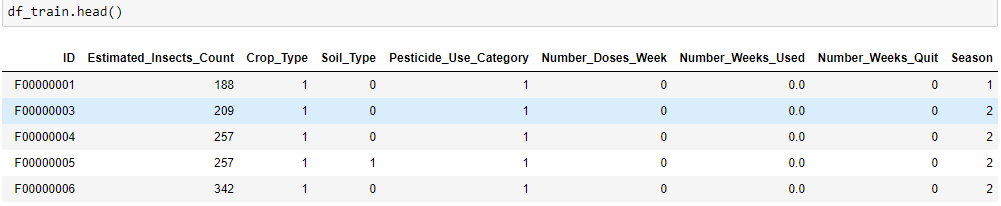
Agriculture Crop Damage Classification

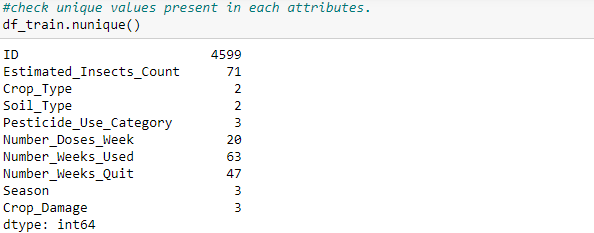


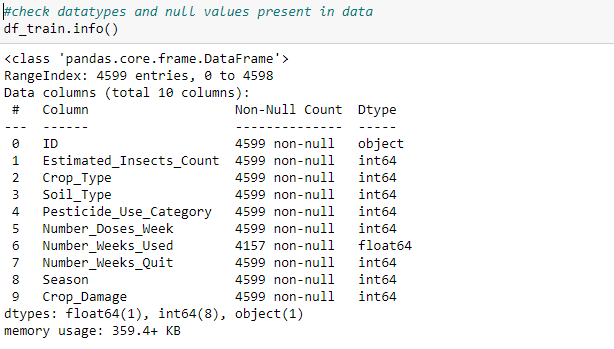
* **Problem Statement:**
* A good harvest is based on various factors like protecting crops from insects, soil type, use of pesticides and many other factors.
* Use of pesticides is crucial factor for crop, because if uses of pesticides are more than requires amount then it damage the crop.
* Here, we have dataset of agriculture and purpose of this dataset is to predict the crop condition, i.e. whether the crop would be healthy, damaged by pesticides and damaged by other reasons. This is classification issue.
* Train set and test set is provided for this dataset. Train set have 4599 records and 10 variables, while test set have 1199 records.
* All variables are related to farming. Out of 10 variables, one variable is dependent (target) variable and 9 variables are independent variables.
* Independent variables contain data like, crop type, amount of insects per square meter, soil type, use of pesticides etc. Target variable is crop damage with possible outcome of 0, 1 and 2, where 0 = crop is healthy or alive, 1 = crop is damage due to other reason, 2 = crop is damage due to pesticides.
* **Exploratory Data Analysis (EDA):**

**Data Preparation:**

Data preparation is one of the most crucial aspects of machine learning. It takes more time if data is more complex or uncleaned. Let’s look at the data.



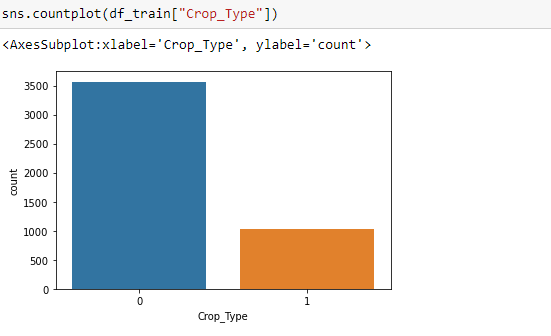




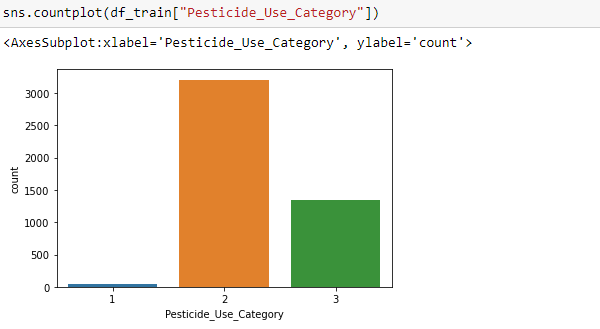
* From the above observation, we can see that null value is present in number\_weeks\_used column. This column depicts number of weeks we used pesticides. We fill null values of this column by its most frequent value.
* We also drop ID column because it is just identification number and not useful for our model.

