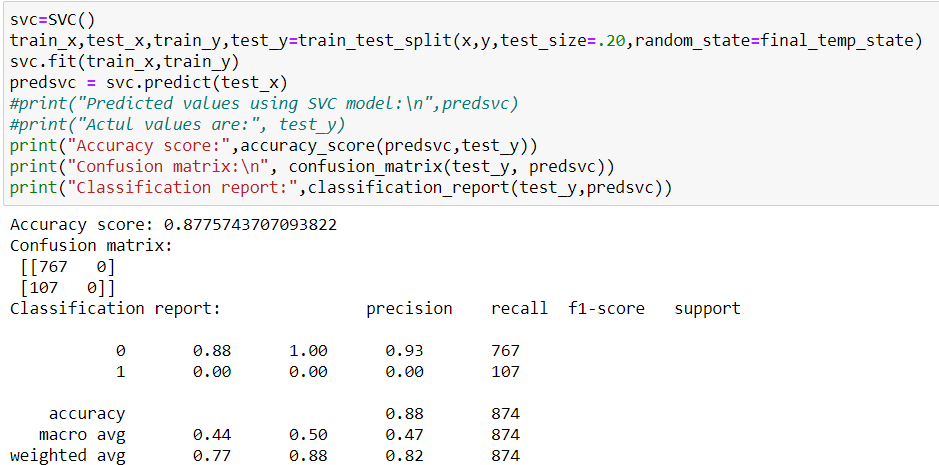
Here we are using ‘train\_test\_split’ method to split data into train and test parts by using test size of 20%. So here 80% of data is used for training and then remaining 20% is used for testing.

We are testing our model with KNeighborsClassifier(), DecisionTreeClassifier(), SVC(), RandomForestClassifier() methods to check which algorithm gives us the best performance and for which best random state value between 1 and 100.

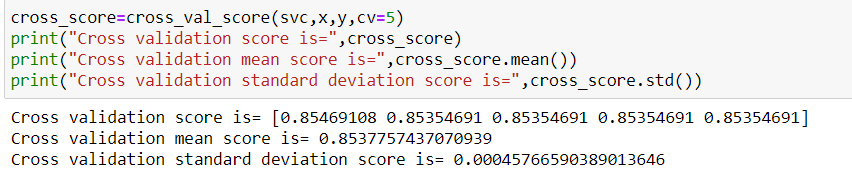
We can see that Support Vector Classifier (SVC) algorithm gives us the best score of 87.75% for random state value of 71.



Now we are checking accuracy score of SVC for both its kernel and can see that score is same for both kernel and hence we will go ahead with default value of ‘rbf’.

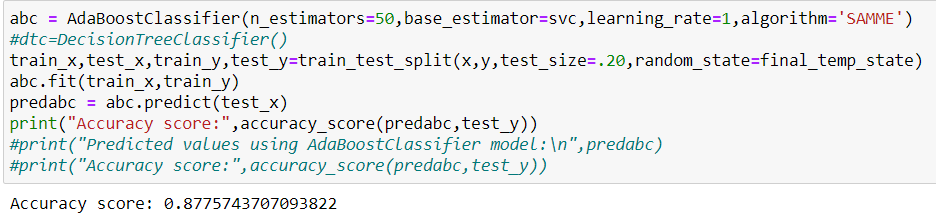


Here we are predicting the output by using ‘predict’ method and then using this predicted values for finding out the accuracy score, confusion matrix and classification report.

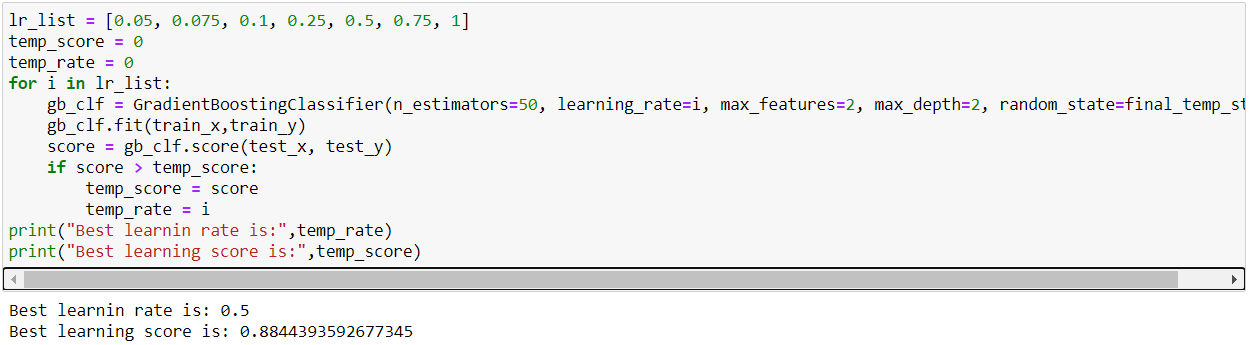


Now we are checking cross validation (score, mean and std) values of SVC with cv value of 5 so that our entire dataset is distributed into training and testing values in 5 iterations and can see score of **85.37%** for cross validation mean score.

