CROP DAMAGE PREDICTION

USING MACHINE LEARNING ALGORITHM

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

In

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

By

KAGITHAPU ARCHANA (2103A52140)

PINGILI NITHYA (2103A52181)

Under the guidance of

Mr. D. Ramesh

Assistant Professor, Department of CSE.

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that project entitled **"CROP DAMAGE PREDICTION USING MACHINE LEARNING ALGORITHMS"** is the bonofied work carried out by **KAGITHAPU ARCHANA,PINGILI NITHYA.**

As a course project of the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY in ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year **2022-2023** under the guidance and supervision.

Mr. D.Ramesh

Asst. Professor,

SR university ,

Ananthasagar, Warangal.

Dr. M Sheshikala

Asst.Prof .&HOD(CSE) ,

SR University,

Ananthasagar, Warangal

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr.D.Ramesh, Asst. Prof** for guiding us from the beginning through the end of the course Project . We express our gratitude to Head of the department CS&AI, **Dr. M. Sheshikala** , **Associate Professor** for encouragement , support and insightful suggestions,. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean , School and Artificial Intelligence , **Dr C.V. Guru Rao** , for his continuous support and guidance to complete this project in the institute.

ABSTRACT

We all know that Agriculture is the back bone of our country. Cultivating crops i.e farming plays an important role all over the world. Without farming there will be no scope for surviving .

But now-a-days crop damage became a major problem for many of the farmers. This is happening due to many factors like Temperature, Season, Types of crops growing and many other factors. Many of the cultivators are not showing interest towards agriculture due to these factors which are becoming the major problems for getting less yield. So to get proper yield with in a desired time farmer should know the consequences before selecting the crop.

This research will help the farmer to know about the crop production rating before the farmer gets crop yield. Depending on the parameters which have been provided by the farmer we can predict the output in the form of accuracy using some classification algorithms which will give the farmer, whether the farmer will get good yield or damaged crop.

CONTENTS

CHAPTER-1

INTRODUCTION

Agriculture is an important occupation all over the world. Agriculture will help the country to grow its economic status. It is the large commercial sector among all other sectors. One can definitely get more profits if they had a good command in farming. As the population is growing across the world, to meet everyone's needs the crop production should be increased. There is huge demand for agriculture lands now-a-days. But there is lack of technology in agriculture sector so, we need to improve our technologies to get more yield in short span.

When the farmers get awareness regarding the technology then there will be more benefits to farmers as well as our country economic condition. We all know that our country will do more and more imports than exports this will effect the country's development. So, when we have a good crop production we can reduce the imports.

Hence, to improve crop production and predict the crop damage before its yielding time we came with a solution using machine learning. Machine Learning techniques can help agriculture experts make early decisions during complex situations. Classifying the stages of crop production is a challenging task, but this can be outstandingly handled by ML Classification techniques. Some of the standard methods used for classification are Logistic Regression and KNN etc…

1.1-PROBLEM STATEMENT:

Generally most of the farmers do not have any idea about crop choosing. They generally go with the flow that means they cultivate the crops according the trend going on. This may effect crop yield because they literally don't know which crop is suitable to that soil, temperatures, weather conditions and season etc…

The farmers now-a-days are habituated to use hybrid varieties which will give fast production. This will help the farmer for only one type of crop but as the time passes that will definitely effect the soil fertility. If the farmer want to rotate the crop the soil may not give better yield. So, to avoid such consequences the farmer should use some technologies to know about crop prediction.

By implementing few algorithms of Machine Learning we can find the accuracy and it predicts whether the crop gets damaged or we get a good yield. This will become a boon to many farmers which help them to get benefit and earn benefits.

The main motive is to prove the prediction accuracy using the different classification models and compare which model performs better regarding the problem.

1.2-EXISTING SYSTEM:

The existing solution available in now –a-days is created by the experts on considering the present demanding crops and they worked on them by taking the factors like climatic conditions ,soil type and availability of water etc, to know whether the crop will sustain or not

1.3-PROPOSED SYSTEM:

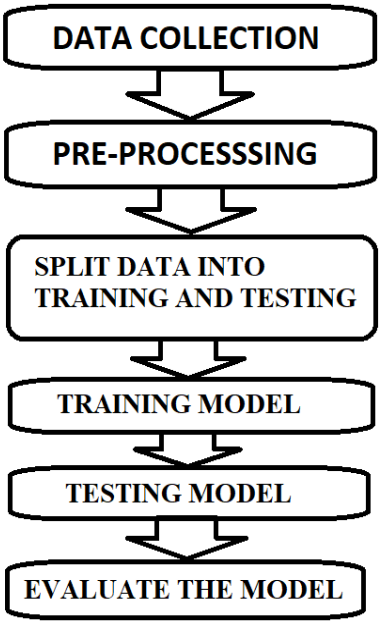
With the assist of data set obtained we create 5 different machine learning algorithms and they are **KNN,SVM,LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST.** We examined the outcomes of accuracy and found that which models performs the best**.**

1.4-OBJECTIVES:

- We are going to compare each and every accuracy.
- We are going establish the machine learning algorithm which gives the best accuracy.

1.5-ARCHITECTURE:

This is a Supervised learning approach. As the data have categorical values we used classification based machine learning algorithms to predict the condition of crop, which might be normal,suspect or damaged.



We collected the data set firstly and after collecting the data pre-processing is done.Then the data set is get divided into 2 sets(i.e training and testing). Using classification based machine learning models we trained the model after finding the accuracy on the training data set we found the accuracy on the testing model. Based on those conclusions we evaluated the model. And this is the architecture followed by us.

CHAPTER-2

LITERATURE SURVEY

Reference Number	Title	Description	Machine Learning Algorithm	Accuracy
[1]	**C**rop **Y**ield **A**nalysis **U**sing **M**achine **L**earning **A**lgorithms	Water, UV, Pesticides, Fertilizers are the features considered. They used SVM and LR and we get best accuracy for Linear Regression.	- **S**upport **V**ector **M**achine - **L**inear **R**egression	**80%** **81%**
[2]	**I**ntelligent **C**rop **R**ecommendation **S**ystem using **M**achine **L**earning	Parameters which are considered are Rain fall, Temperature, Geographical Location. Here, the system is divided into 3 modules namely crop analysis, crop recommended and crop sustainability predictor.	- **D**ecision **T**ree - **K N**earest **N**eighbour - **L**inear **R**egression - **N**aïve-**B**ayes - **N**eural**N**etwork - **S**upport **V**ector **M**achine	**81%** **85%** **88%** **88.26%** **82%** **89.88%**
[3]	**A**n **A**pplication of **M**achine-**L**earning **T**echnique-in **F**orecasting-**C**rop **D**isease	Crop patterns, Crop rotations, weather-parameters, environmental conditions, soil types, soil nutrients and etc., Most accurate results are got when ANN and LR are used. Activation functions used are Sigmoid, TanH and ReLu.	- **S**upport **V**ector **M**achine - **A**rtificial **N**eural **N**etworks - **L**ogistic **R**egression	**87%** **98%** **88%**