**Data Analysis:**

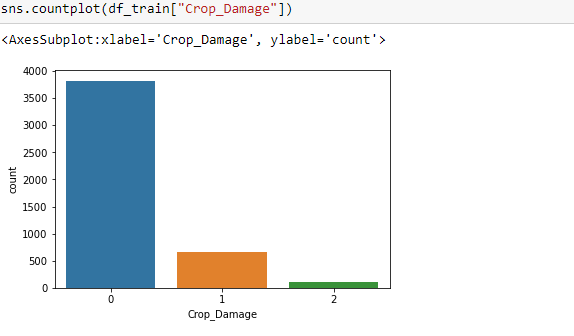
Now, let’s look at the data and gain some useful information by analysis:



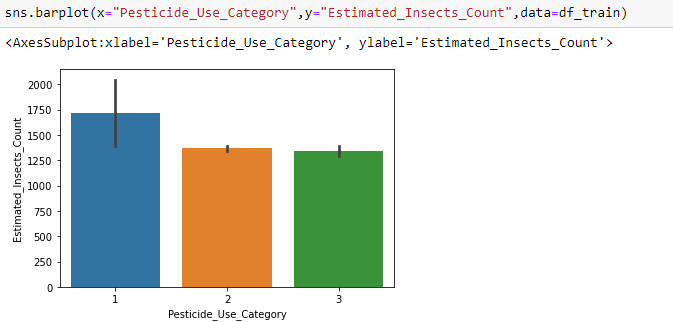
From the above count plot, it is clear that there are two types of crop 0 and 1. 0 type of crop is more preferred by farmers.



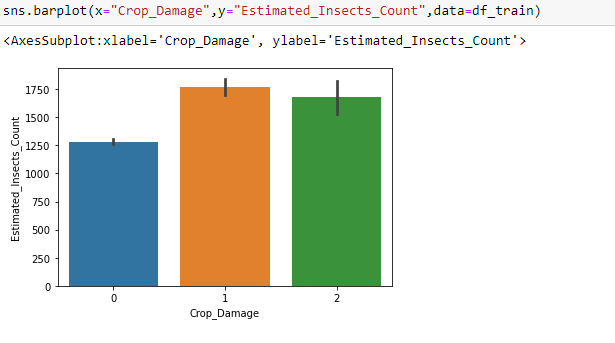
There are three categories in pesticides\_use\_category column. Where 1 = farmers never used pesticides, 2 = previously used pesticides and 3 = currently used pesticides. Maximum number of farmers used pesticides previously, while very few farmers never used pesticides for farming.



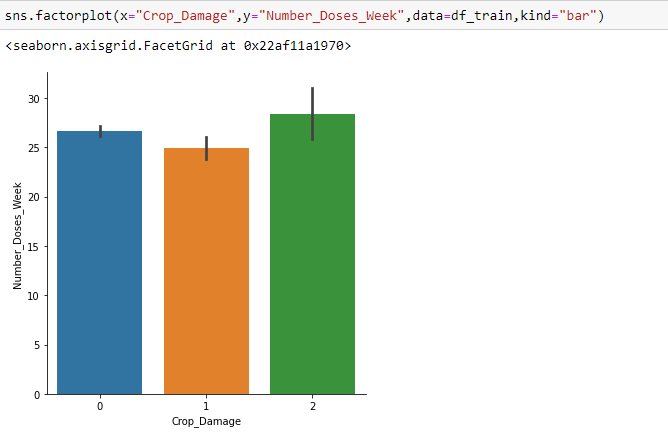
Crop damage column has three categories 0, 1 and 2, which is described in problem statement. From the above plot we can see that, there is imbalanced classification in target variable and this classification badly affect our model and also reduce accuracy of model, so we solve this issue in afterword process.



From the above bar plot of pesticides\_use\_category vs. Estimated\_insects\_count, we can see that insect’s counts are maximun, where pesticides are never used and insects counts are equal, where pesticides previously used and currently using.