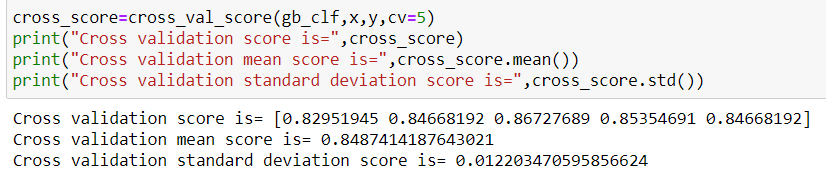
Now we will try to boost the performance of our model using AdaBoostClassifier and GradientBoostingClassifier algorithms.



We can see we are getting score of **87.75%** using AdaBoostClassifier.

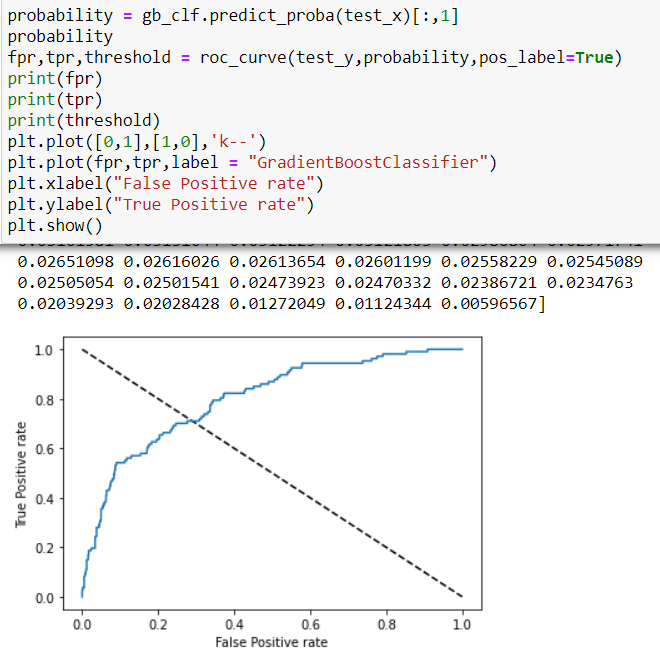


Here we can see that for GradientBoostingClassifier, best learning score is found to be **88.44%** for learning rate of 0.5.



Here we can see that Cross validation mean score of GradientBoostingClassifier is **84.87%.**

By looking at accuracy score and cross validation mean score of SVC and GradientBoosting which are almost same we will be using GradientBoosting algorithm for our final model.

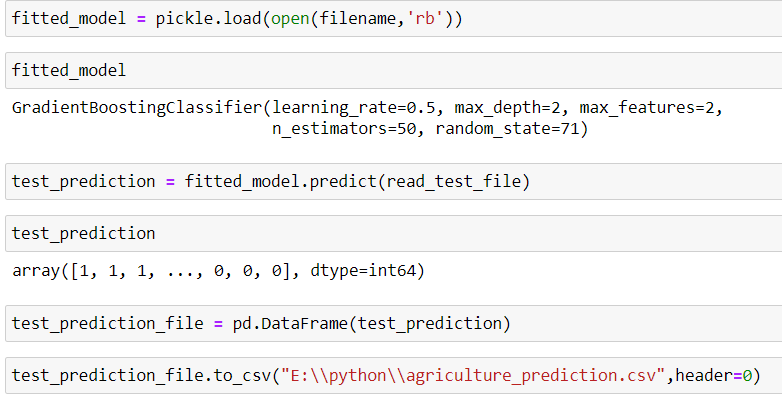


Here we are drawing the AUC-ROC curve for our model which shows a good amount of data lies within our curve which indicates our model is giving a good performance.



Now we are saving our model in a pickle object so that it can be used for our test dataset.

From our training data set we have dropped column ‘ID’ and have replaced null value in column 'Number\_Weeks\_Used' with mean value. We will have to perform same operations on test data also.



After performing data cleaning process on test data also we are now using our saved model on test data.

Here we are using load method of pickle library to do this.

We have loaded our model in ‘fitted\_model’ variable and then used predict method to find value of output variable ‘Crop\_Damage’ for our test data.

Complete source code of my model can be found at path:

<https://github.com/japitale/DT_Evaluation_projects/blob/main/ProjectsDynamicsM20_project10_agriculture.ipynb>

**We can see how machine learning technology can be used to predict the outcome of the harvest season, i.e. whether the crop would be healthy (alive), damaged by pesticides or damaged by other reasons.**