[4]	**M**achine **L**earning **A**pplications for **P**recision **A**griculture	Soil, Temperature, Rain fall, Humidity, pH, Nitrogen status, Sun shine hours, Fertilizers, Harvesting schedule are the features used here. By using few machine learning algorithms one can predict the crop damage before harvesting. DT will give the best accuracy.	- **S**upport **V**ector **M**achine - **D**ecision **T**ree - **R**andom **F**orest	**90.99%** **92%** **87%**
[5]	**R**eview on **C**rop **P**rediction **U**sing **D**eep **L**earning **T**echniques	Features used are Soil, climatic-conditions, water, temperature, rainfall, vegetative index, soil type, texture and nutrients. The study was done by the methods used were ANN, CNN, RNN which are deep learning methods. Out of these methods RNN gives the best accuracy.	- **A**rtificial **N**eural **N**etwork - **C**onvolutional **N**eural **N**etworks - **R**ecurrent **N**eural **N**etwork	**70%** **87%** **89%**
[6]	**C**rop **Y**ield **P**rediction **U**sing **M**achine **L**earning **A**lgorithms	Temperature , Rainfall , Area, Season are features which are considered. Random Forest gives the highest accuracy than other algorithms.	- **R**andom-**F**orest - **XGB**oost - **KNN** - **L**ogistic **R**egression	**67.80%** **63.63%** **43.25%** **25.81%**
[7]	**C**rop **Y**ield **P**rediction using **M**achine **L**earning **T**echniques	Naive Bayes method and K-Nearest-neighbour methods are used to find the accuracy. Here, accuracy of different crops are considered and compared. These will give the crop damage prediction. Out of both algorithms KNN gives best accuracy.	- **N**aive **B**ayes - **K**-**N**earest **N**eighbour	**69%** **78%**

[8]	**S**upervised **M**achine **L**earning **A**pproach for **C**rop **Y**ield **P**rediction in **A**griculture **S**ector	The proposed system mainly focus on yield, weather predictions and crop type which is going to cultivated. Algorithms proposed are RF & DT. Out of them RF will give the best accuracy than DT.	- **R**andom **F**orest - **D**ecision **T**ree	**Best accuracy** **Not-best accuracy**
[9]	**C**rop **Y**ield **P**rediction **B**ased **O**n **I**ndian **A**griculture **U**sing **M**achine **L**earning	Root mean square error is the metric which is used in this project. State, District, Crop, Season, Year, Area and Production are the features used in this project. Advanced regression techniques like Lasso, Kernel Ridge and stacked regression to minimize the error and they give best predictions.	- **S**tacked **R**egression - **K**ernel **R**idge - **L**asso	**96%** **98%** **99%**

2.1-Analysis of the Survey:

As we have seen above results we came to know that there is no model which is giving best accuracy to all the crops. By the survey we concluded that the people who had done all the experiments were did only on specific crops using specific parameters. So, we came up with an an idea which is going to predict any type of crop damage with good accuracy.

CHAPTER-03

DATA PRE-PROCESSING

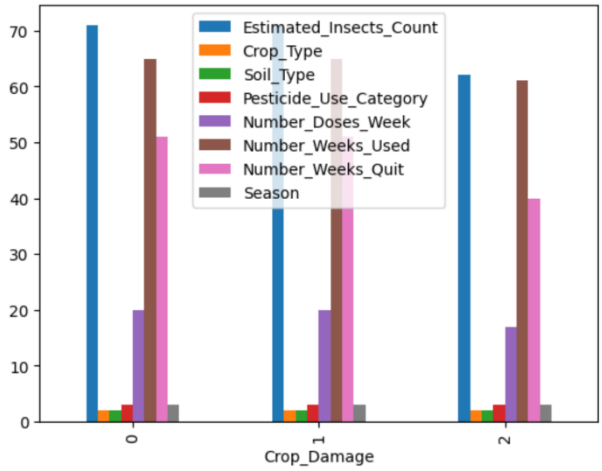
This data set contains 88858 records, which were then classified into 3 classes: Normal, Damage, Pathological Values(0,1,2).

3.1-Data Description:

The data set has a column with string values hence,we removed that column i.e, ID. The data set contain some missing values in the column so, we used mean of that particular column and filled the missing values.

ID	Estimated	Crop_Type	Soil_Type	Pesticide_	Number_I	Number_V	Number_V	Season	Crop_Damage
F00000001	188	1	0	1	0	0	0	1	0
F00000003	209	1	0	1	0	0	0	2	1
F00000004	257	1	0	1	0	0	0	2	1
F00000005	257	1	1	1	0	0	0	2	1
F00000006	342	1	0	1	0	0	0	2	1
F00000008	448	0	1	1	0		0	2	1
F00000009	448	0	1	1	0		0	2	1
F0000001C	577	1	0	1	0	0	0	1	2
F00000012	731	0	0	1	0	0	0	2	0
F0000002C	1132	1	0	1	0	0	0	1	2
F00000021	1212	1	0	1	0		0	3	0
F00000023	1575	0	0	1	0	0	0	1	1
F00000024	1575	0	1	1	0	0	0	2	1
F00000028	1575	1	1	1	0	0	0	2	1
F00000029	1575	1	1	1	0	0	0	2	2
F0000003C	1785	1	1	1	0	0	0	2	1
F00000035	2138	0	1	1	0	0	0	1	1
F00000037	2401	0	1	1	0		0	1	1
F00000038	2401	1	1	1	0	0	0	2	1
F00000039	2401	1	1	1	0	0	0	2	1
F00000045	2999	0	1	1	0	0	0	3	1
F00000048	3516	1	0	1	0	0	0	2	0
F00000049	3895	1	1	1	0	0	0	1	1
F0000005C	4096	1	1	1	0	0	0	2	1

The data set contains 88858 rows and 10 columns including target column i.e, Crop_Damage. There are 9000 missing values in Number_Weeks_Used.We replaced that missing values using mean fill method.



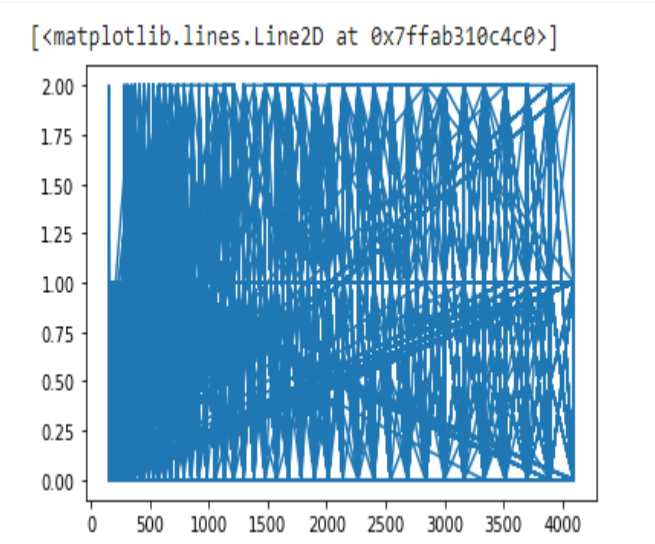
This graph represents the distribution of target variable with every feature in the data set.