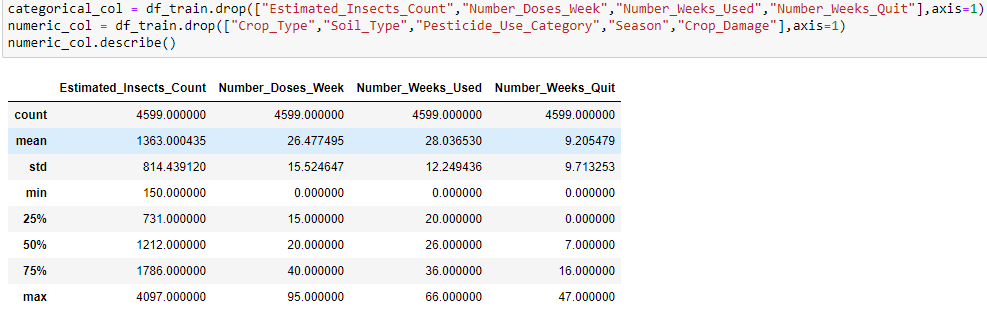


From the above bar plot, it is clear that numbers of insects are less where crop is healthy, which is around 1250 insects per square meter. Numbers of insects are more, where crop is damaged due to other reasons and damaged due to pesticides, which is above 1700 insects per square meter.



From the above bar plot, we can see that crop damage due to pesticides is high, where pesticides number of doses per week is high.

* Now, split the numerical columns and categorical columns for further analysis and for removing outliers. For removing outliers we only consider numerical columns because remove outliers from categorical columns is not a good idea and badly affect our model.
* First, we check the statistical summary of numerical columns.

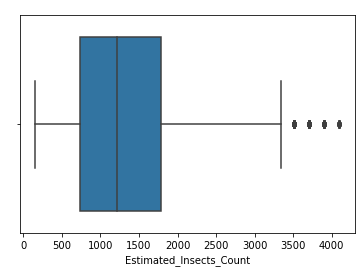


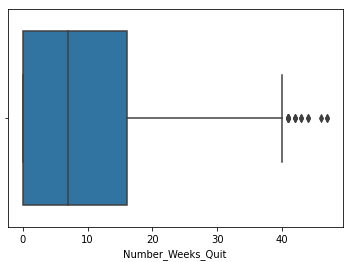
From the above observation we can conclude that:

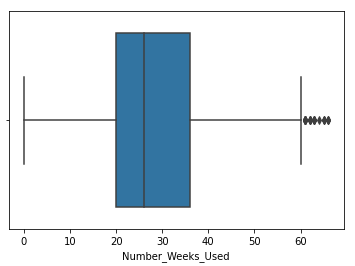
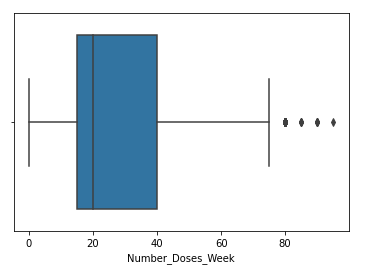
* Minimum insect’s counts per square meter are 150 and maximum insect’s counts per square meter are 4097.
* Pesticides minimum doses per week are zero and maximum doses per week are 95.
* There is large difference between 75 percent of the data and maximum value of the data in every attributes. It represents that outliers present in data and need to remove. Outliers are extreme values that fall a long way outside of the other observations.
* There is large difference between mean and standard deviation in every attributes except Number\_Weeks\_Quit. It depicts that skewness is present in data and need to remove. Skewness refers to a distortion that deviates from the symmetrical bell curve or normal distribution in a set of data.

**Handling Outliers:**

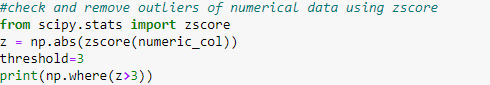
Now, we check and remove outliers from data:

First, we check outliers by draw boxplots.