COLUMNS AND NO.Of VALUES

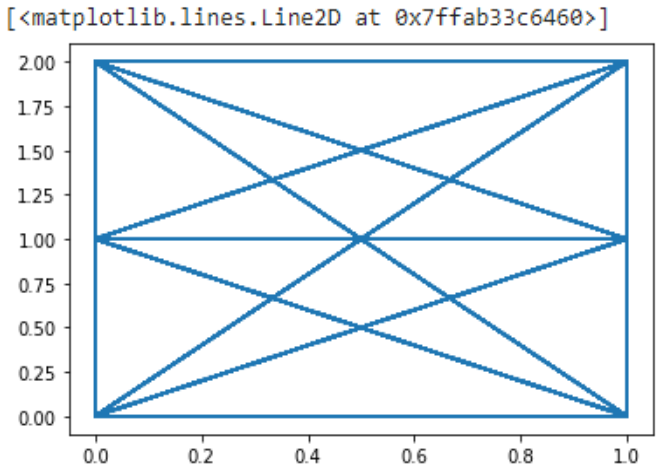
ID	88858
Estimated_Insects_Count	88858
Crop_Type	88858
Soil_Type	88858
Pesticide_Use_Category	88858
Number_Doses_Week	88858
Number_Weeks_Used	79858
Number_Weeks_Quit	88858
Season	88858
Crop_Damage	88858

3.2-Data Visualization:

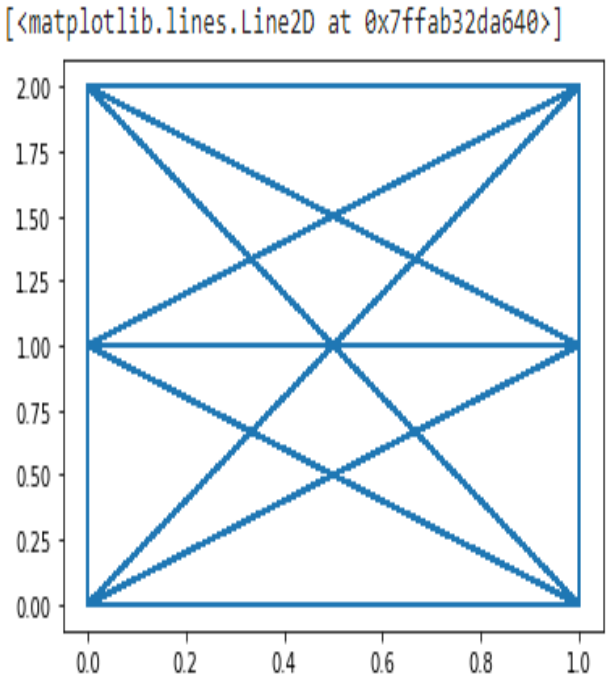
The following are the graphs which represent the distribution of data



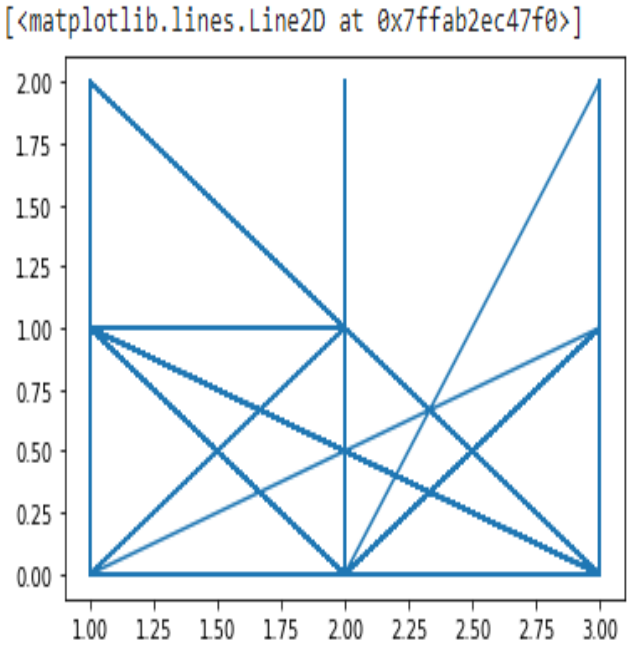
Estimated_Insects_Count



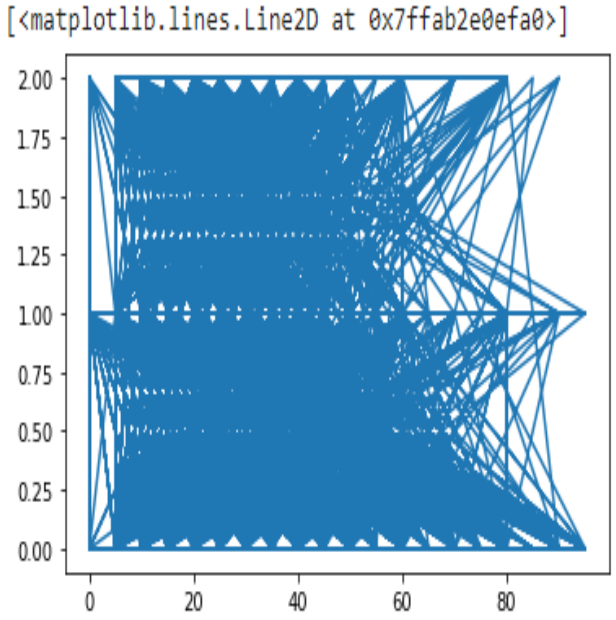
Crop_Type



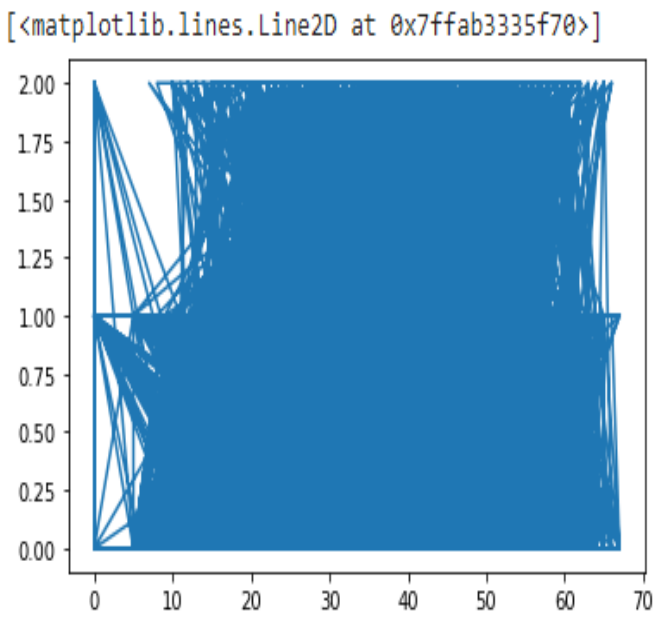
Soil_Type



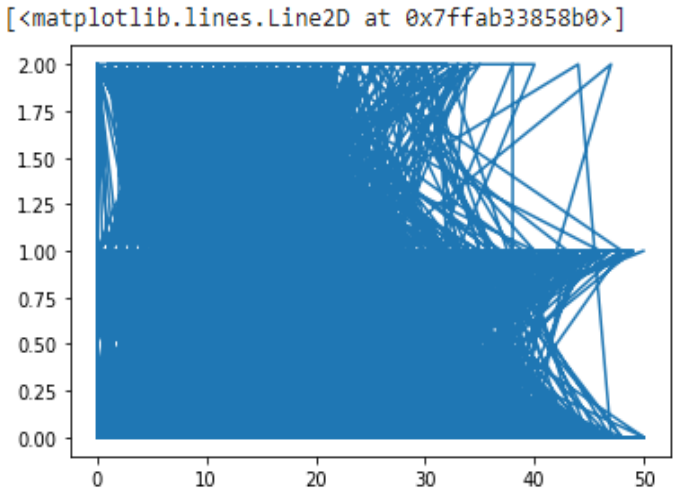
Pesticide_Use_Category



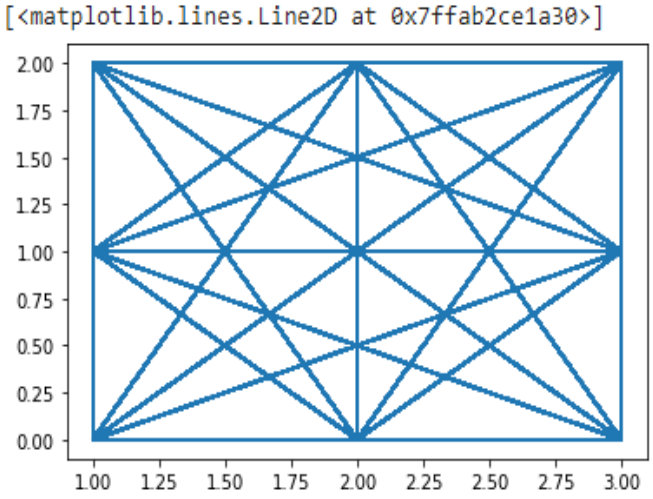
Number_Doses_Week



Number_Weeks_Used

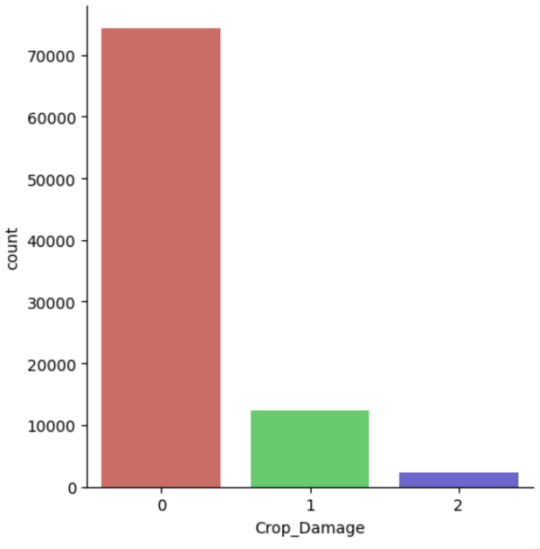


Number_Weeks_Quit

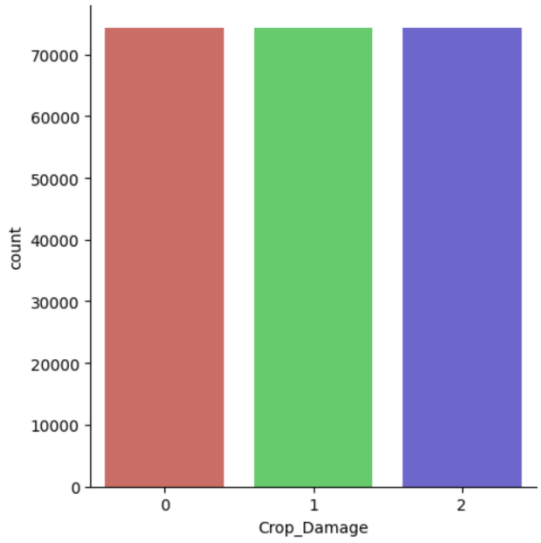


Season

TARGET VARIABLE DISTRIBUTION GRAPH

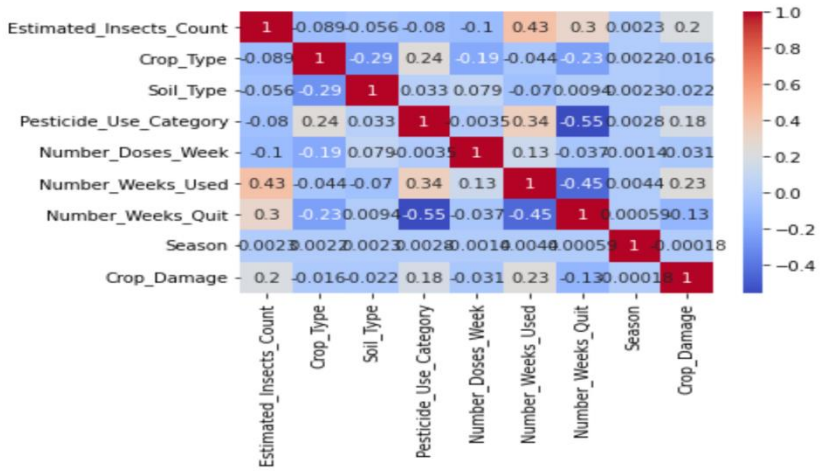


This graph represents the distribution of target variable.From the graph we can conclude that the data set is imbalanced and to balance the data set we use a method called random sample which is going to balance the data



This is the graph which is obtained after balancing the data using random sampler method.

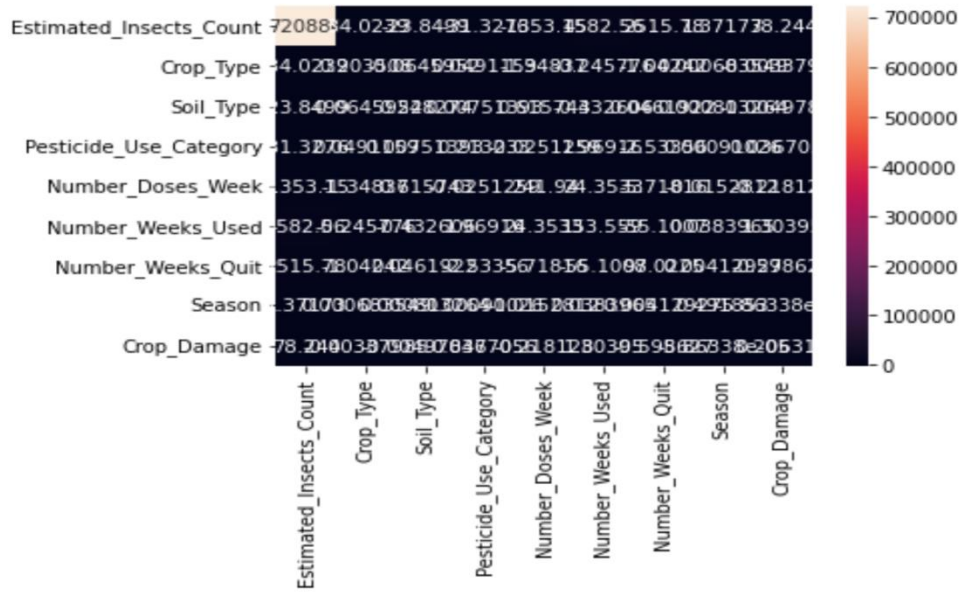
CORRELATION MATRIX



Correlation matrix is used to evaluate the relationship between the variables.