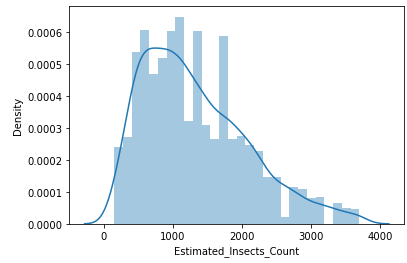
* From the above box plot, we can see that black dots are known as outliers. Outliers present in every numerical attributes and need to remove, because outliers badly affect our model and predict wrong results. So it is necessary to remove outliers.
* Here, we use Zscore method for removing outliers. Code of our dataset is as below:

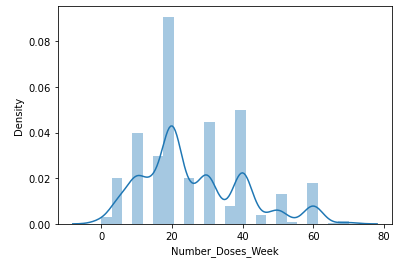
  

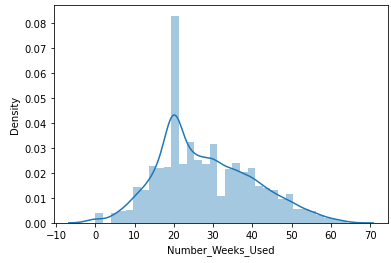
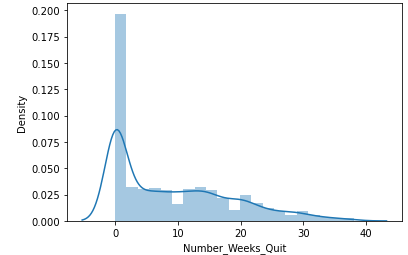

* We remove data from numerical columns as outliers by using zscore and then we remove data from categorical columns, which we already remove from numerical columns.
* After removing outliers, we merge numerical columns and categorical columns.
* Then, we split our independent variable as x and target variable as y for further process.
* As we see in data analysis, there is imbalanced classification in our target variable and to cop up with that issue we use oversampling technique.
* Oversampling uses synthetic techniques to increase the number of minority class samples to equal the number of majority class samples.

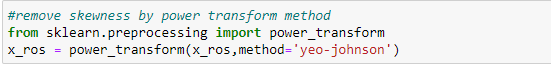
**Handling skewness:**

* As we see during symmetrical analysis, skewness present numerical attributes and need to remove.
* First, we check skewness using distribution plot.



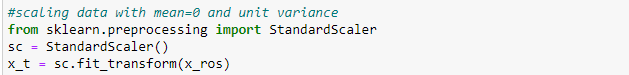


  
From the above plot we can see that skewness is present in each numerical attributes or data is not normal distributed. All the plots are right skewed, that means most of the data falls to the right. Thus, the distribution plot skews in such a way that its right side is longer than its left side. For removing skewness we use power transform method. This method removes skewness and makes data normally distributed. Below is the code for this method:



**Scaling:**

After removing skewness, we need to scaling our data. For this we use standard scaler method. Scaling is a method to normalize our independent variables. Scaling is essential for machine learning algorithms that calculate distance between data. For example, most of the classifier calculates the distance between two points by the distance. If one of the features has large value, then distance consider this particular feature. Scaling in necessary, where large and small values present in our data. Below is the code for scaling:



Here, we use standard scaler method. This method transform our data with mean = 0 and standard deviation = 1.

**EDA concluding remarks:**

Here, we complete our exploratory data analysis process. In conclusion, firstly, we do some data preparation for better data analysis and we draw some graphs for better understanding of the data. Secondly, we get some insights from the statistical summary and remove the imbalance classification from the target column. At the end, we remove outliers and skewness from the data, and scaling data by using standard scaler method.

* **Pre-processing pipeline:**

Here, we consider and explain topics which we covered above:

1. Data Preparation: In this section, we check null values and fill null values. Also drop ID columns because it is just identification number and not useful for model.
2. Data Analysis: In this section, we analyze data with crop damage variable in graphical format and do some other analysis which help in better understanding of data. After that we also see statistical summary of the data and this summary gives better insight of the data.
3. Handling Outliers: First we check outliers by plotting box plot and remove outliers by using zscore method because outliers impact badly on our model, so it is necessary to remove outliers.
4. After removing outliers, we use oversampling method to cop up with imbalanced classification of target attribute. This method increase the number of minority class samples to equal to the number of majority class samples.
5. Handling Skewness: We check skewness by plotting distribution plot and see that plot is right skewed. For removing skewness we use power transform method.
6. Scaling: At last, scaling of the data is important, where small and large values present in columns because it badly affect our model and give unnecessary results. For scaling data we use standard scaler method. This method transform data with mean = 0 and standard deviation = 1.