Here 1 represents strong correlation,0 represents neutral relation,-1 represents not strongly related

COVARIANCE MATRIX



Covariance depends on the sign. And it is used to determine how much a variable changes randomly.

Positive sign indicates variables in the same direction and negative indicates the opposite direction.

CHAPTER 4

METHODOLOGY

4.1. PROCEDURE TO SLOVE THE GIVEN PROBLEM:

4.1.1 K-Nearest Neighbors:

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

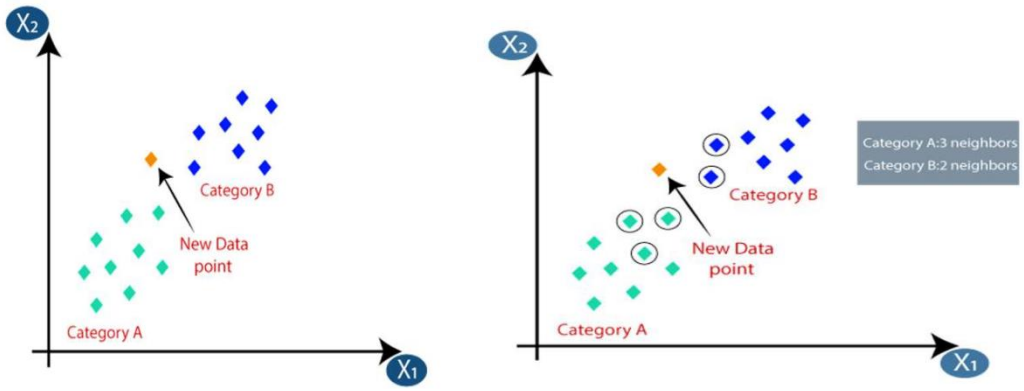
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



Firstly, we will choose the number of neighbors ,so we will choose the k=5.Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between points, which we have already studied in geometry. By calculating the Euclidean distance, we get the nearest.Neighbors,as three nearest neighbors in category A and two Nearest neighbors in category B.

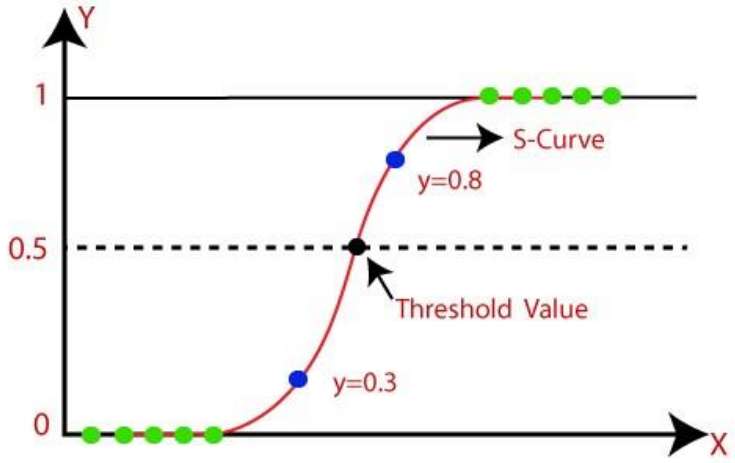
As we can see the 3 nearest neighbors are from category A,hence this new data point must belong to categ

4.1.2. Logistic Regression:

Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

Logistic regression is used when the dependent variable is binary such as click on a given advertisement link or not, spam detection, Diabetes prediction, the customer will purchase or not, an employee will leave the company or not.

Logistic regression uses Maximum Likelihood Estimation (MLE) approach i.e., it determines the parameters (mean and variance) that are maximizing the likelihood to produce the desired output.



Logistic Regression uses a sigmoid or logit function which will squash the best fit straight line that will map any values including the exceeding values from 0 to 1 range. So, it forms an "S" shaped curve.

Sigmoid function removes the effect of outlier and makes the output between 0 and 1.

The logistic function is of the form:

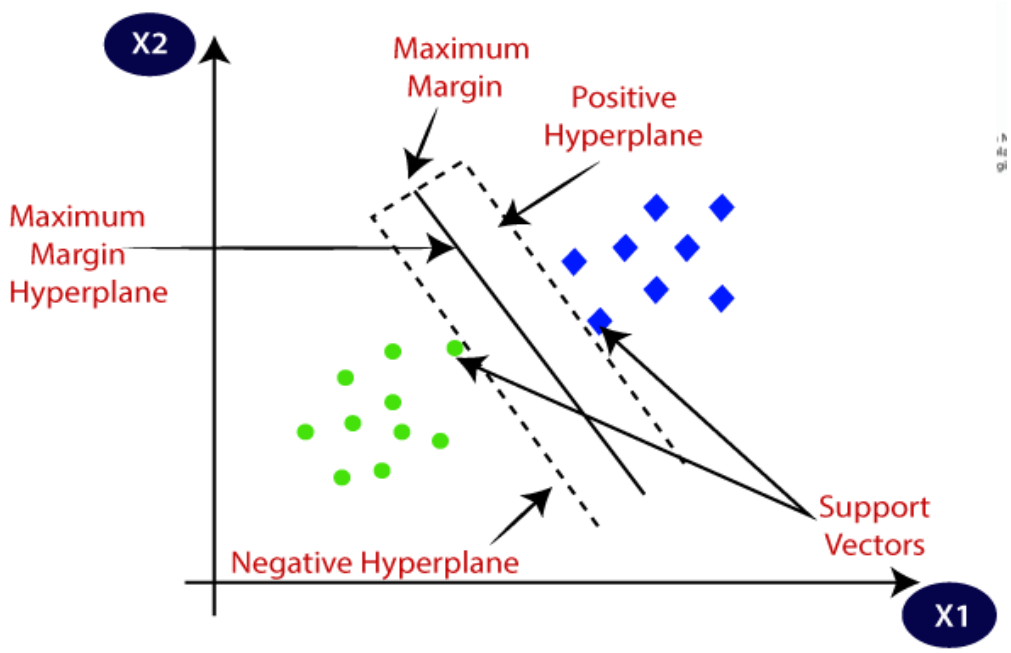
P(X)=1/1+e^-z

where μ is a location parameter

s is a scale parameter.

4.1.3 SVM(Support Vector Machine):

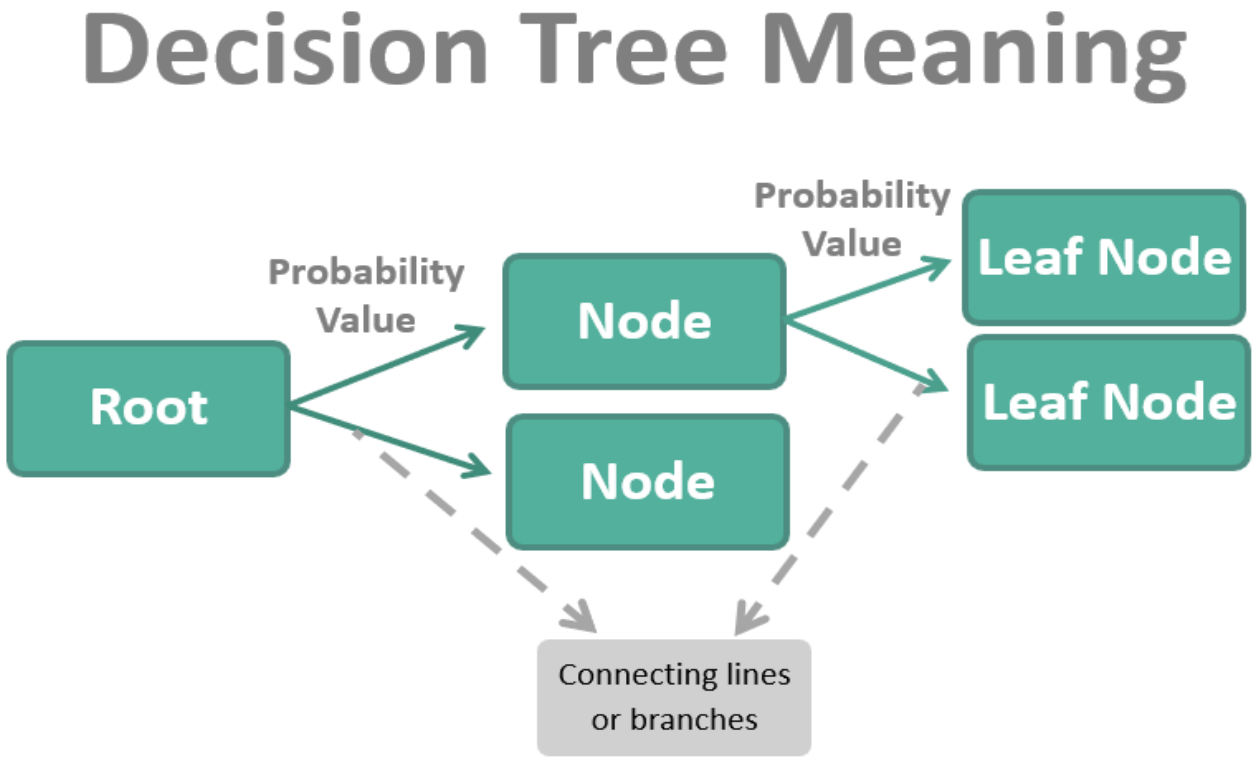
- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. In the SVM method,we plot each data item as a point in n-dimensional space(where n is no.of features you have).
- We perform classification by finding the hyperplane that differentiates the two classes very well.
- We have three hyper planes(A,B and C). Now,identify the right hyper plane to classify star and circle.
- We want our data points to be as far away from the hyperplane as possible,while still being on the correct side of it.
- The distance between the hyperplane and the nearest data point from either set is known as margin.
- The goal is to choose a hyperplane with the greatest possible margin.
- There will never be any data point inside the margin.



4.1.4 Decision Tree:

- It is a non parametric method used for supervised learning method used for both classification and regression.
- It uses tree representation to solve the problem.As deeper the tree goes, the more complex the decision rules and the fitter the model.
- Entropy and Information Gini are used to calculate root node among all the nodes.
- Hence,an optimal tree can be formed.
- From given data a tree can be formed and using entropy & information gini we can calculate accuracy.
- Tree is a hierarchical representation(pictorial representation).
- ENTROPY:Entropy is the measure of uncertainity of a random variable.The higher entropy results in more information.
- **Entropy**=summation(-p*log(p))
- INFORMATION GINI:Information gini is the measure of changes in the entropy.

- **Information Gini**=E(T)-E(T,X)

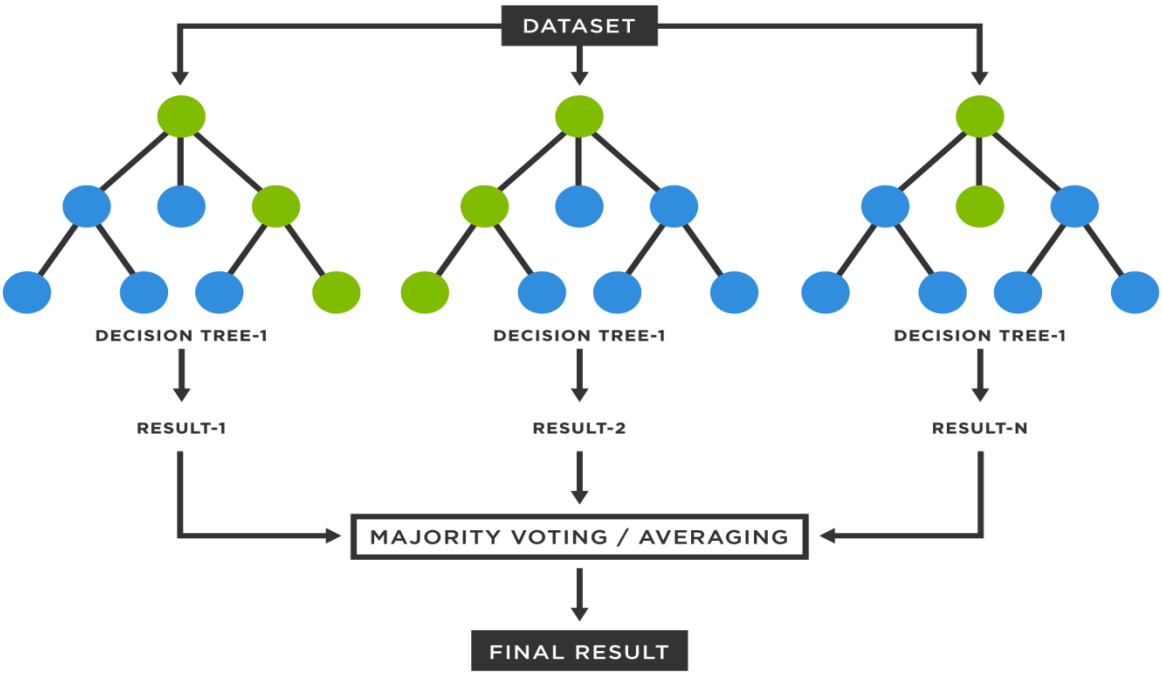


4.1.4 Random Forest:

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Random forest algorithms have three main hyper parameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged.



4.2-SOFTWARE DESCRIPTION:

We used the Google collab service to test our machine learning algorithms written in Python. The notebooks produced the following results.