* **Building Machine Learning Model:**

The model building process consists of selecting model that is based on various machine learning methods used for experimentation. In this model building process we trained machine learning algorithm to predict the labels from the features. The aim is to find the best classifier model for analyzed problem and classifier with best classification results is used for prediction. The classification algorithm that we used is:

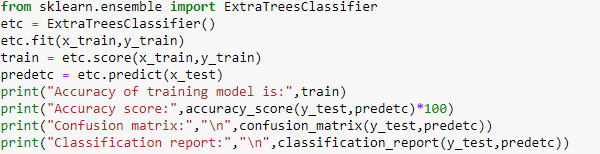
1. Logistic Regression
2. Decision tree classifier
3. SVC (support vector classifier)
4. GaussianNB
5. K neighbors classifier
6. Random forest classifier
7. Ada boost classifier
8. Gradient boosting classifier
9. Extra trees classifier
10. Catboost classifier
11. Xgboost classifier

We need to train the model to classify new observations. For training and testing we divided our original dataset into two parts training set and testing set. The accuracy of the machine learning algorithms increases with the amount of data using during training. The original dataset divided into two parts with 80:20 ratio. 80 percent of the data is use for training and 20 percent of the data is used for testing. Below is the code for splitting the data:

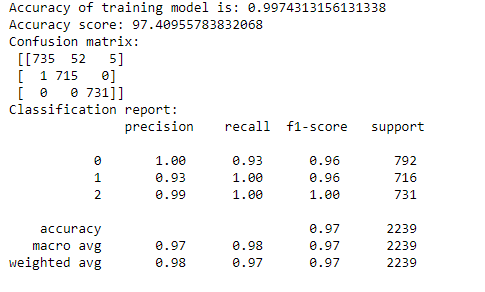


Model allow training set to find relation between variables and based on that model predict test result.

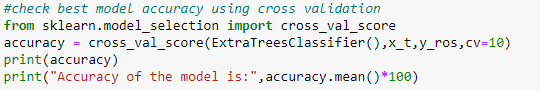
* Here, we check training set accuracy and testing set accuracy. For checking testing set accuracy we use accuracy score, confusion matrix and classification report. We use many algorithms to find best model, but here we describe only best model.
* We find our best model is Extra trees classifier. This classifier is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a forest to output its classification result. By using extra trees classifier we get best result. Below is the code for our best model:



Output:

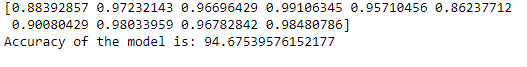


* After trying many algorithms we get good accuracy using extra trees classifier. We get 99.74 % training model accuracy and 97.40 % testing data accuracy with good confusion matrix and classification report.
* Now, we check extra trees classifier accuracy using cross validation to confirm that our model is not going through underfitting or overfitting. Below is the code for cross validation:



* Here, we use cv = 10, that means our training set is divided into 10 parts and provide mean accuracy of those 10 parts.

Output:



* We get 94% accuracy using cross validation that means our model is not underfitted or overfitted.
* Now, check accuracy of all used algorithms:

|  |  |  |
| --- | --- | --- |
| Algorithms | Training Accuracy | Testing accuracy |
| Logistic Regression | 0.52 | 0.52 |
| Decision tree classifier | 0.99 | 0.93 |
| SVC (support vector classifier) | 0.66 | 0.65 |
| GaussianNB | 0.52 | 0.53 |
| K neighbors classifier | 0.90 | 0.85 |
| Random forest classifier | 0.99 | 0.96 |
| Ada boost classifier | 0.55 | 0.56 |
| Gradient boosting classifier | 0.73 | 0.72 |
| Extra trees classifier | 0.99 | 0.97 |
| Catboost classifier | 0.96 | 0.92 |
| Xgboost classifier | 0.97 | 0.92 |

* Now save best result and best model. Below is the code:



* After saving best result and best model, we test our original test data using best model.



* **Conclusion:**
* To conclude, first we assess the data statistically and then we classified them. The dataset is processed, dividing into train set and test set. Then, we selected various classification algorithms, which provide different results and able to evaluate best classification result by using extra trees classifier. By using extra trees classifier, we predict that whether crop is healthy, damaged by other reasons and damaged by pesticides.
* Crop damage prediction helps farmers to when they are going to wrong and which type of changes they required for healthy crop.