#LOADING THE DATA SET:

```
import pandas as pd
from matplotlib import pyplot as plt
import numpy as np
train = pd.read_csv('/content/train.csv')
#FILLING NULL VALUES Using mean of that column:
train['Number_Weeks_Used'].fillna(train['Number_Weeks_Used'].mean(),inplace=True)
#DROPPING THE COLUMN 'ID' due to string values:
train.drop(columns=["ID"],axis=1,inplace=True)
#DROPPING TARGET COLUMN FROM FEATURE COLUMNS:
X=train.drop(columns=['Crop_Damage'])
Y=train[["Crop_Damage"]]
#SPLITTING THE DATA SET INTO TRAINING AND TESTING:
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25,stratify=Y, random_state=2)
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
scaler.fit(X_train)
X_train=scaler.transform(X_train)
X_test=scaler.transform(X_test)
```

4.2.1 THROUGH KNN:

```
from sklearn.neighbors import KNeighborsClassifier
classifier=KNeighborsClassifier(n_neighbors=5)
classifier.fit(X_train,Y_train)
```

#Above is the python code which loads the k nearest neighbors model and performs knn on the given training and testing data.

CONFUSION MATRIX:

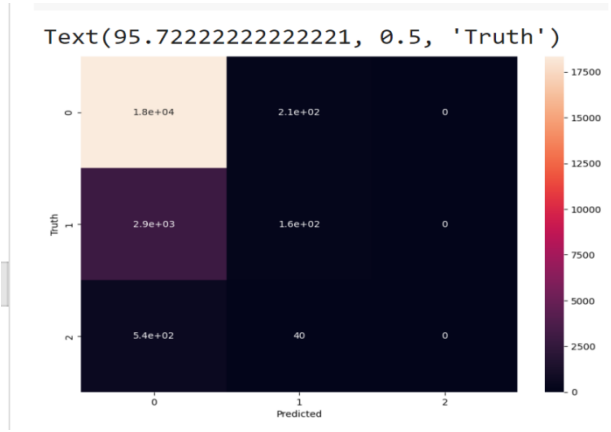
```
y_pred=model.predict(X_test)
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test,y_pred)
cm
```

OUTPUT:

```
array([[18350, 210, 0],                                                    [
2920, 157, 0],                                                             [
538, 40, 0]])
```

```
import seaborn as sn
plt.figure(figsize = (10,7))
sn.heatmap(cm, annot=True)
plt.xlabel('Predicted')
plt.ylabel('Truth')
```

OUTPUT:



- This picture represents the confusion matrix which is obtained after performing KNN model.
- Here, the matrix is imbalanced(I.e, the diagonal is less than the other values.

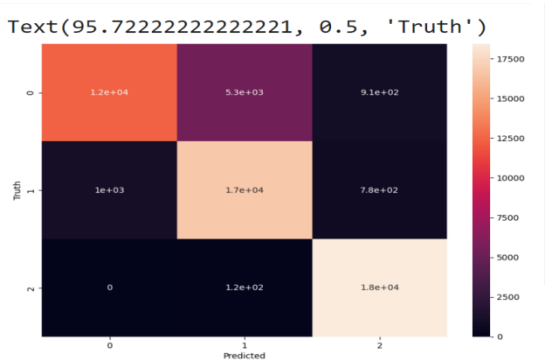
AFTER BALANCING THE DATA SET:

#Code to balance the data set:

```
from imblearn.over_sampling import RandomOverSampler
ros=RandomOverSampler(sampling_strategy="not majority")
x_res,y_res=ros.fit_resample(X,Y)
```

CONFUSION MATRIX:

```
array([[12333, 5315, 912],                                    [
1026, 16749, 785],                                              [
0, 118, 18441]])
```



- This picture represents the confusion matrix which is obtained after balancing the data set and performing KNN model.
- Here, the matrix is balanced(I.e, the diagonal is greater than the other values.

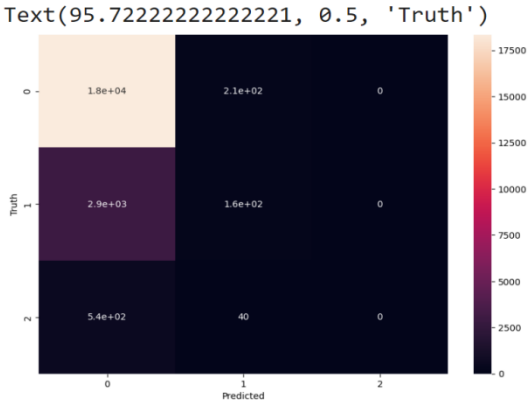
4.2.2 THROUGH LOGISTIC REGRESSION:

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
model = LogisticRegression()
model.fit(X_train, Y_train.values.reshape(-1,))
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

#Above is the python code which loads the logistic regression model and performs logistic regression on the given training and testing data.

CONFUSION MATRIX:

array([[18350, 210, 0], [
2919, 158, 0], [538,
40, 0]]

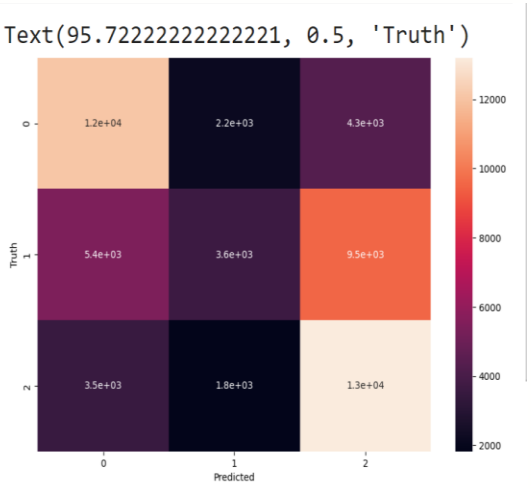


- This picture represents the confusion matrix which is obtained after performing LOGISTIC REGRESSION model.
- Here, the matrix is imbalanced(I.e, the diagonal is less than the other values.

AFTER BALANCING THE DATA SET:

CONFUSION MATRIX:

array([[12070, 2185, 4305], [
5556, 3608, 9396], [
3539, 1746, 13274]])



- This picture represents the confusion matrix which is obtained after balancing the data set and performing LOGISTIC REGRESSION model.
- Here, the matrix is balanced(I.e, the diagonal is greater than the other values.

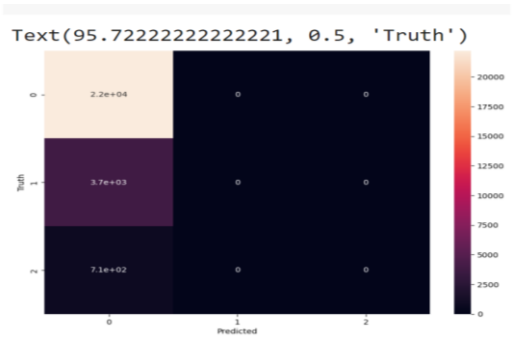
4.2.3 THROUGH SVM:

```
from sklearn import svm
clf = svm.SVC(kernel='linear')
clf.fit(X_train, y_train.values.reshape(-1,))
SVC
SVC(kernel='linear')
```

#Above is the python code which loads the supporting vectors model and performs svm on the given training and testing data.

CONFUSION MATRIX:

array([[22203, 0, 0], [
3747, 0, 0], [
708, 0, 0]])



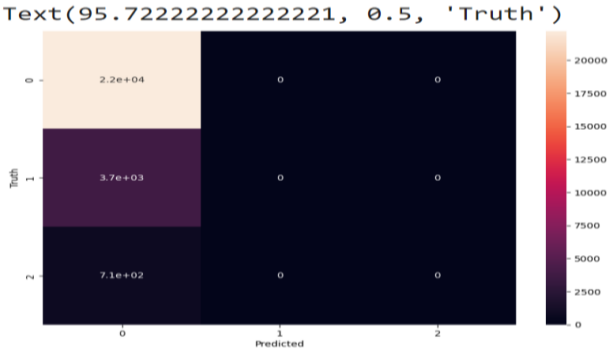
- This picture represents the confusion matrix which is obtained after performing SVM model.
- Here, the matrix is imbalanced(I.e, the diagonal is less than the other values.

4.2.4 THROUGH DECISION TREE:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25,stratify=Y, random_state=2)
d=DecisionTreeClassifier()
d=d.fit(X_train,Y_train)
```

CONFUSION MATRIX:

array([[15786, 2382, 392], [
2064, 823, 190], [321,
210, 47]])



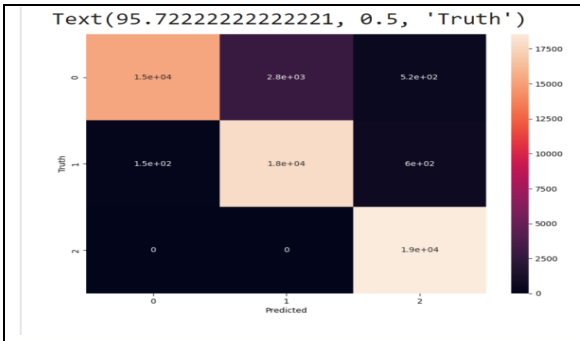
- This picture represents the confusion matrix which is obtained after performing DECISION TREE model.
- Here, the matrix is imbalanced(I.e, the diagonal is less than the other values).

AFTER BALANCING THE DATA SET:

CONFUSION MATRIX:

array([[15251, 2792, 517], [
152, 17806, 602], [0,
0, 18559]])

- This picture represents the confusion matrix which is obtained after balancing the data set and performing DECISION TREE model.
- Here, the matrix is balanced(I.e, the diagonal is greater than the other values).

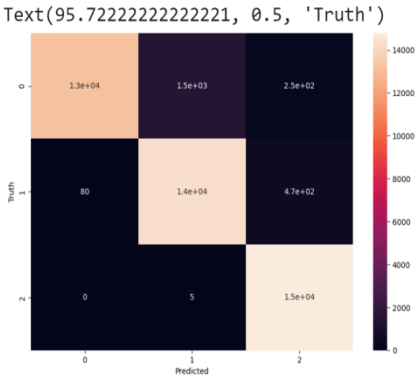


4.2.5 -THROUGH RANDOM FOREST:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score, recall_score, ConfusionMatrixDisplay
from sklearn.model_selection import  train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(x_res, y_res, test_size=0.2)

rf = RandomForestClassifier()

rf.fit(X_train, Y_train.values.reshape(-1))

y_pred = rf.predict(X_test)
```

CONFUSION MATRIX:

```
array([[13154, 1463, 248],
80, 14326, 474],
0, 5, 14793]])
```



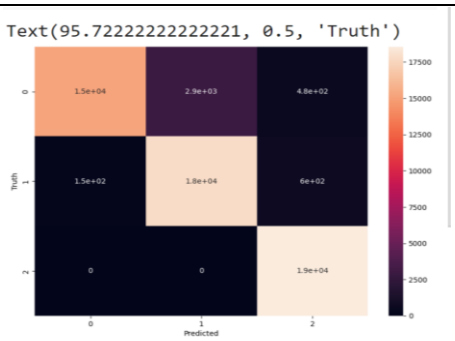
- This picture represents the confusion matrix which is obtained after performing RANDOM FOREST model.
- Here, the matrix is imbalanced(I.e, the diagonal is less than the other values).

AFTER BALANCING THE DATA SET:

CONFUSION MATRIX:

```
array([[13154, 1463, 248],
80, 14326, 474],
5, 14793]])
```

- This picture represents the confusion matrix which is obtained after balancing the data set and performing RANDOM FOREST model.
- Here, the matrix is balanced(I.e, the diagonal is greater than the other values).



CHAPTER-5

RESULTS

MACHINE LEARNING ALGORITHMS	ACCURACY BEFORE BALANCING THE DATA SET	ACCURACY AFTER BALANCING THE DATA SET
LOGISTIC REGRESSION	**Accuracy:** 0.8333508395480396	**Accuracy:** 0.5180111952584787
K-NEAREST NEIGHBORS	**Accuracy:** 0.8298897141571011	**Accuracy :** 0.8535174841502182
SUPPORT VECTOR MACHINE	**Accuracy :** 0.8328831870357867	**Accuracy :** 0.8354715282466801
DECISION TREE	**Accuracy :** 0.7505739365293721	**Accuracy :** 0.9270281434652203

RANDOM FOREST	**Accuracy :** 82.29237002025658	**Accuracy :** 94.98013155826955

After balancing the data set we got better accuracy than before.Highest accuracy is acquired in Random Forest.

CHAPTER-6

CONCLUSION

This study has explored the impact of features like season , crop type and soil type etc on crop damage.This work advocates for early information sharing specifically on expected yield so one can ensure a proper planning before growing the crop .

The entire aim of this project is to predict crop damage before getting yield . This can be done using some machine learning techniques like KNN , Logistic Regression , Support Vector Machine, Random Forest.

This research will help the farmer to know about the crop production rating before the farmer gets crop yield. Depending on the parameters which have been provided by the farmer we can predict the output in the form of accuracy using some classification algorithms which will give the farmer, whether the farmer will get good yield or damaged crop.

CHAPTER-7

FUTURE SCOPE

In future, new features from the fields can be gathered to get a perfect image of the crop damage using other machine learning algorithms and deep learning algorithms such as ANN or CNN to get more accurate predictions.

In coming years, this approach can be developed for any type of users like low level farmers who are uneducated , by adding some additional features into the machine learning model which can be learned by any type of user in hardly 1 hour . This research work can be enhanced to higher level by availing it to whole India for country's development.

CHAPTER-8

REFERENCES

[1] https://ieeexplore.ieee.org/abstract/document/9221459

[2] https://ieeexplore.ieee.org/abstract/document/9418375

[3] https://dl.acm.org/doi/abs/10.1145/3372454.3372474

[4] https://ieeexplore.ieee.org/abstract/document/9311735

[5] https://iopscience.iop.org/article/10.1088/17426596/1767/1/012026/meta

[6] https://ieeexplore.ieee.org/abstract/document/9221459

[7] https://ieeexplore.ieee.org/abstract/document/9033611

[8] https://ieeexplore.ieee.org/abstract/document/9137868

[9] https://ieeexplore.ieee.org/abstract/document/9154036